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 十一 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

November 12, 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。社交媒体用户群体庞大，观点涉及话题广泛，使得网络成为能够挖掘出关于各种话题的大众观点的宝贵资源库。然而社交媒体中的观点常常是通过带有噪声的非结构化文本碎片中表达出的，并且这些碎片化文本分散在不同的来源（不同的用户），人工就某一话题去浏览所有的文本碎片并分析总结出相关的观点是非常困难的，需要以计算手段自动分析，整合并总结出所有文本中的观点信息。本文主要研究社交媒体用户观点的自动分析（包括观点挖掘和整合集成），主要目标是对用户在社交媒体上就所关注话题发表的大量观点更好地建模，并基于此模型进一步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，本文确定了观点分析的三个主要步骤：情感知识的抽取，观点文本情感极性分类，用户观点的集成。这三个步骤组成了一个观点集成综合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协同关联的观点分析与应用任务提出了新的通用的处理方法：

- **中文情感词典的抽取和构建**：目前表示情感知识的词典主要是在英文中构建的英文情感词典，这些词典在观点文本识别、极性分类等任务中起到了重要作用，是进行观点分析的基础。中文情感词典抽取和构建方法研究相对较少，还没有形成比较全面可靠的情感词典。靠人工编辑形成的情感词典费时费力，覆盖度偏低，因此本文根据不同语言间表达情感知识词汇间的对映性，借鉴已有的英文情感词典中的词语情感知识，使用 HowNet 语义知识库中词语的双语语义描述转化英文情感词典的情感知识，抽取中文情感词汇并计算情感极性值，形成了自动构建的中文情感词典 SentiHowNet。为了提高词典的覆盖度以及领域情感知识的适应性，分析验证了基于语料资源中连词语言规则和上下文语境统计特征的情感词典扩展方法，并提出了混合两种方法的扩展方法对 SentiHowNet 在领域语料内进行了扩展。使用本文方法得到的中文情感词典可以自动构建无需人工标注，与其他几个词典相比覆盖度和领域适应性更好。

- **基于特征空间划分的情感极性分类**：情感极性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，可以看作是一种特殊的文本分类。

情感极性分类最常用的词袋模型中，用以表达情感的词语特征常常起到不同的作用，有些词语具有通用情感表达作用，能在不同领域和语境中表示不变的情感极性，而有些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。因此本文提出了将特征空间划分为领域独立和领域依赖两部分的情感极性分类方法，该方法分别在两部分特征空间上训练分类器，然后将两个分类器组合在一个框架中形成一个更强的情感极性分类器，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，可以在无需标注数据进行训练情况下达到有监督机器学习方法的性能。

- **用户观点的集成建模**：社交媒体中用户产生的内容往往是短小而又分散的信息碎片，因此用户针对所关注话题的观点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所关注话题以及用户针对话题的观点组合在一起进行建模，模型中将观点按照话题的不同方面进行整合集成，并提出一种通用的可扩展观点表示方法，将同一话题的观点表示为在一个可扩展的情感值空间的分布，这种表示能够表达出用户更详细和多视角下的观点信息。

- **用户交互行为分析**：作为用户主观模型的直接应用，本文对用户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。

在对以上四个观点分析与应用任务的研究中，本文侧重于使用通用的鲁棒性好的无监督或弱监督方法，因此本文的方法适用于话题广泛的大量观点的自动分析，这也使我们的方法区别于针对特定领域精心进行特征设计并使用大量标注数据进行充分训练的有监督机器学习方法，因为这些方法转换到新领域就会变得性能下降，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。基于这种思路我们的方法显示出良好的通用性并达到一定的评测性能，能够在多个研究领域（比如商业智能和社会学研究）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 noisy text fragments scattered in different sources(i.e., different users), 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 all the text fragments. This thesis focuses on the problem of automatic opinion analysis including opinion mining, integration and summarization, whose goal is to better support modeling huge amounts of opinions for all topic of interests of social media users, and further to analyze their interaction behaviors based on these opinions.

To systematically study this problem, we have identified three important steps of opinion analysis: extraction of sentiment knowledge, sentiment polarity classification of opinionate text, and opinion integration of users. These steps form three key components in an integrated opinion summarization system，the results of which are used to promote online behavior analysis of users. 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 analysis and play important roles in tasks such as opinionative text identification and feature selection of sentiment classification, etc. There are relatively few studies on extraction and construction of Chinese sentiment lexicon, and there is no comprehensive and dependable Chinese sentiment lexicon available yet. The sentiment lexicon compiled by human is time-consuming and laborious, while has a low coverage. Therefore based on the sentiment knowledge mapping between words of different languages, and drawing from current English sentiment lexicons, we proposed a novel method to identify a number of Chinese sentiment words and calculate their sentiment polarity value using bi-linguistic semantic definition of HowNet knowledge resources, which formed a Chinese sentiment lexi-

con named SentiHownet. In order to improve coverage and domain adaptability of SentiHownet, we analyzed and verified language rules based extension method and corpus based statistical context features extension method with experiments, and proposed a hybrid method by combining two methods. The SentiHownet lexicon is constructed automatically without human annotation, which has wider coverage and better adaptability for domain opinion analysis than other Chinese sentiment lexicons.

- **Sentiment polarity classification based on feature space division:** Sentiment classification classifies the text into predefined categories according to features co-occurence, and can be regarded as a kind of special text classification. The bag-of-words features of sentiment classification are often used with different functions: some features represent the same general sentiment polarity across different domains and context, while others represent specific sentiment polarity only in specific domain or context. Therefore, we proposed to divide the feature space of sentiment classification task into two separate parts, including domain-dependent part and domain-independent part. Two different classifiers are learned using two feature parts, and then combined together into a stronger sentiment polarity classifier in a bootstrapping framework.The framework started training on an off-the-shelf idiom resources without annotation in a bootstrapping way. The proposed method can achieve the performance of supervised methods without any annotation dataset.

- **Integration of opinions of users:** User-generated content(UGC) of social media are often short and dispersed text fragments, so that the opinions of users about topic of interests are scattered in the unstructured fragmented short text. To be able to digest opinions of users comprehensively and accurately, we proposes the concept of subjectivity model by combining the topics and opinions together, in which the opinions are integrated according to the different aspects of the same topic articulated in the UGC. We also put forward a general representation of opinion, which defined opinion as sentiment distribution over a scalable sentiment value space, and provided a more detailed and informed multi-perspective view of the opinions.

- **Interaction behaviors analysis of users:** As direct applications of subjectivity model, we analyzes the subjective motivation of the information dissemination behavior 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 for conformity needs because the message is popular, we proposed three subjectivity similarity measurements. For retweeting behavior analysis, the three subjectivity similarities are verified to be correlated to the retweeting behavior, and can serve as 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 analysis. 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 sociological Research.

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，互联网已经开始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[1]，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，互联网普及率已达到 47.4%[2]。随着互联网技术的迅猛发展，网络进入 Web2.0 时代，出现了各种各样吸引广泛用户参与的社交媒体（Social media）平台。社交媒体逐渐开始将网络和人类社会生活融合在一起，已经开始成为工作、学习以及日常生活中必不可少的重要部分。

Web2.0 时代的互联网用户不再只是单纯的信息接收者，同时也成为网络内容的产生者，用户可以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[3]所做的网络访问量统计，从表中可以看出，流量前十的互联网网站中社交媒体就有七个。

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] 表中加黑的网站为社交媒体

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[4]：

[1]http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201407/P020140721507223212132.pdf

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

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

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

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，社交媒体是以网络技术和互联思维为基础的一项应用，用户因此可以非常方便进行内容创作、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。

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

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，主要包括文本、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），这些信息依靠用户间建立的社交关系以及信息交互形成相互关联的庞大数据库。Kaplan 和 Haenlein[1] 从信息产生和流动角度定义了社交媒体：首先是作为媒体（media），社交媒体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），社交媒体中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产生者之间不停地改变自己的角色；其次，之所以称这种新媒体是社会化的（social）的媒体，因为作为社交工具，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，也成为用户间信息交流、信息共享以及信息传播的便利工具。

虽然社交媒体的出现为用户社会生活提供了便利，但是随着用户数量不断增加，产生的内容不断达到新的量级，导致用户作为信息消费者遇到了一些新的挑战，用户从“信息海洋”中找到有用信息变得更加困难，这种现象称为信息超载

（information overload）[5]。同时，由于社交媒体发布信息的快捷和方便, 使得网络内容产生门槛降低，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[6]，每分钟有超过 3000 张照片上传到 Flickr[7]，每年有超过 160 多万的博客（blogs）发表[3]。

- **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提出了新的挑战[4, 5]。

- **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[6]。不仅于此，社交媒体中的网络结构也充满噪声，这主要是因为：一是网络中存在一些传播虚假和垃圾信息的用户[7]；二是用户间建立关系的便捷性很容易使得用户将各种社交关系放在一起，并不区分好朋友和陌生人[8]。

- **非结构化（Unstructured）**：社交媒体中用户产生数据，特别是文本数据，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[8]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[9]，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（How are you）?）等[10]。

- **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[11]，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、经济以及教育等多方面发挥着越来越重要的作用。一些互联网公司以社交媒体大数据资源为支撑，以 SaaS

[5]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。
[6]http://www.twitter.com/
[7]http://www.flickr.com/
[8]http://www.facebook.com/

（Software as a service）形式为用户提供在线服务。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[9]。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挑选了粉丝在 Facebook 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[10]：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，主要是互联网数据，来实现的一些相关应用。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，由于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。

[9]http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[10]来源：工业和信息化部电信研究院于 2014 年 5 月发布的《大数据白皮书》

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[12] 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出的对于所谈论目标事物的态度、情感或者看法[13]。观点常常简化为对目标事物的认同或不认同（或者认为目标事物好或者坏，持积极（Positive）态度还是消极（Negative）态度）[14] 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（Artitles）、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：可以在 Amazon[11], Yelp[12], 以及 TripAdvisor[13]上发表各种商品和服务的评论；可以在 Twitter[14]和 Facebook 上对最新话题表达自己的观点。社交媒体因为拥有庞大的用户群以及用户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。社交媒体中的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[15] 发现网络中进行信息宣导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[16] 以及政策制定者[17] 不得不考虑的重要因素，有证据显示这种观点的相互影响过程具有显著经济效应[18–20]。此外，大规模的用户意见整合形成的观点可以反映民众政治倾向[21]，甚至可以提高股票市场的预测效果[22]。

社交媒体海量的用户产生内容不可能依靠人工地去发现和总结其中的观点信息，需要能够从文本中自动对观点信息进行挖掘和分析的计算方法。观点分析[15][23] 就是对文本中带有情感色彩的主观信息进行分析、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。尽管计算语言学和自然语言处理已经有很长的研究历史，但

[11] www.amazon.com

[12] www.yelp.com

[13] www.tripadvisor.com

[14] www.twitter.com

[15] 本文中观点分析（Opinion analysis）综合了情感分析 (Sentiment analysis)，观点挖掘（Opinion mining）以及主观性分析（Subjectivity analysis）等任务，是对文本中的观点信息进行挖掘、分析以及总结的过程。

是直到 2000 年才开始进行观点挖掘和情感分析等观点分析任务的研究，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，研究者首次拥有了大量的具有观点信息的文本数据，也正是因为有了这些数据，规模性的观点分析研究才成为了可能。可以说观点分析与社会媒体是一同起步和成长的，是社交媒体中数据分析的核心研究。观点分析研究不仅对于自然语言理解（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，因为这些领域都受到人的主观性的影响。

为了便于后续阐述，本文首先要针对观点分析研究中要用到的以下几个概念进行定义加以明确。

定义 (文档 (Document)): 文档指的是自然语言中的文本片段，一般一篇文档中至少会讨论一个话题。

定义 (话题 (Topic)): 本文中的话题概念比较广泛，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)): 情感极性值指的是评价观点积极（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)。 ■

对观点有不同的定义方法，例如 Liu[24] 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（或方面，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[25] 也对观点做了定义，认为观点由四个元素组成：即主题（Topic)、持有者（Ho1der)、陈述（Claim）以及情感（Sentiment)。无论对观点如何定义，一般都认为观点分析就是发现和确定各个元素的过程。总体来说，比较全面的观点分析可以认为是由三个主要步骤组成：

- **文本中观点信息识别**：需要识别文档中涉及的话题信息，将不同话题对应的文本片段按照话题对应联系起来，并且需要对文本片段进行主客观类找到主观性文本，因为观点一般是从主观性文本中确定的。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建模表示的情感词典。

- **文本情感极性分类**: 从文本中抽取有用的特征将文本分为不同的情感极性类别，一般是将文本分为积极或消极极性（或者中性，即客观文本），主要使用各种机器学习方法，或者使用基于规则的方法。

- **观点的整合集成**: 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，观点整合集成是在更高的层次上分析过程，是将文本碎片中分散的观点整合集成，并根据不同的应用需求以一种合理的方式表示，比如将文本片段按照时间顺序串联起来形成观点随时间的演化表示，或者本文研究的将同一用户所有文本片段中的观点集成起来用以对用户的主观性进行建模。

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；而观点分析主要是识别文档中表达的观点、喜好、立场和态度等主观信息，需要对文档进行词语语义、词性、甚至句法和篇章结构等深入分析。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情感极性，如 “举止轻薄”，而在电脑领域，“轻薄” 却表示褒义，具有积极情感极性。

2. **词性依赖性**: 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等方式发表对于关注话题的看法和观点，因此网络充满了各种各种意见，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信息消费能力，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，每个人观点分散在信息碎片海洋中，因此满足用户真正的信息需求（能针对目标主题快速准确地从网络中找到需要的观点）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[26–28]，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“查找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）。从信息产生者的角度来说，用户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，才能代表用户对于话题的完整观点。因此本文首先从下面一个科学问题出发，来研究观点分析：

科学问题 1：怎么样才能准确对社交媒体中用户层面的观点信息进行分析和表示？

这个问题需要从挖掘用户产生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，情感极性分类以及观点信息集成等问题。

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看法等主观性的影响。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分析有帮助。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，并分析其对用户在线交互行为的影响：

科学问题 2：用户的主观性如何表示和建模？

科学问题 3：怎么样使用主观模型分析用户的在线交互行为？

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方面观点的集成及表示问题。在使用社交媒

体过程中，用户作为带有自己主观判断的主体，会在不同情况下产生不同的交互行为，因此回答第二个问题需要首先确定用户在线交互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进行度量分析。

1.3 相关研究

本节主要介绍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, 29–31]，领域适应化（Domain adaptation）[30–33, 33–42]，预处理（preprocessing）[38, 43, 44] 以及进行一些结构化处理（词性、句法、标注等）[33, 41, 45–54]。本文在进行观点分析研究时，需要借鉴已有工作对社交媒体的非结构化噪声文本进行预处理，但是这些预处理方法不属于本文研究内容，因此不作详细介绍。

1.3.1 观点挖掘

观点挖掘主要研究识别文档中针对某一话题表达出的观点以及判断观点的极性（例如，是积极还是消极）[24]。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，因为已有研究表明将文档中的客观文本过滤掉后会有助于提高观点挖掘的准确性[55]，目前主客观分析主要采用机器学习方法进行主客观分类[56–61]。极性分类是将文档就话题表达出的情感极性进行分类，一般是分为积极与消极极性，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方法、基于特征共现统计方法以及基于机器学习方法等三类。

1.3.1.1 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[16]，DAL（Dictionary of Affect of Language）[17]，WordNet-Affect[18]以及 SentiWordNet[62]。基于情感词典的方法一般是用词典确定文本中的带有情感极性的词语，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[63–65]。一个句子或文档的情感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 1.1来表示：

$$SD = \frac{\sum_{w \in D} S_w * weight(w) * modifier(w)}{\sum weight(w)} \tag{1.1}$$

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函数。典型工作如 Zhu 等[66] 首先将文档或句子中词语的极性值累加在一起，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[67]，或 Linguistic Approach[68] 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。

1.3.1.2 基于特征共现统计的观点挖掘

这种方法是基于语料中表达相似观点的词语经常会在一起出现这一假设基础上的，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些极性恒定的词语（比如“good”）共现的频率来确定。Turney[69, 70] 提出使用点对点互信息（Point-wise mutual information (PMI)）[71] 作为统计依据来计算词语的共现：

$$PMI(x, y) = \log_2 \frac{F(x, y)}{F(x)F(y)} \tag{1.2}$$

其中 $F(x, y)$ 表示两个词语的共现频率，$F(x)$ 表示词语 x 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基准词语的互信息差值来确定：

$$PMI\_IR(x) = \sum_{p \in pWords} PMI(x, p) - \sum_{n \in nWords} PMI(x, n) \tag{1.3}$$

[16] http://www.wjh.harvard.edu/~inquirer/
[17] http://www.hdcus.com/
[18] 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[72] 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[73] 使用语义空间相似性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[74]，他们提出了使用搜索引擎确定共现频率可能会存在的一些问题。Ben 等[26] 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博客文档。

1.3.1.3 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：

$$g: D \to S, \quad g(d) = \arg max_S f(d,s) \tag{1.4}$$

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）在文档集中抽取有用特征，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集极性标签；（4）最后使用训练得到的分类器对新文档极性预测给出概率值。

Pang 和 Lee[75] 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极性分类，实验结果显示三种分类器性能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[58] 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[59] 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[76–79]，训练数据的构建[80, 81] 以及机器学习方法的选择[82]。

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 1.1所示产品评论的观点集成为例[83]，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[84–86]。

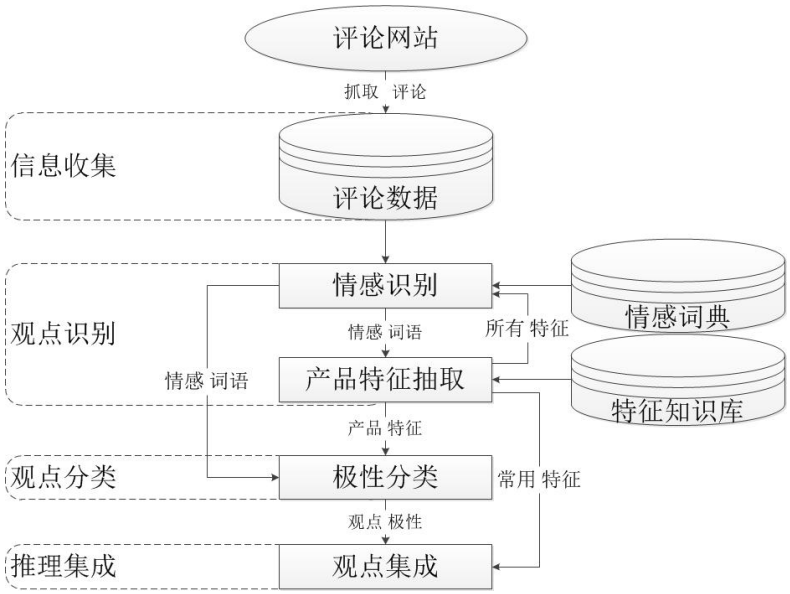


图 1.1 产品评论的观点集成框架

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 的综合观点。

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[87] 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[88] 使用监督学习方法从文本中抽取关键短语并将其聚类用以表示话题的方面。随着话题模型的引入，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[89, 90]，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[85, 91]。

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。

相关的研究有：Boyd 等[92] 通过实际数据研究了 Twitter 中转发类型和转发的原因，分析了不同用户群体，用户特性以及交流方式对于转发行为的影响，同时也从内容上分析了用户喜欢转发原因，发现微博文本中的 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[93] 研究了 Twitter 中用户之间提及（“@username”）关系网络，并分析了信息在提及网络上是如何传播的，发现大约有 25% 的微博是被作者的朋友转发，大部分是被粉丝而非朋友转发；Macskassy 和 Michelson[94] 分析了大量 Twitter 用户的近一个月的数据，解释了各种信息传播的方式，尤其是转发的行为模式，发现微博的内容是被转发的决定因素；Starbird 等[95] 对危机事件发生时 Twitter 上的信息传播进行了深入研究，具体分析了 2011 年埃及的政治事件，并演示了事件的相关信息在 Twitter 上是生成，发展，传播过程，对于信息传播的理解有帮助作用；Comarela 等[96] 研究了影响用户回复或转发的因素，发现先前行为，信息发布频率，信息的时效性以及信息长度决定了用户是否回复或转发；除了以上的工作，一些研究还从不同角度对 Twitter 中的转发行为机制进行了深入的研究[97–100]。

综上所述，影响用户转发行为的因素主要包括微博文内容、微博社交媒体特性（如是否包含链接、hashtag、提及等)、微博作者的用户特性以及社交关系等虽然已有的 Twitter 转发研究从许多不同的角度进行了分析，但是还没有工作从用户的主观动机角度进行研究，本文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，目前的转发行为分析大多都是从微博本身考虑，并未从微博接收者角度进行分析，本文将对微博、作者、接收者三个角色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挖掘观点信息，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挖掘集成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交互行为，本文主要分析用户在微博的转发行为。研究框架如图1.2所示。

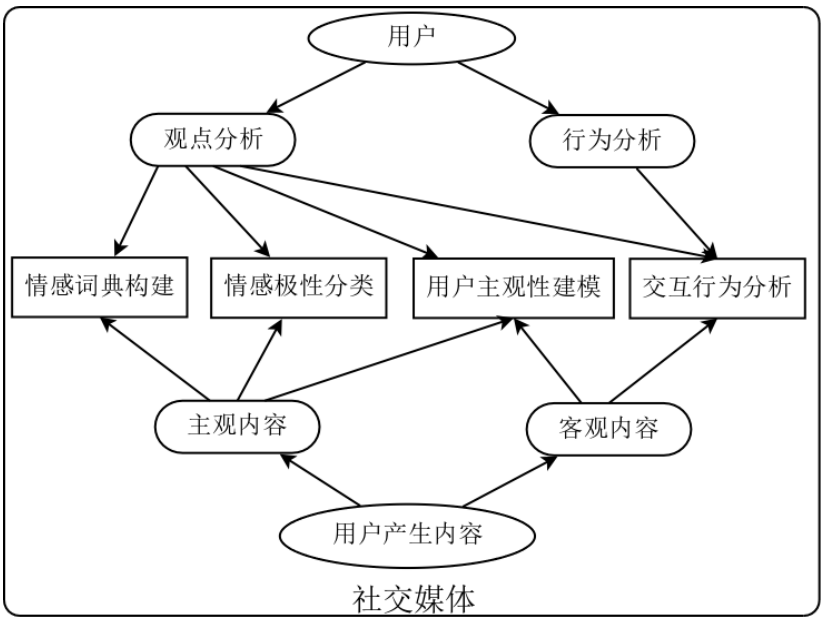


图 1.2 本文研究框架

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**情感极性分类研究**；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在线交互行为进行分析，属于**交互行为分析**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

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. **情感极性分类**：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. **交互行为分析**：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相结合，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，要面对以上问题和挑战，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一定领域独立能力，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的微博作为训练数据，是基于微博短小，观点表达会集中在一些极性相同的词语上这一假设，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。无监督或半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跨语言转换。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 1.3所示。，每个章节内容具体安排如下:

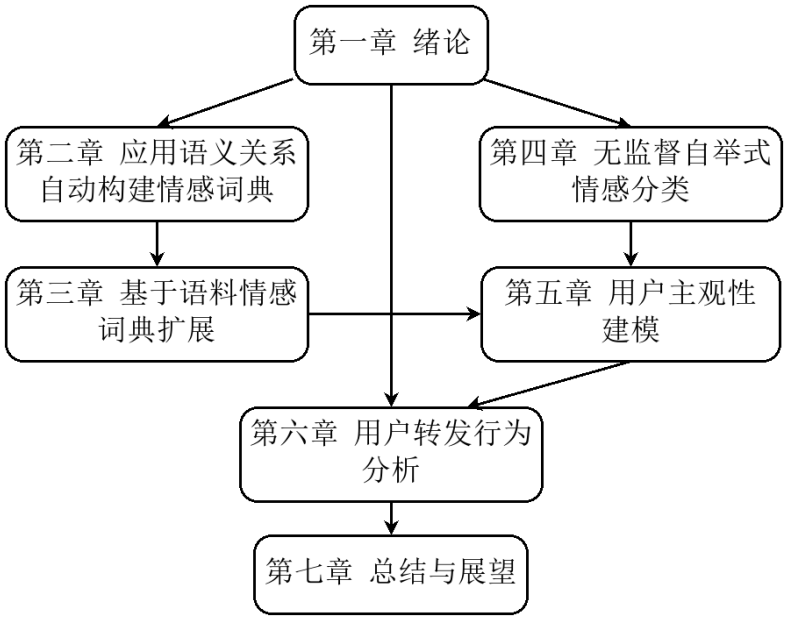


图 1.3　论文整体结构图

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，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

上一章主要介绍了本文的研究背景，要研究的科学问题，研究内容与方法，并指出观点分析研究一项基础的工作就是研究如何针对不同应用构建具有足够覆盖面并且良好适应性的情感知识词典。人在使用语言表达情感或观点时，最基本的方式是使用具有明确情感色彩的词汇，因此为了分析用户的观点，最直接的方法应该从用户产生文本中使用的词语开始，将语言中经常使用的词语所表达的情感信息进行汇总形成的词典就是情感词典。观点分析研究首先是在英文文本上开始的，情感词典相关研究也是从英文为主，方法相对比较成熟，形成了一些经常使用的英文情感词典资源。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[101-103]。目前研究使用主要有 HowNet 情感词典[104]，NTUSD 情感词典[105] 以及大连理工大学的情感词汇本体词库[106]。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中文词典进行适应性的转化，构建相应的中文情感词典资源，既可以省去耗费大量人力的人工标注过程，又可以克服目前中文情感词典自动或半自动构建方法的可靠性和覆盖度问题。因此本章提出将英文情感词典资源情感知识转化为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，可靠性更高。下一章将该方法构建的情感词典在领域语料中进行扩展，以提高其领域覆盖度和适应性。

本章具体安排如下：首先对情感词典构建需要考虑的问题以及相关工作进行全面的介绍，接着对本章要使用的词典资源进行介绍，然后详细阐述如何利用一个双语语义知识库将英文情感词典情感知识转化为中文情感词典相应的词语情感信息，最后对该方法进行实验验证和说明。

2.2　相关工作

构建情感词典需要考虑词典覆盖面（Coverage）、词典内容（Content）以及构建方法（Acquisition）三个方面问题，具体内容可以用图 2.1框架来展示。

2.2.1　词典覆盖面

就词典的覆盖面来讲，情感词典可以分为通用的词典以及领域专用的词典。构建通用情感词典的主要假设就是希望词语表达的情感独立于具体的领域和应用，

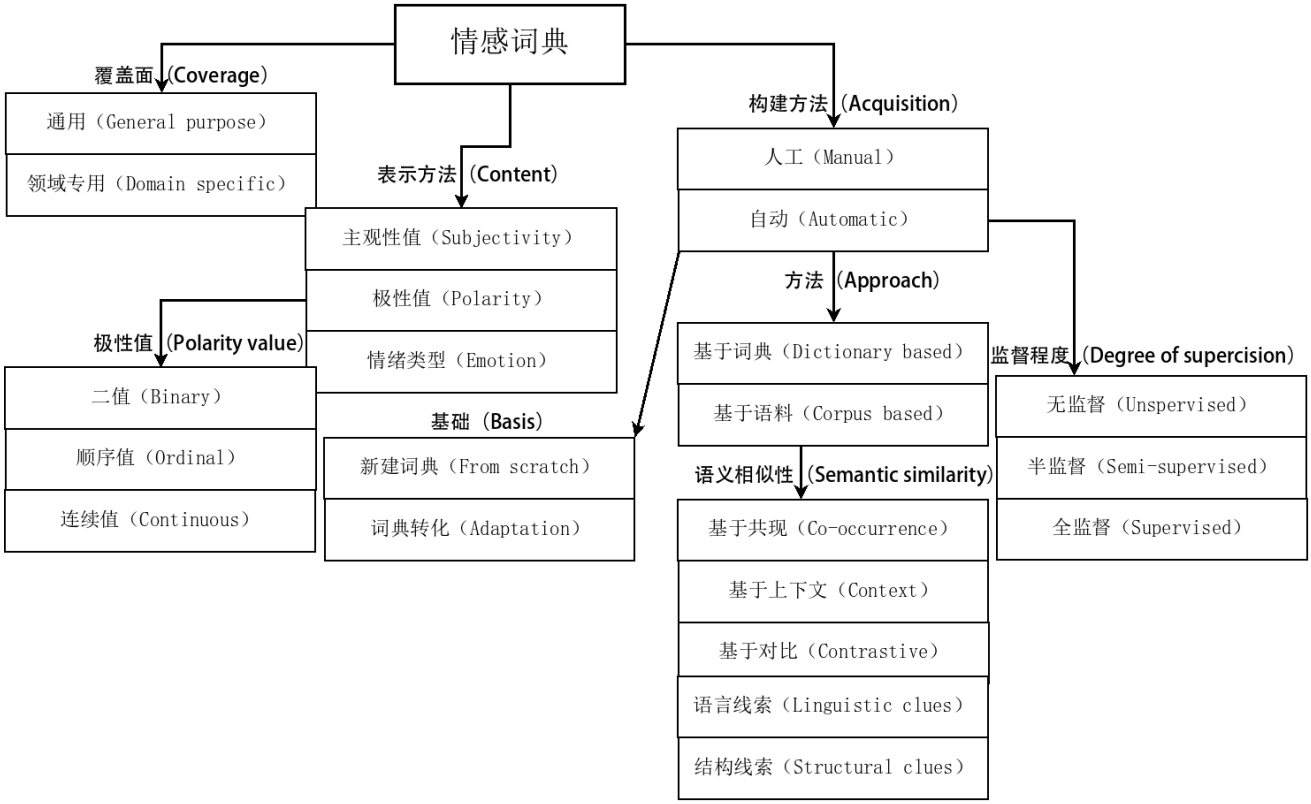


图 2.1　情感词典相关研究

这对一部分词语 (比如:“赞”、“爱”、“憎恨”等), 目前经常使用的情感词典基本都是通用的情感词典。但是实际上很多词语表达的情感是依赖于领域和具体的语境的, 比如“轻薄”一词, 一般情况表示负面情感, 但是描述手机时却可以表示正面的评价。而且, 一些本身不表示具体情感的词语 (比如:“长”、“短”、“老”、“经典”等), 用到一些特殊的语境时也会表达出一些具体情感 (比如:“手机待机时间长”与“手机待机时间短”)。最近开始有一些研究开始针对具体领域和应用需求构建一些领域专用的情感词典[107–109]。

2.2.2　词典内容

情感词典可以就其表示情感知识内容进行细分, 当然情感知识表示方法与具体的应用场景是密切相关的, 我们不考虑具体的应用场景, 而仅仅对词典的情感表示方法进行划分。从这个方面来看, 情感词典表示情感知识的方法可以分为三种: 词语表示主观性的程度 (Degree of subjectivity)、表示情感的极性 (Polarity) 或者表示的情绪类型 (Emotionmood, 比如喜、怒、哀、乐等)。表示主观性程度的情感词典主要用于文本主观性的探测任务[110–114], 而主观文本表达观点的具体情感类型的识别, 需要表示情感极性或情绪类型的情感词典。在情感分类中经常使用是表示情感极性的词典, 其中的词语都标注了表达的情感极性是积极的还

是消极的，这对于想要确定文本所表达观点的倾向性是非常重要的。对这类情感词典还可以根据表示的情感极性值大小进一步划分：比如使用二值情感极性（也就是积极的和消极的）的情感词典[115–118]；使用顺序值（ordinal）表示极性值的情感词典，比如常使用 1-5 整数值区分情感极性的强度值[119]；还有一些使用连续数值（continuous）表示情感极性强度的情感词典[70, 120–122]。目前有很多这样的英文情感词典，比如：OpinionFinder（OF）[123]，Appraisal Lexicon（AL）[124]，SentimentWordNet[62] 以及 Q-WordNet[125] 等。

如果为了分析更细粒度（fine-grained）的情感，需要将词语表达的情感根据情绪类型进行表示，比如 Bollen 等[22] 通过分析 Twitter 中大众表达出的不同情绪来预测股票指数的变化，Garcia 和 Schweitzer[126] 对产品评论中的情绪类型进行了细致研究，类似的工作还有 Davidov 等[127] 以及 Strapparava 和 Mihalcea[128] 在 Twitter 上的工作。这种类型的情感词典都是靠人工编辑形成的，比如 GI（General Inquirer）[1][129]，ANEW(Affective Norms for English Words)upciteBradley1999，WordNet-Affect[2]upciteValitutti2004,Valitutti2004a，DAL (Dictionary of Affect in Language)upciteWhissell1989 等词典。

2.2.3 词典构建方法

从情感词典的构建方法来看，可以分为人工构建和自动构建两种类型。目前公开可用的人工编辑的情感词典基本都是通用的情感词典（比如：OF 词典和 GI 词典），人工构建情感词典主要面临的问题除了需要耗费大量的人力，还有覆盖面相对较低，以及需要对不同的领域进行适应性扩展才能达到好的观点分析效果。

而对于自动构建情感词典方法，还可以按照方法（Approach）、监督程度（Degree of supervision）和构建基础（Basis）三个维度进行区分。

2.2.3.1 构建方法

情感词典主要的构建方法分为两类：一是基于词典（dictionary-based）方法，根据已有词典的词语之间的语义关系判断词语的情感极性或计算情感极性值；二是基于语料（corpus-based）方法，根据词语在语料中的分布特点推导出情感极性并计算极性值。两类方法共同特点是都需要一个预先标注的种子词集（seed set），然后通过不断迭代计算词语与种子词集词语之间的某种语义相似性，推导词语情感极性值并扩充种子词集，直到收敛。

基于词典方法： 基于词典的方法通常会使用一个词库（thesaurus）或语义知识库(比如常用的是 WordNet[130])，并且常用的假设是词语间的语义关系转换词语的情

[1] http://www.wjh.harvard.edu/~inquirer/Home.html
[2] http://wndomains.fbk.eu/wnaffect.html

感信息，最常用的语义关系是词语间的同义和反义关系[115, 131, 132]。例如形容词“lovely”会将积极极性通过同义关系传递给“admirable”、“adorable”’、“amiable”和“pretty”，反过来会将消极极性转换给反义词语“awful”、“unlovely”和“ugly”。但是这种转换会随着语义距离增加而弱化，比如在 WordNet 中从“good”到“bad”的同义关系距离长度只有 3[115]，因此方法设计时需要采取适当措施将语义距离考虑在内[115, 131–134]。除了同义和反义关系，一些研究提出使用 WordNet 中的其他语义关系，比如“similarity”，“derived-from”，“pertains-to”，“also-see”或“attribute”等关系[135, 136]。Takamura 等[137] 以及 Andreevskaia 等[138] 使用了并不直观的下位关系（hyponymy）构建情感词典。还有一些方法通过计算词语在词典中解释的相似性来度量词语间的语义相关性，然后根据这种语义相关性构建情感词典[62, 135, 139]。

基于语料方法： 和基于词典方法一样，基于语料方法一个基本思想就是通过某种方法度量词语间的语义相关性，然后从标注好的种子集中推断出词语的情感信息。这些度量方法可以分为以下四种：

- **基于词语共现方法**：代表性的工作是 Turney 等[69, 70]，主要是假设“一个词语的语义倾向性（semantic orientation）[3]往往与其相邻的词语的语义倾向性相关”，因此他们使用点互信息（pointwise mutual information）PMI 统计对词语和种子词集的相关性进行度量，推导出词语的情感极性。

- **基于上下文方法**：除了直接通过共现来度量两个词语的相关性，在统计语义学还有还有一种常用的方法就是使用词语的上下文信息。在 Firth 的《Contextual Theory of Meaning》一书中，提出一个基本的假设就是“a word is characterized by the company it keeps”[140]，因此词语的语义信息是与上下文语境紧密相关的。因此一些基于语料的情感词典构建方法利用这一假设，提出在相似上下文出现的词语很有可能具有相似的情感信息，因此可以从情感极性已知的种子词集推导出其他词语的情感极性[113, 121, 141]。

- **基于对比方法**：该类方法将前台（foreground）语料和背景（background）语料对比分析进行情感词语的抽取构建情感词典。比如 Maks 和 Vossen[112] 研究了对数似然和相对频率比提取主观词构建语情感词典，他们使用报纸新闻以及新闻评论作为主观前台语料，维基百科文本作为客观背景语料。相似的工作还有 Stepinski 和 Mittal[142]。

- **基于语言线索**：前面几种方法单纯依靠语料中统计出的信息，不考虑对文本的深层次语言学分析。其实已经有工作确认一些常用的语言模式有助于词语

[3]Turney 使用语义倾向性指代词语的情感极性

的情感信息的探测。Hatzivassiloglou 和 McKeown[116] 发现一个句子中连词（“and”和“but”）对与所连接的两个词语的情感极性具有一定的限制作用，出现在“and”两边的词语一般具有相同的极性，而出现在“but”两边的词语极一般性相反，他们利用这种限制从文本语料中抽取并构建情感词典。在产品评论的观点挖掘研究中一些工作扩展了这种连词语言线索，同时考虑了跨句子的连词[63, 91, 143, 144]。

- **基于结构线索**：代表性的工作是 Kaji 和 Kitsuregawa[145, 146] 的工作，他们利用 HTML 文档中的结构线索分别抽取情感极性为积极和消极的句子集，从大量 HTML 文档中抽取出大概 500,000 主观句子用于训练情感分类器并构建情感词典。

2.2.4 词典转化

上述的所有情感词典构建方法都是从头开始（from scratch）构建新的情感词典，最近也有一些方法研究已有情感词典进行转化，主要是增强通用情感词典的领域适应性或者从单语言情感词典扩展到多语言。通用情感词典进行领域转化方法，主要有 Choi 和 Cardie[107] 提出的基于线性规划方法，Du 等[108] 提出的基于信息理论框架以及 Qiu 等[147] 使用语言模式的扩展方法等。Mihalcea 等[148] 提出了基于词典和基于语料的方法将英文情感词典通过翻译转化为其他语言的情感词典。

2.2.5 混合方法

混合方法指的是构建情感词典时将多种词典和语料资源结合起来。例如 Hoang 等[149] 提出使用 WordNet 的语义关系产生初始情感词典，然后使用从网络语料中获取的统计信息对其进行完善，词典资源和语料资源用一个错误最小化算法（error minimization algorithm）结合起来。Lu 等[150] 提出将四种信息组合起来确定词语的情感极性，包括从一个通用情感词典得到的信息，从一个词库（thesaurus）得到的信息，以及从一个领域文档集中得到的语言线索和结构线索信息，这四种信息通过一个基于线性规划的优化框架结合起来确定词语的情感极性。

综上所述，目前有很多种构建情感词典方法，以针对英文资源的构建方法研究为主。中文情感词典的构建方法研究还相对较少，而且基本上是借鉴英文的构建方法，而且形成的中文情感词典表示的情感知识是简单的二值极性。本章主要研究如何从英文词典进行转化得到中文情感词典，属于词典转化方法，但是我们的实现方法是借助于双语语义知识库中的语义关系实现这种转化，而不是使用翻译的方式，而且我们形成的情感词典能够通过计算将英文情感词典的情感极性值同时转化过来，情感知识更加丰富。

2.3 词典资源简介

本节简要介绍要用到的一些词典和知识库资源，主要是针对本章研究相关部分作介绍。

2.3.1 HowNet 语义知识库

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[151]。HowNet 两个重要名词是“义原”和“概念”：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[152]；义原是最小语义单元，用于定义和描述概念。

2.3.1.1 义原

HowNet 设计了大概 2200 多个义原，这些义原分为几个大类，具体参见表 2.1。

表 2.1 HowNet 义原分类

义原	数量	示例	语义倾向性[1]
Event| 事件	819	blame| 埋怨	一般有倾向性
Entity| 实体	142	human| 人	不具倾向性
Attribute| 属性	117	length| 长度	一般不具倾向性
aValue| 属性值	899	good| 好	一般有倾向性
Quantity| 数量	3	rate| 比率	一般不具倾向性
qValue| 数量值	13	ufficient| 足	一般有倾向性
SecondaryFeature| 次要特征	100	desired| 良	一般有倾向性
Semanticroles| 语义角色	90	StateFin| 终状态	一般不具倾向性[2]

[1] 语义倾向性即情感极性。

[2] 虽然语义角色类不具有倾向性，但是代表的语义关系可以影响其他义原的倾向性。

表中可以看出，除了事物类、属性类以及数量类义原，其他义原一般都具有情感极性，并且义原都是由中英双语标识，因此可以通过英文标识从英文情感词典中获得其情感极性值。但是有一部分义原的英文标识不是一个单词（比如：FondOf| 喜欢，WhileAway| 消闲等），无法直接从英文情感词典直接获得情感极性值。实际上义原之间并不是独立的，义原之间存在复杂的关系，HowNet 中描述了义原之间的主要的 8 种关系：上下位关系、同义关系、反义关系、对义关系、属性 -宿主关系、部件 -整体关系、材料 -成品关系、事件 -角色关系。义原之间组成的是一个复杂的网状结构，而不是一个简单树状结构。不过，义原关系中最重要

关系是上下位关系，义原根据上下位关系形成了如图2.2树状层次体系，我们可以借助其对无法直接转化得到情感值那部分义原的情感极性值的计算。

图 2.2　HowNet 义原层次结构

除了义原以外，HowNet 还有一些符号（或称为符号义原）对概念语义描述，可以把这些符号归为几类：第一类包括“,”（表示“和”的关系）、“~”（表示“或”的关系）、“∧”（表示“非”的关系），用来表示语义描述式之间的逻辑关系；第二类包括“#,%,$,&,*,+,?,!,@”，表示概念之间以及概念的属性之间的关系；第三类包括几个无法归入以上两类的特殊符号“{},(),[]”。这些符号义原中第一类描述逻辑关系的三个符号会引起所描述义原情感极性的变化，尤其是“∧”会引起情感极性的反转。

2.3.1.2　概念

如图 2.3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。

```
W_C=一蹶不振
G_C=V
E_C=
W_E=collapse after one setback
G_E=V
E_E=
DEF={decline|衰败:SincePoint={fail|失败},scope={Circumstances|境况}}
```

图 2.3　HowNet 中概念描述方式

DEF 是 HowNet 对于概念的定义，称为语义表达式，是知网的核心。HowNet 知识描述语言是比较复杂的，为了后续分析计算，我们归纳为以下几条：

1. HowNet 收录词语主有两类，即实词和虚词；

2. 虚词描述比较简单，用“{句法义原}”或“{关系义原}”进行描述，虚词不具情感极性;

3. 实词的描述就比较复杂了，由一系列用逗号隔开的语义描述式组成，其中语义描述式分为三种形式：

   (a) 独立义原描述式：用“基本义原”，或“(具体词)”描述;

   (b) 关系义原描述式：用“关系义原 = 基本义原”，或“关系义原 = (具体词)”，或“(关系义原 = 具体词)”描述;

   (c) 符号义原描述式：用“关系符号基本义原”或者“关系符号（具体词)”加以描述;

4. 实词描述代表了该词的语义知识，因为实词一般具有语义倾向性（如果将中性也视为倾向性的话)，因此实词的描述式可以帮助我们确定语义倾向性。

2.3.2 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[153]。WordNet 是根据词义而不是词形来组织词汇信息。如图 2.4所示，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。

```
Help:
  Noun
    S: (n) aid, assist, assistance, help
    S: (n) assistant, helper, help, supporter
    S: (n) avail, help, service
  Verb
    S: (v) help, assist, aid
    S: (v) help, facilitate
    S: (v) help oneself, help
    S: (v) serve, help
    S: (v) avail, help
```

图 2.4 WordNet 单词描述方式

2.3.3 SentimentWordNet 情感词典

SentimentWordNet 是 Baccianella[62] 等在语义词典 WordNet 基础上使用随机游走的图算法计算得到的情感词典。如图 2.5所示，SentimentWordNet 的每条记录都是一个 WordNet 的 Synset 条目，并且每个 Synset 都计算出了褒义、贬义情感极性的强度值（简称情感极性值），本章就是利用 SentimentWordNet 的情感极性值以及 HowNet 概念的语义关系进行计算得到中文词语的情感极性值，实现从英文情感词典到中文情感词典的情感知识转化。SentimentWordNet 共有收录了 117,000 多个 Synsets，约 192,493 单词。

```
healthy:
   P: 0.75 O: 0.25 N: 0
   healthy#101170243

   P: 0.5 O: 0.5 N: 0
   healthy#202273838

   P: 0.875 O: 0.125 N: 0
   salubrious#1  healthy#3  good_for_you#102558184

   P: 0.75 O: 0.25 N: 0
   sound#2  levelheaded#1  level-headed#1  intelligent#3  healthy#401944088

   P: 0 O: 1 N: 0
   tidy#3  sizeable#2  sizable#2  respectable#3  hefty#3  healthy#5  goodly#1  goodish#200624576
```

图 2.5 SentimentWordNet 情感词描述方式

2.4 基于语义关系的情感词典构建方法

如本章相关工作部分所述，要构建一个新的情感词典有两种方式，一是从头开始（from scratch），另外一种就是通过转化（adaption）其他词典资源的方式。中文观点分析的研究在最近几年才开始受到重视，并且主要是借鉴英文研究已有的资源和方法。中文和英文语法结构和语义表示上存在很大的差别，直接套用英文的资源和研究方法会出现“水土不服”，比如直接将英文情感词典通过翻译方式转化的中文情感词典，是一个从英文情感知识到中文情感知识“给”的方式转化，是将英文情感词典内容映射到中文情感词典，因此存在歧义较大，覆盖度较低以及可靠性不高等问题，并且词典中不可避免存在翻译带来的错误。本章我们提出一个从中文到英文情感词典去“取”的方式转化情感知识，是从中文情感词典内容到英文情感词典的逆映射，因此可以根据中文词语的语义单元选择英文对应语

义单元然后转化情感知识，有效避免了歧义，而且不受覆盖度的限制，可靠性也更高。同时可以直接将英文中对情感极性值的计算结果直接转化为中文词语的情感极性值，减少了计算开销。本章研究正是基于这种动机展开的，提出的解决方案如图 2.6框架所示。

具体来说，我们使用了双语语义知识库 HowNet 作为我们中文词语的来源以及对应英文查询词语的来源。HowNet 对义原和概念（大部分都有英文标注）进行了英汉双语标注，可以作为中英文情感知识转化的“桥梁”。我们的计算框架中，每个词语的情感极性值的计算都是由三部分组成，首先是词语对应的英文标注可以从英文情感词典中查询获得情感极性值，第二部分是词语的语义描述 DEF 中会有义原的英文标注，也可以查询得到情感极性值，第三部分通过对语义描述 DEF 的语义关系分析，按照义原在 DEF 中的语义角色对其情感极性值加权后与第一部分进行组合计算词语的最终情感极性值。

HowNet 中词语本身有些没有英文标注，无法通过查询英文情感词典获得情感极性值。有些词语虽然有英文的标注，但是查询英文情感词典时候会遇到一词多义问题，不同语义对应的情感极性值不尽相同，得到的情感极性值也会因为存在歧义而不准确。还有一些词语标注的英文是多个单词，无法直接得到情感极性值。HowNet 中词语的语义由概念表示，每个概念都有对应的语义描述 DEF，DEF 是由一个到多个义原按语义关系组合在一起的，因此概念在中文中是没有歧义的。我们正是利用了 HowNet 中概念语义描述部分，分析其中义原以及义原之间的关系，将义原情感极性值组合计算出词语的情感极性值。HowNet 中义原是描述语义最基本的单位，因此我们假设义原的情感极性是确定的，因此通过语义关系组合计算出的情感极性值也应该是确定的，因此可以词语的情感极性进行“消歧”。

综上所述，本章构建情感词典需要解决的问题描述如下：

1. 如何基于英文情感词典计算义原的情感极性值？

2. 如何通过概念描述中的语义关系分析组合计算概念的情感极性值？

3. 最后如何确定词语的情感极性值？

针对这的三个问题，我们通过情感词语及义原抽取、义原情感极性值计算、以及词语情感极性值计算三个部分进行阐述。

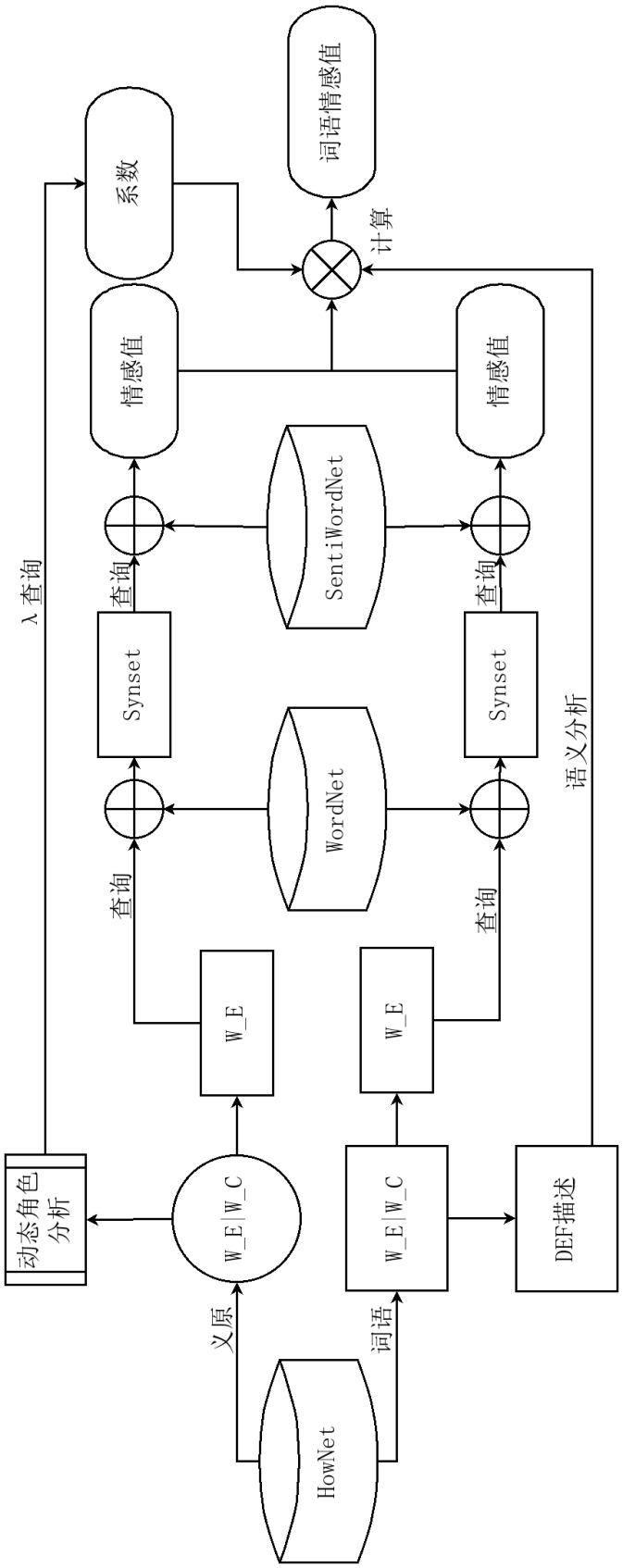


图 2.6　基于语义关系情感词典构建方案

2.4.1 词语和义原抽取

词语抽取主要是从 HowNet 中抽取词语（W_C）和概念描述（DEF），并对 DEF 进行分析得出其组成义原及语义关系描述符。在进行词语情感极性值计算时，需要根据 DEF 中义原和语义关系描述符进行词语的语义分析和极性值计算。情感词语和义原抽取处理流程如图 2.7所示。

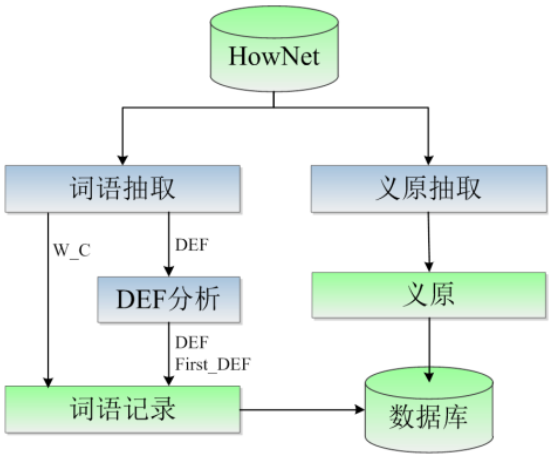


图 2.7　词语和义原抽取处理流程

从 HowNet 中抽取出的词语，定义其记录格式如图 2.8所示。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。

```
NO.=035390
W_C=好
G_C=ADJ
W_E=kind
G_E=ADJ
DEF=aValue|属性值,behavior|举止,kindhearted|善,desired|良
First_DEF=aValue|属性值
PosScore=
NegScore=
```

图 2.8　抽取词语记录格式

从 HowNet 中抽取得到的义原的记录格式如图 2.9所示。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层

次关系，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)”。

```
NO.=33
Category=Event
W_C=依靠
W_E=depend
DEF=depend|依靠
Layer=5
Father=32
PosScore=
NegScore=
```

图 2.9　抽取义原记录格式

2.4.2　义原情感极性值

这三个部分中义原的情感极性值的确定是非常关键的，是计算词语情感极性值的基础。在 HowNet 中义原都使用中英双语标注，基本都可以根据英文标注查询得到情感极性值（称为查询类义原极性值)。也有一部分义原英文标注由多个单词连接组成（如“FreeOfCharge| 免费”)，无法直接查询得到情感极性值，可以通过在上下位关系树中与其他义原的语义距离进行计算获得（称为计算类义原极性值)。最后可以通过义原间的反义和对义关系对计算出的义原情感极性值进行校正。

2.4.2.1　查询类义原极性值

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 SentimentWordNet 就可以得到对应的情感极性值。基于 WordNet 和 SentimentWordNet 的义原极性值计算过程如图 2.10所示。

在 HowNet 中获取义原后将义原对应英文单词（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 再映射到 SentimentWordNet 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极性值（如“good”的极性值为 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 i 的 Sense 中的 p 类型极性值。

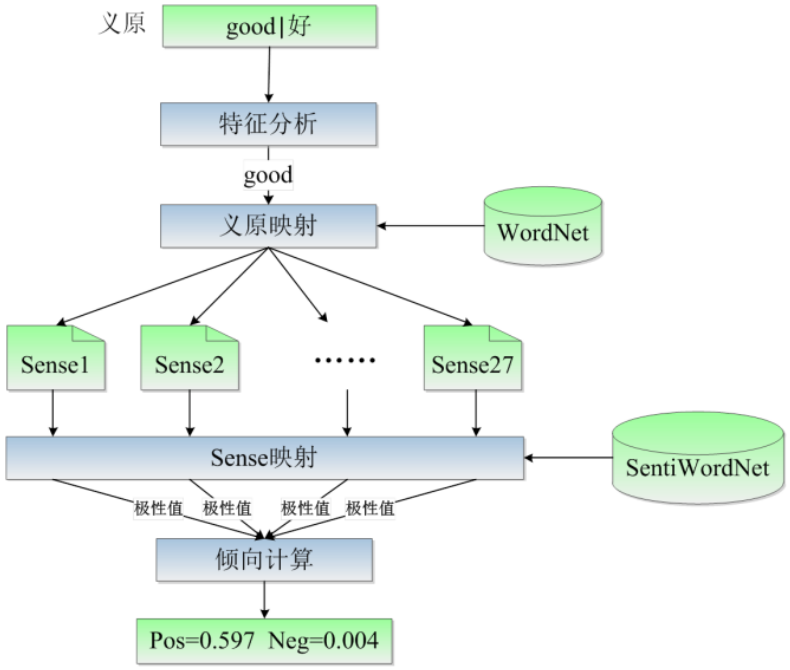


图 2.10　义原情感极性值计算过程

2.4.2.2　计算类义原极性值

经过上面的查询计算过程，可以得到大部分义原的情感极性值。由于所有的义原根据上下位关系构成了一个树状的义原层次体系，针对一些无法查询计算得到的义原，我们采用简单的通过语义距离计算相似度的办法间接计算出情感极性值。假设两个义原（一个情感极性值已知，一个未知）在层次体系中的路径距离为 d，根据公式 2.2，我们可以得到这两个义原之间的语义距离：

$$sim(s_i, s) = \frac{\alpha}{d + \alpha} \tag{2.2}$$

其中 s_i 是情感极性值已知义原，s 表示需要情感极性值计算的义原,d 是 s_i 和 s 在义原层次体系树中的路径长度。α 是一个可调节的参数，一般 $\alpha = 0.5$。

为了能够在上位和下位义原的情感极性值取得平衡，对任意一个情感极性值未知义原 s，都要计算 s 与最靠近 s 极性值已知的上位义原 s_1 和下位义原 s_2 之间的语义距离：$sim(s_1, s)$ 和 $sim(s_2, s)$。然后从上位义原 s_1 和下位义原 s_2 的情感极性值加权平均得到 s 的情感极性值。

$$\varphi(s, p) = sim(s_1, s)\varphi(s_1, p) + sim(s_2, s)\varphi(s_2, p) \tag{2.3}$$

2.4.2.3　情感极性值校正

所有义原在前面两种方法计算得到情感极性值会存在一些偏差（bias），有些义原偏差会比较大，甚至计算得到的极性值与义原的真实语义倾向相反（如“FreeOfCharge|免费”义原计算得到极性值为 PosScore=0.07，NegScore=0.236），因

此需要通过利用 HowNet 中的其他语义关系对上述计算方法进行校正。我们采用了基于 HowNet 中对义和反义语义关系进行义原情感极性值校正。对于任一义原 s，对义或反义义原为 $\overline{s}$，对 s 情感极性值修正为：

$$\varphi(s,p)=\frac{|\varphi(s,p)-\varphi(\overline{s},p)|}{2} \tag{2.4}$$

2.4.3 词语情感极性值

词语情感值可以通过两种途径获得，一是通过词语本身的英文标注直接查询英文情感词典，这种方式并不可靠而且存有歧义；二是根据词语的语义描述 DEF 中的义原的情感极性值计算得出，这种方式相对可靠，每个义原都有确定的情感极性值，因而不存在歧义。为了计算词语的情感极性值，需要对语义描述 DEF 中的语义关系进行分析，因为义原间的语义关系会引起义原情感值的反转以及在描述语义倾向时的权重变化。因此首先对 HowNet 中因为 DEF 的语义关系不同引起的情感极性值变化提出如下定义：

定义 (情感极性值反转)：义原 s 的 p 极性值 $\varphi(s,p)$ 取反运算是，将 s 的积极极性值和消极极性值互换，过程如公式 2.5：

$$\overline{\varphi(s,p)}=\varphi(s,q),\quad (p,q)\in P\&\&p\neq q \tag{2.5}$$

■

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”等相当于句子中的否定词，会引起其他词语情感极性值反转，因此我们从 819 个事件类义原中挑选了在语义描述中起否定作用的义原称为极性反转语义角色，并加以标记。

定义 (情感极性值加权)：λ 因子与义原 s 的 p 极性值的加权运算定义为 λ 乘法运算，过程如公式 2.6：

$$\lambda\times\varphi(s,p)=\begin{cases}\lambda\varphi(s,p), & \lambda>0\\ 0, & \lambda=0\\ |\lambda|\varphi(s,p), & \lambda<0\end{cases} \tag{2.6}$$

■

公式 2.6中 λ 取值范围为 $-1,0,1$，具体值需要根据关系义原描述式中的关系义原（动态角色义原）和符号义原描述式中的符号义原确定。符号义原中只有“∧”（表示“非”的关系）会改变语义倾向，因此“∧”所修饰的义原在计算中权重为 $\lambda=-1$。HowNet 中共有 90 个动态角色义原，我们分别对每个义原进行了分

析，确认了其语义关系角色所确定的基本义原的权重取值 λ。如词语“扭亏为盈”的 DEF 描述为“DEF=alter| 改变，StateIni=InDebt| 亏损，StateFin=earn| 赚”，义原“InDebte| 亏损”为初始状态（StateIni），“earn| 赚”为最终状态（StateFin），经过分析后，StateIni 描述的“InDebte| 亏损”的 λ 取值为 0，StateFin 描述的“earn| 赚”的 λ 取值为 1。

最后词语的情感极性值计算总结为公式 2.7。其中 $\varphi(w,p)$ 表示词语 w 的 p 极性值，s_i 表示词语 DEF 中第 i 个义原，n 为词语 DEF 中义原总数。

$$\varphi(w,p)=\frac{\sum_{i=1}^{n}\lambda_i\times\varphi(s_i,p)}{\sum_{p\in P}\sum_{i=1}^{n}\lambda_i\times\varphi(s_i,p)} \tag{2.7}$$

其中：$\sum_{p\in P}\varphi(w,p)=1$。

对于已经通过查询得到情感极性值的词语 (有多个英文 Sense 对应的词语的情感极性值 $\varphi(\overline{w},p)$ 可以取所有 Sense 对应极性值的加和平均)，可以和通过语义描述 DEF 计算得到的极性值加权累加，计算公式为：

$$\Psi(w,p)=\alpha\varphi(w,p)+(1-\alpha)\varphi(\overline{w},p) \tag{2.8}$$

其中 $\alpha\in(0,1)$ 的取值要考虑那种方式得到的情感极性值更准确，一般将 α 取偏大些以反映语义描述 DEF 对词语语义倾向性的影响更大。

2.5 实验

情感词典的实验评测有两种方法，一是直接评测，将情感词典与人工编辑的或者其他可靠性较高的词典进行对比评测；二是间接评测，将情感词典应用到文本情感分类任务评测其性能。本节我们使用直接评测的方法。

2.5.1 直接评测

在实验评测时，采用 HowNet 评价词词典基准。HowNet 评价词词典是 2007 年发布的人工标注的词典，每个词语只标注积极和消极极性，没有极性值。HowNet 评价词词典中情感词共有 6497 项，其中积极极性词语 3436 项，消极极性词语 3061 项。

我们将本章中利用 HowNet 中的语义关系构建的情感词典命名为 Senti-HowNet。SentiHowNet 的词语都有积极极性值和消极极性值，为了和 HowNet 评价词词典作对比评测，按照公式 2.9对 SentiHowNet 的每个词计算得出一个极性。

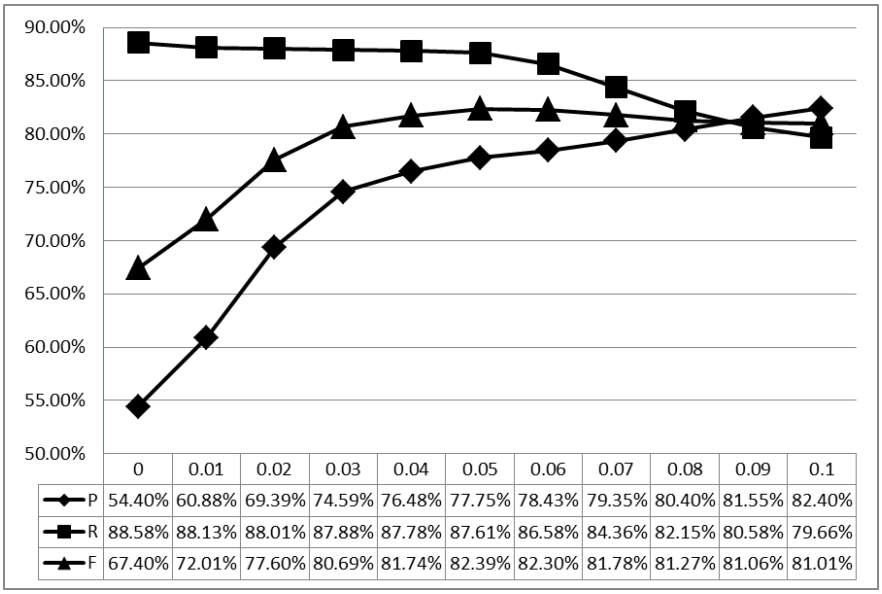
$$\begin{cases} positive & score(w)>T \\ negative & score(w)<-T \end{cases} \tag{2.9}$$

其中 $score(w) = \varphi(w, P) - \varphi(w, N)$ 为词语 w 积极极性值与消极极性值相减得到的差值，T 为阈值，差值 $score(w)$ 高于 T 则词语 w 情感极性为积极极性，差值 $score(w)$ 低于 $-T$ 则词语 w 情感极性为消极极性。当 $T = 0$ 时，我们构建的 SentiHowNet 有积极极性情感词 12433 项，消极极性情感词 11148，词典的整体规模达到 23581 个条目。

评价指标采用常用的精准率（P）、召回率（R）以及 F 值。

2.5.1.1 阈值 T 的设定

首先考察不同的阈值 T 对词典评价指标的影响，以确定一个合理的阈值。图2.11为 T 的不同取值对词典性能指标的影响，实验中我们将 T 的值由 0 逐渐增加，当 T 大于 0.1 时性能指标基本没有变化或成下降趋势。从图中可以看出，在 $T = 0$ 时，虽然召回率最高达到 88.58%，但精准率最低仅有 54.40%，F 值仅为 67.40%。当 $T = 0.05$ 时，精准率提高到 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 2.11　不同 T 值时的性能指标

2.5.1.2 与其他词典性能对比

确定了阈值 T 以后，我们将 SentiHowNet 与目前常用的中文情感词典 NTUSD 词典（11086 个中文词汇，2810 积极极性词语，8276 消极性词语）[105] 以及大连理工大学的情感词汇本体词库（用 DLLEX 标记，17156 条目，10627 个积极极性词语，6529 个消极极性词语）[154] 进行对比评测，首先是三个词典对评价基准词典的覆盖度对比，结果如表 2.2所示。从表中可以看出，按照积极和消极极性分

来来看，在积极极性词中 SentiHowNet 词典覆盖度最好，NTUSD 词典由于包含的积极极性词数量少因此覆盖读最低；在消极极性词中，NTUSD 词典的覆盖度最好，SentiHownet 和 DLLEX 词典覆盖度基本一样。四种个词典中除了作为基准词典的 HowNet 评价词词典，采用我们设计方法在 HowNet 上自动构建的情感词典 SentiHowNet 包含情感词条目（注意经过阈值 T=0.05 过滤后）最少，但是对基准词典的覆盖度最高（总体准确标注数达到 6092），主要原因是 SentiHowNet 本身就是从 HowNet 自动产生，是 HowNet 包含词语的子集，而 NTUSD 和 DLLEX 词典中词语的来源不同，因此会有覆盖度的偏置。

表 2.2　词典覆盖度

词典	积极极性		消极极性		总体统计	
	标注	正确	标注	正确	标注	准确
HowNet	3436		3061		6497	
NTUSD	2810	2204	8276	3022	11086	5226
DLLEX	10627	3020	6529	2876	17156	5896
SentiHownet	4256	3218	5113	2874	9369	6092

在 T=0.05 时，SentiHowNet 与其他词典性能比较如表 2.3所示。在积极极性词语的性能对比中，SentiHowNet 词典的精确率和 F 值最好，分别达到了 93.66% 和 83.67%，但是召回率（75.%61）比 NTUSD 词典（78.43%）略差；在消极极性词语的性能对比中，三个词典的精确率都比较高（都超过 90%），但是召回率比较低，NTUSD 词典精确率最好，达到 98.73%，而 SentiHowNet 词典在召回率和 F 值上性能最好；总体来看，宏平均指标中 SentiHowNet 词典精准率为 93.77%，召回率为 65.02%，F 值为 76.79%，均为三个词典中最高。

表 2.3　词典性能对比

词典	积极极性			消极极性			宏平均		
	P	R	F	P	R	F	P	R	F
NTUSD	0.6414	**0.7843**	0.7057	**0.9873**	0.3652	0.5331	0.8044	0.4714	0.5944
DLLEX	0.8789	0.2842	0.4295	0.9396	0.4405	0.5998	0.9075	0.3437	0.4985
SentiHownet	**0.9366**	0.7561	**0.8367**	0.9389	**0.5621**	**0.7032**	**0.9377**	**0.6502**	**0.7679**

以上两个对比表明相对于需要人工干预的人工或半自动方法构建的情感词典，使用我们设计的自动构建情感词典方法可以构建一部覆盖度比较好，性能又可靠的情感词典，同时节省了不必要的人力开销。

2.6 小结

本章对中文情感词典构建相关研究进行了分类和详细阐述，基于目前中文情感词典研究现状，提出了一种新的词典跨语言自动转化方法。该方法以双语语义知识库 HowNet、WordNet 语义词典和 SentimentWordNet 情感词典为基础，根据 HowNet 对词语语义定义和描述的特点，通过双语标注将 HowNet 义原和词语情感值转换为英文情感词典对应词语的查询与计算。具体来说，中文词语的情感极性值计算被分解为三个部分，分别为义原情感值的计算与校正，词语语义描述的语义关系分析以及最终的词语情感值的计算。按照该方法最终形成情感词典 SentiHowNet，词典的规模为：有积极极性情感词 12433 项，消极极性情感词 11148，词典的整体规模达到 23581 个条目。在实验部分，采用了直接评测的方法，以人工编辑的 HowNet 评价词典为基准，在覆盖度、精确率、召回率以及 F 值等指标上与现有的常用情感词典 NTUSD 和 DLLEX 词典进行了实验对比。实验结果表明，SentiHowNet 对基准词典的覆盖度最好，宏平均的精确率、召回率以及 F 值最高，证明了该自动构建情感词典方法的有效性，避免了人工和半自动方法的人工干预开销。

第三章　基于语料情感词典扩展

3.1　引言

上一章内容主要介绍了进行观点分析的所需的一项基础工作，也就是通用情感词典的构建方法的研究，提出了根据语义词典 HowNet 语义关系将英文情感词典情感知识跨语言转换为中文情感词典的自动构建方法，并构建了情感极性值标注的中文情感词典 SentiHowNet[155]。本章将对通用情感词典 SentiHowNet 在领域内进行基于语料的扩展，以增加该词典的领域覆盖度和适应性。

基于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直接从词典中抽取情感词语，词典中词语间的显式的语义关系也有助于进行情感极性至计算。但是基于词典方法构建情感词典覆盖度受限于词典资源规模，并且只能表示词语通用的情感知识，因而在实际应用中受到多方面的挑战：一方面词典中的词语往往具有多个义项，义项之间的情感极性不尽相同，在实际应用中会存在歧义问题；另一方面，语言的词语本身是开放的集合，会随着人类社会的发展而不断变化，尤其是对于专业领域中不断涌现的新词语，对情感词典在领域内的覆盖度提出了严峻的挑战；还有众所周知的领域适应性问题，也就是相同词语在不同的领域表达出的情感极性也会不一样。这些问题与挑战已经受到越来越多的研究者的注意[63, 107–109, 147, 156, 157]，并且提出了很多基于语料的情感词典构建方法，这些方法已在上一章的相关工作部分进行了详细介绍，在此不再赘述。随着互联网发展，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，网络上的各种类型用户产生内容（User Generated Content，UGC）包括评论（review）、博客（blogging）、微博（Twitter）以及社交网络的状态（status）等不断涌现并以指数级速度增长。这些类型的 UGC 中的语言具有不同于传统媒体和知识资源的一些新特点，表达情感的词语数量上达到了新的规模，形式上发生了巨大变化，表达情感极性更加依赖于具体领域和语境，因此研究社交媒体中观点信息需要对在传统知识资源上构建的情感词典进行扩展，提高在 UGC 数据上的覆盖度以及适应性。基于这种需求，本章提出并验证基于产品评论语料资源对 SentiHowNet 情感词典进行领域内无监督的扩展方法。

基于语料资源的情感词典构建方法需要解决两个问题，一是情感词语的选择，就是从语料中确定除了现有情感词典或种子词集以外的带有情感极性的词语，二是对这些词语在语料中表达的情感极性（值）进行计算。在英文中通常将这两个问题合并进行考虑，相关研究主要有两种思路：一是基于语料中文本的语言学线索，代表性研究是 Hatzivassiloglou 等[116] 基于并列或转折连词所连接词语的情感

极性制约规律确定新的情感词并判断极性；二是基于语料中词语出现语境中的统计信息，代表性的研究是 Turney 等[69] 采用点互信息（PMI）统计语料中与种子词集的词语共现度高的情感词，使用共现统计值并计算其极性值。因此借鉴英文的情感词语选择与极性计算研究方法，本章提出基于中文语料资源对 SentiHowNet 情感词典扩展的三种方法，并进行实验的验证。

1. 基于中文语言中的并列、递进以及转折关系的连词进行情感词的发现以及极性值计算方法；

2. 统计词语在语料中上下文信息进行情感词语选择和极性值计算方法；

3. 将基于连词线索方法和上下文信息统计方法相结合的方式进行情感词语选择和极性计算。

本章主要内容介绍采取与上一章不同的结构，首先介绍研究使用的词典和语料资源以及预处理，然后按照方法和实验想结合的方式分别介绍三种情感词典的扩展方法以及实验效果，最后作小节。

3.2 数据集及预处理

3.2.0.3 词典和语料

本章使用的词典和语料资源如表 3.1所示，选取的语料资源是谭松波博士提供的酒店、书籍和电子商品三个领域评论语料[158]，每个领域有文本 4000 篇，其中积极极性评级和消极极性评价各 2000 篇。

表 3.1 词典及语料资源

资源	名称	说明
词典	SentiHowNet 情感词典	基于上一章工作
语料	酒店（Hotel）评论	4000 篇，积极评价和消极评价各 2000 篇
	书籍（Book）评论	4000 篇，积极评价和消极评价各 2000 篇
	电脑（NoteBook）评论	4000 篇，积极评价和消极评价各 2000 篇

3.2.0.4 数据预处理

对数据的预处理主要包括分词与词性标注以及词语抽取形成结构化记录。对中文文本进行深入分析前需要进行分词和词性标注，本章直接使用现有的分词工具，也就是广泛使用的中科院 ICTCLAS[1]分词软件[159]，该工具可以对文本同时

[1]目前名字为 NLPIR/ICTCLAS，具体见http://ictclas.nlpir.org/

进行分词和词性标注。分词和词性标注后，过滤掉一些停用词[2]，语料中所有的词语都有可能成为被扩展对象，因此从语料中抽取出所有词语形成结构化的记录，记录格式如图 3.1所示，主要内容主要有词语编号（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，有些词在 SentiHowNet 中出现过，直接可以通过查询进行极性和极性值标注；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。

```
ID=135
Category=ADJ
W_C=有趣
Word_Tag=2
File_Tag=40.txt
Sentence_Tag=0
Senti_Tag=No
PosScore=0.6458333333333334
NegScore=0.0
```

图 3.1　语料预处理记录格式

3.3　基于连词情感词典扩展

Hatzivassiloglou 等[116] 通过对英文语言的连词在句子中的语法和语义作用研究发现，一个句子中由连词（如 and 或 but）连接的两个形容词或副词的情感极性存在一定的关联性，如并列连词“and”连接的两个形容词（如“nice and good”）情感极性相同，而转折连词“but”连接的两个形容词（如“nice but unnatural”）情感极性相反，否则就会引起语义上的错误（如“nice and unnatural”语法上没有问题，但是语义上存在不正确。)。有研究[160] 也发现中文语言也会有相同的规律，并且中文连词类型更丰富（有并列、转折、递进、总结、让步等类型的连词)，而且数量更多，因此可以选择更多的连词进行情感词的抽取。基于连词的情感词词典构建方法一般只能判断情感词的情感极性，如何能够计算得到情感词的极性强度值需要不同的方法设计。

3.3.1　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，对连词上下文的语义倾向性

[2]使用哈工大 1208 词的停用词表进行过滤

有限制作用的连词一般为并列连词、转折连次以及递进连词。在此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值计算，通过筛选选择的连词集合为：

- **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。

3.3.2 极性值计算

SentiHowNet 情感词典中的每个词都标注了情感极性值，因此语料中未标注的情感词的情感极性值可以根据所有与其在同一句子的 SentiHowNet 词语情感极性值计算获得。基于连词的情感词语极性计算基本思路是，对抽取到的待标注情感词语（目前之考虑形容词和副词）所在句子进行分析，确定连词及其类型，找到句子中极性值已知的情感词语并分析在句子中与连词和待标注词语的相对位置，然后依据以下原则进行计算：

1. 位于连词同一侧的形容词或副词具有相同极性；
2. 位于并列连词和递进连词两侧的形容词或副词具有相同极性；
3. 位于转折连词两侧的形容词或副词具有相反极性。

对未标注词的情感值计算，对极性相同的词语情感值累加，极性相反的词语极性值相减，然后取均值。具体情感极性值计算为：

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

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 分别表示与待标注词 w_t 在连接词同侧词语，在并列或递进连接词两侧词语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：

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

情感极性值具体计算过程如算法 3.1所示。

算法 3.1 基于连词的极性值计算

已知:

待标注词语集 $\{w_1\}$;

连词集合 $\{c\}$;

极性值已知词语集合 $\{w_2\}$;

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

3.3.3 实验

为了评价所提出的基于连词的情感词典扩展方法的性能，从酒店、书籍和电子商品三个领域评论语料中个随机选取了 200 篇评论（积极和消极极性各 100 篇）进行了人工标注。标注过程为：首先从评论中抽取其中的形容词和副词，过滤掉 SentiHowNet 中已有的词语，然后对剩余未标注词语根据在语料中的上下文标注积极和消极情感极性，标注好的词语作为评测基准。评价指标采用精确率（P）、召回率（R）以及 F 值（F）作为评测指标。使用基于连词的算法抽取计算并判断极性后，在三个领域扩展得到的情感词数统计如表 3.2所示。

表 3.2　情感词典扩展统计

领域	积极极性			消极极性			总体统计		
	基准	扩展	正确	基准	扩展	正确	基准	扩展	正确
Hotel	103	88	75	98	124	98	201	212	173
Book	166	166	112	245	196	196	411	362	308
NoteBook	61	61	41	66	58	50	127	119	91

从表中可以看出，在书籍领域扩展得到的情感词语较多，主要是因为书籍的评论篇幅较长，而且有更丰富的词汇来表达对书籍内容的评论，基于连词的扩展方法在消极极性词语的判断准确性较高；电子产品领域扩展的情感词较少，主要是因为对电子产品评论一般较短，表示评价观点的词汇相对比较少，专业化程度更高些，基于连词的极性计算和判断方法在该领域的准确性都比较低；而对于酒店领域基于连词的极性计算和判断方法在该领域的准确性最好。

针对三个领域的情感词典扩展实验性能评测结果如表 3.3所示，三个语料中与上面统计相对应，基于连词的极性计算方法在酒店领域的性能最好，无论是微平均还是宏平均，精确性都高于 72.82%，在消极极性的词语精确率甚至达到 100%，召回率都高于 79.03%，F 值都高于 78.53%；电子产品领域的性能指标相对较低，宏平均和微平均三个性能指标较低；在书籍领域，对消极极性情感词的判断召回率达到 100%，精确率以及 F 值在三个领域中都是较高，总体性能居中。总体来看，基于连词的方法能够对情感词典进行有效扩展，性能较好。

表 3.3　性能评测结果

领域	积极极性			消极极性			宏平均		
	P	R	F	P	R	F	P	R	F
Hotel	0.7282	0.8523	0.7853	1.0000	0.7903	0.8829	0.8607	0.8160	0.8378
Book	0.6747	0.6747	0.6747	0.8000	1.0000	0.8889	0.7494	0.8508	0.7969
NoteBook	0.6721	0.6721	0.6721	0.7576	0.8621	0.8065	0.7165	0.7647	0.7398

3.4　基于上下文情感词典扩展

词语的上下文是词语在实际应用中的语言环境，对于词语的语义理解有着重要作用，是自然语言处理经常使用的信息，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，在自然语言处理的具体应用问题解决过程中，上下文扮演着问题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[161]。对于本章要解决的情感词语抽取和极性值计算任务来说，统计情感词语出现的上下文特征可以为情感极性值的计算提供有用信息，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[162, 163]。

词语的上下文可以利用的信息有很多，一般是直接将上下文中出现的词作为有用特征使用[164, 165]，但是这种方法需要面临的一个问题是上下文特征的稀疏性，尤其是对社交媒体文本来说，篇幅一般较短，而且词语的不规范使用现象普遍，稀疏问题会更严重。因此我们采用了上下文词语的词性信息进行统计和计算。具体来讲，首先是对待标注情感词语，分析其上下文中词语的词性，获取一个词性特征向量，然后根据其上下文特征向量进行情感词语极性计算。

3.4.1 上下文特征向量

对于词语 w，通过分析在一定窗口范围内上下文中词语的词性形成词语的特征向量，下面给出形式化定义。

定义 (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:

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

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。

3.4.2 基于词性特征向量的情感词极性值

基于上下文的情感词极性值计算的一个假设就是具有相同的上下文特征向量的词语具有相同的极性，根据这一假设，统计出 SentiHowNet 中与待标注词 w_t 词

性特征向量相同的情感词，然后计算 w_t 的情感极性值。如果待标注词 w_t 出现的上下文中有 N 中词性特征向量，则 w_t 的情感极性值计算公式为:

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

其中 $W_{positive} + W_{negative} = M$，表示与待标注词 w_t 具有同一特征向量 $V(w_t)$ 的 SentiHowNet 中的情感词集，$W_{positive}$ 和 $W_{negative}$ 分别为极性为积极和消极的词语。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。

以书籍语料中形容词“亮丽”为例，设定 $W = 2$，在语料中抽取“亮丽”所有的 N 种词性特征向量，其中的一个的特征向量为 < *daaun* >，从 SentiHowNet 中获取与“亮丽”的这一词性特征向量相同的所有情感词如表 3.4所示，共有 16 个情感词与“亮丽”有相同的词性特征向量 < *daaun* >，其中积极极性词语 10 个，消极极性词语 5 个，于是可以使用公式 3.3计算其情感极性值为积极极性值为 0.4297，消极极性值 0.0421。当然这是词性特征向量 < *daaun* > 计算出的极性值，将所有 N 种词性特征向量下计算的极性值取平均就可以得到“亮丽”在书籍领域最终情感极性值。

表 3.4　计算示例

情感词	情感词上下文	情感词	情感词上下文
大	没有多大的可读性	弱小	再贫困弱小的人
简单	太浅显简单的东西	有趣	又生动有趣的绘画
真实	比较客观真实的角度	直观	没有明显直观的效益
新鲜	更多新鲜的元素	奇妙	更多奇妙的东西
严格	太多严格的界限	好	更多好的作品
石破天惊	颇多石破天惊之语	肮脏	太多肮脏的东西
柔软	最脆弱柔软的心脏	多	反正好多的故事
粗俗	原本低级粗俗的俚语	曲折	明明惊险曲折的战争

更一般的，基于词性特征向量的情感极性值具体计算过程如算法 3.2所示。

3.4.3　实验

评测实验使用的数据、基准与评测指标与基于连词的方法 3.3节相同。对三个领域的情感词典扩展实验结果如图 3.2、图 3.3和图 3.4所示，三个图中分别显

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 $\{V(w)|w \in \{w_1\} \cup \{w_2\}\}$;

1: **for** 每一待标注词语 $w_1 \in \{w_1\}$ **do**

2: **for** w_1 每一特征向量 $V(w_1)$ **do**

3: **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**

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

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

6: **else**

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

8: **end if**

9: **end for**

10: 对各个特征向量下的情感值累加

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

12: **end for**

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

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

15: 将 w_1 加入到集合 $\{w_2\}$;

16: **end for**

示了精确率、召回率以及 F 值的积极和消极极性的宏平均结果。图中可以看出当窗口 $W = 1$ 时，在三个领域的数据中，基于词性特征向量方法都能够准确抽取所有的情感词，因此精确率、召回率以及 F 值相同，分别为 67.65%、72.89% 和 72.13%；当窗口增加到 $W = 2$ 时，召回率都略有上升，精确率都有所下降，酒店领域基本不变，书籍领域下降 6.62%，电子产品领域下降最多（8.1%），F 值都变化不大；当窗口增加到 $W = 3$ 时，除酒店领域的召回率上升外，其他两个领域和其他指标都出现明显下降。因此通过分析各个领域的性能指标，采用上下文窗口 $W = 2$ 时，基于词性特征向量的情感值计算方法性能最好，这与在英文领域的研究结论[162, 163] 是一致的。

当窗口 $W = 2$ 时在三个领域扩展得到的情感词数统计如表 3.5所示。

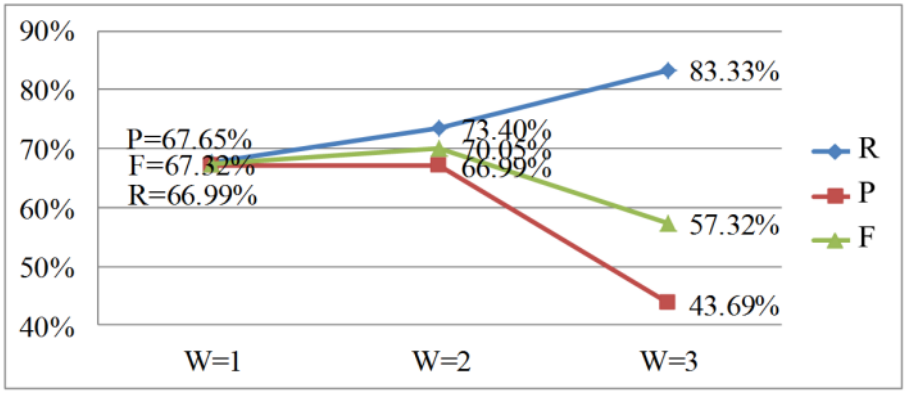


图 3.2　Hotel 语料评测结果

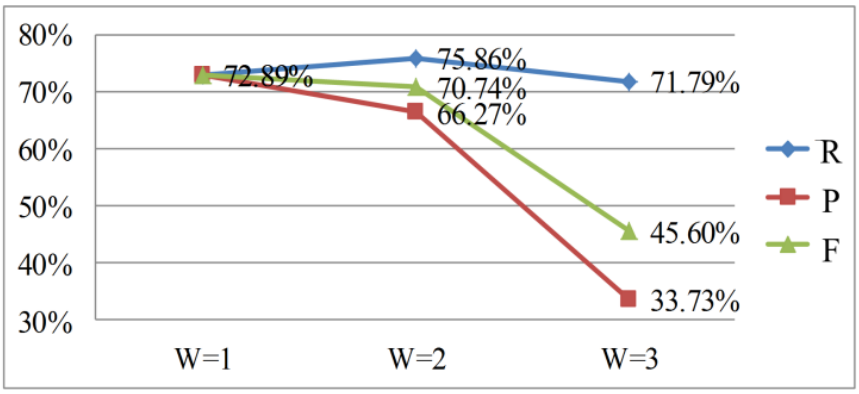


图 3.3　Book 语料评测结果

3.5　基于混合方法情感词典扩展

对基于词性特征向量的情感词典扩展方法的实验结果分析发现，在统计词语的上下文词性特征向量时，有些词在 SentiHowNet 中没有发现与之词性特征向量相同的情感词，无法采用基于词性特征向量方法计算情感极性值，这种情况下可以考虑采用基于连词的方法进行计算。同样的，对基于连词的情感词典扩展方法实验结果分析发现，在语料中有些情感词出现的所有句子中没有任何并列、递进或者转折连词出现，无法采用基于连词方法计算情感极性值，这种情况下可以考虑采用基于词语上下文的词性特征向量方法进行计算。两种方法可以相互补充，因此本节提出将两中方法混合使用的方法，并实验验证混合方法的性能。

3.5.1　基于混合特征的情感词极性计算

对任一选取的情感词 w，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：

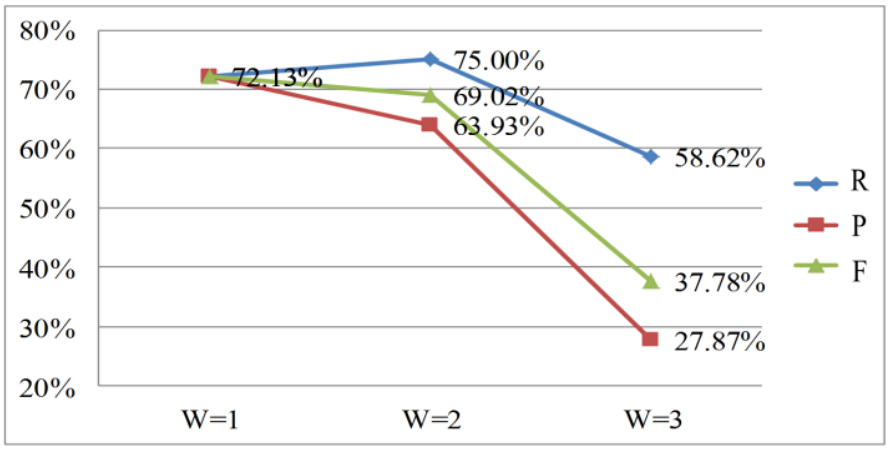


图 3.4　NoteBook 语料评测结果

表 3.5　情感词典扩展统计

领域	积极极性			消极极性			总体统计		
	基准	扩展	正确	基准	扩展	正确	基准	扩展	正确
Hotel	103	98	64	98	86	71	201	184	135
Book	166	128	115	245	231	157	411	359	272
NoteBook	61	50	42	66	58	39	127	108	81

1. 优先采用基于词性特征向量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. 当采用基于词性特征向量的方法进行计算时，优先设置上下文窗口大小为 2，其次为 1。

3. 当采用基于词性特征向量的方法无法对待标注词语进行情感值计算时，采用基于连词的方法进行计算。

具体的混合情感词语极性计算如算法 3.3。

3.5.2　实验

基于混合方法评测实验使用的数据、基准与评测指标与基于连词的方法 3.3节相同。在三个领域扩展得到的情感词数统计如表 3.6所示。与表 3.5以及表 3.2进行对比，可以看出基于混合方法在每个领域的正确标注词书目数目比基于连词和基于词性特征向量的方法都有明显增加。

对三个领域（Hotel、Book、NoteBook）的情感词典扩展实验精确率、召回率以及 F 值的微平均及宏平均结果如表 3.7所示。表中可以看出，基于混合方法在三个领域的性能指标都比较稳定，宏平均的精确率、召回率以及 F 值都在 74% 以

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

表 3.6　情感词典扩展统计

领域	积极极性			消极极性			总体统计		
	基准	扩展	正确	基准	扩展	正确	基准	扩展	正确
Hotel	103	103	64	98	100	88	201	203	152
Book	166	175	125	245	236	192	411	411	317
NoteBook	61	66	48	66	61	52	127	127	100

上，电子产品领域性能最好（三个指标都达到 78.69%），酒店领域稍低，数据领域居中。

表 3.7　各个领域性能评测结果

领域	积极极性			消极极性			宏平均		
	P	R	F	P	R	F	P	R	F
Hotel	0.6214	0.6214	0.6214	0.8980	0.8800	0.8889	0.7549	0.7476	0.7512
Book	0.7837	0.8136	0.7983	0.7530	0.7143	0.7331	0.7711	0.7711	0.7711
NoteBook	0.7869	0.7273	0.7559	0.7879	0.8525	0.8189	0.7869	0.7869	0.7869

为了综合比较本章的三种方法，基于连词的情感词典扩展、基于词性特征向量的情感词典扩展和混合方法的情感词典扩展的实验评测宏平均结果对比情况如图 3.5、图 6 3.6和图 3.7所示，从比较结果来看，混合方法在性能是在三个领域语料中均是最优的，是比较理想的情感词典扩展方法，其次为基于词性特征向量的

方法（取窗口为 2 时），基于连词方法性能接近于词性特征向量的方法。因此在选择对 SentiHowNet 情感词典扩展方法时，优先选择混合方法。

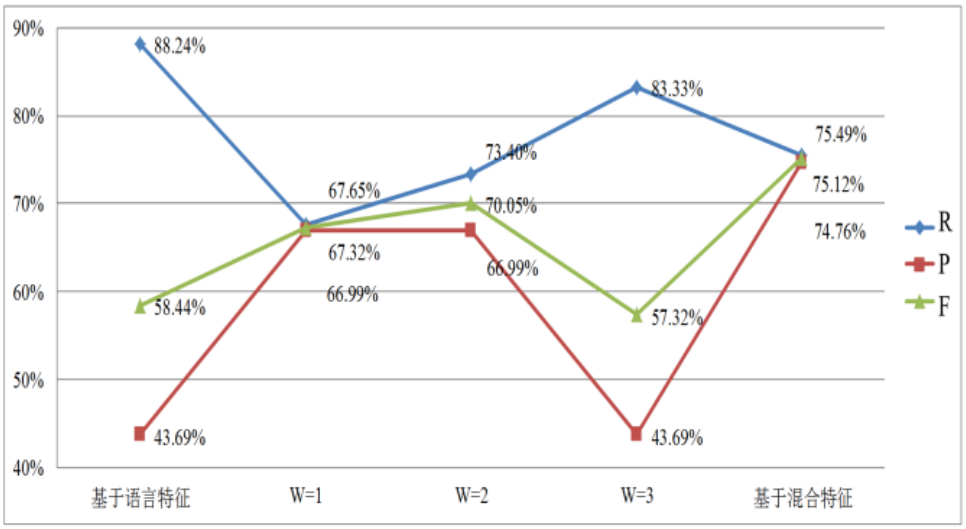


图 3.5　Hotel 语料评测结果综合比较

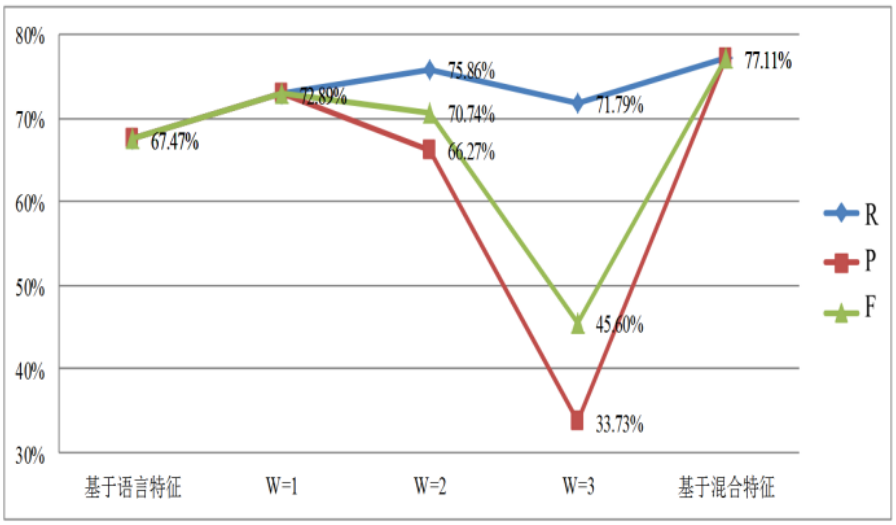


图 3.6　Book 语料评测结果综合比较

3.6　小结

本章详细讨论了基于语料资源的中文情感词典的三种扩展方法，并用三个领域的语料进行了实验验证。第一种方法是基于连词语言线索方法，该方法根据语言中连词对词语语义倾向的限制作用，选择了并列、转折以及递进三类连词来进行词典扩展，设计了基于连词的情感词的极性值计算方法以及详细算法，实验验证了中文连词在情感词典扩展中的有效作用；第二种方法是基于词性特征向量的

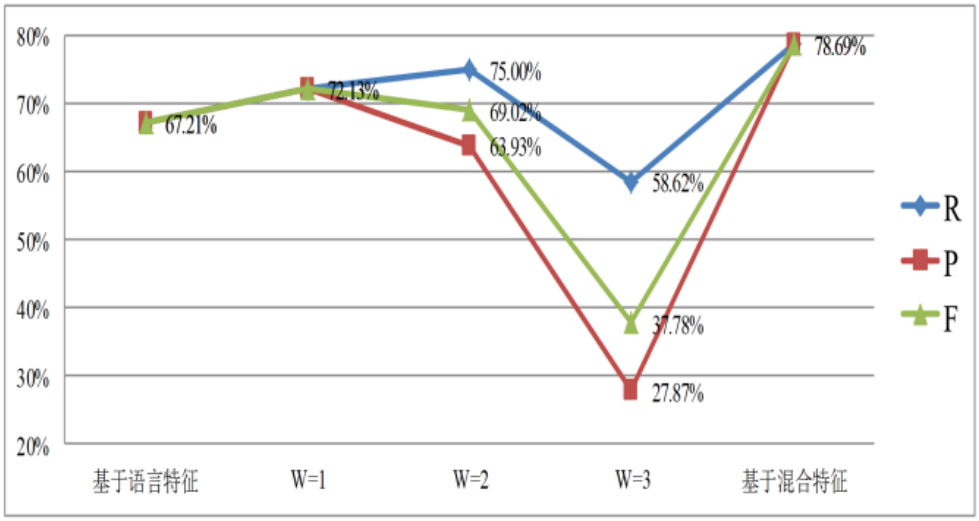


图 3.7　NoteBook 语料评测结果综合比较

方法，根据出现在相同上下文的词语语义相似性，设计词语的词性特征向量，统计具有相同词性特征向量的情感词进行情感极性值计算，并设计了详细算法，实验结果证明了词性特征向量的有效性；第三种方法为一二种方法的混合，可以起到相互补充的作用提高情感词的识别准确性，实验结果证明混合方法在三种方法中在三个领域中最为稳定，达到最好性能。

第四章　无监督自举式情感分类

4.1　引言

前面两章介绍了观点分析一个基础的情感知识表示的问题，也就是如何构建情感词典。本章主要介绍如何确定主观文档的观点倾向性，也就是对文档表达的情感进行极性分类（Plarity classification），或者称为情感分类（Sentiment classification）[58]。情感分类一直是观点分析的主要研究内容，主要方法可以分为基于词典的方法以及基于机器学习方法[24]。基于词典的方法主要是以情感词典为核心，将情感词典表示的情感知识与一些人工设计的规则组合在一起就可以对文档情感极性分类，无需标注语料和训练过程，省去了标注语料的时间人力开销，因此得到广泛的应用[78, 143, 146]。但是这种方法为了获得好的分类效果，需要对情感词典进行领域内的适应性转化，性能受到词典的规模、覆盖度以及规则复杂度的影响。而基于机器学习方法通过标注语料的训练过程可以从文档中学习到一些可能无法用人工设计的规则表示的情感表达模式，也就是从文档抽取的各种特征通过不同权重组合在一起表达文档情感极性的方式[166]，因此基于机器学习的方法一般比基于词典的方法分类性能要好[23]。基于机器学习方法性能主要取决于抽取到的特征，Pang 等[75] 第一次将机器学习方法引入情感分类研究，通过研究发现“词袋（Bag of words）”特征相对于其他特征是比较稳定且有效的分类特征，因此各种机器学习方法都会选择文档中出现的词语作为主要分类特征。情感分类与文本分类类似，都是将文本分为预先设定的类别，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具体语境[167]。尤其是社交媒体语言，存在着大量特有的情感表达方式，包括不断涌现的新词和符号，传统词语的新用法等等，这些特定表达方式只有在特定语境下才能表现出情感极性，对情感分类任务提出了新的挑战，其中微博就是最具代表性的。

随着微博（Twitter、新浪微博等）的出现和广泛使用，微博用户产生内容成指数倍增，这些内容里面有很多用户对于各种话题的观点和情感等有用信息。因此可以从微博中提取出有用观点信息，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[24]。但是对微博进行情感分类具有特别的挑战性，主要是因为：

1. 首先，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. 其次，更具挑战性的是微博语言是动态的，虽然语言都是动态的，但是微博语言的变化更加迅速，因为微博用户群体的复杂性，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，使得微博上情感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和时间为代价，并且微博语言的动态性使得标注数据具有时效性，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。该框架首先使用现成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去训练得到微博分类器；将两个分类器结合起来，迭代使用协同训练（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

情感分类在观点分析研究中越来越受到重视，前期工作主要研究针对（商品或电影）评论进行情感分类。经常使用的方法可以分为基于词典的方法和基于机器学习的方法，其中基于机器学习方法性能一般要好与词典方法，因此常被其他方法用作对比的基准[75]。

目前情感分类研究开始关注于微博的情感分析问题，将前期研究方法结合微博的语言特点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特有的语言特点结合进情感分类方法中。比如，Barbosa 和 Feng [168] 提出了两阶段支持向量机分类器（Support Vector Machine，SVM）对 tweet 进行情感分类，验证了分类器对 tweet 的类别偏置（biased）和噪声具有很好的鲁棒性；Hu 等 [169] 将社交媒体数据中的情感表达分解成情感指征（emotion indication）和情感关联（emotion

correlation）两种信号，通过对两类情感信号进行联合建模方式实现了对微博的无监督情感分类；Jiang 等 [167] 主要关注依赖于特定话题的微博情感分类，提出了通过将话题依赖特征（target-dependent features）和相关微博同时进行考虑的监督学习方法，可以提升情感分类性能；Wang 等 [170] 针对 hashtag 层面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 [171] 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话题的微博，将微博按情感类型区分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 [172] 基于社交理论抽取微博之间的情感关系，提出了新的基于社会学研究的方法并使用情感关系提高了情感分类性能，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 [173] 依据人的观点通常会带有偏执的一致性，提出了基于迁移学习（transfer learning）方法解决微博实时情感分类问题；Thelwall 等 [174, 175] 设计了 SentiStrength 情感分析工具，用于对微博等社交媒体中非正规语言中的情感分析，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确定微博中的情感强度，该工具获得了广泛的应用。

以上这些工作通过利用微博作为社交媒体的一些网络结构和语言特点对情感分类方法进行了适应性的改进，以便于提出的方法能够适应于微博语言环境，但是没有有效解决微博情感分类语言环境依赖问题，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划分为环境依赖部分（Context-dependent part）和环境独立部分（Context-independent part），然后分别进行训练分类器，最后将两种分类器结合进一个自举式（bootstrapping）学习框架中形成一个更强的情感分类器。

4.3 问题的形式化

情感分类主要目的就是将文档分类为预先定义的情感极性类别（一般是积极极性或消极极性）。具体来讲，对于给定的文档语料库 $D = \{d_1, \ldots, d_n\}$，预定义的情感极性类别 $Y = \{1, -1 \mid \text{positive} = 1, \text{negative} = -1\}$，情感分类的任务就是对每一篇文档 d_i 预测一个类别标签 y_i。与文本分类问题一样，每一篇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，特征权值通常定义为二值的，1 表示特征在文档中出现，0 表示没有出现，这种情况下分类性能更好[75]。对于有监督的机器学习，给定训练集 $D = \{x_1, \ldots, x_m\}$，可以通过训练学习到分类器：

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

对于待分类文档 x，同样可以将其表示为特征向量 $x = (w_1, \ldots, w_v)$（w_i 表示第 i 维权重），其情感极性类别可以使用分类器 f 预测：$f(x)$。

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词特征向量中所有的词语在表达情感极性时作用是等同的，也就是抽取到的词语都是潜在的文档情感极性表征，每个词语的出现与否都有可能决定该文档的情感极性。实际上这种假设是有问题的，因为有些词语在文档中表达的是客观信息，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，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.

该微博的词袋模型是将所有的词语都抽取出来作为特征加入到特征向量中(当然会经过去停用词等预处理)，然后这些词语同等地用于对这条微博的情感极性进行建模分析。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极性，而另外一些词语（Happy，birthday，wish，best，thanks）则无论在什么领域或语境下都是积极情感极性的的标识。基于这样的直观认识，我们提出以下特征空间划分的假设：

假设 (特征空间划分)：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：也就是通用的词语特征，该部分词语特征在任何领域和语言环境下都是某种情感极性的表达方式。
- 领域依赖部分：也就是具体语言特征（包括词语以及符号），这部分特征只有在微博这种语言环境下才能表达一定的情感极性。 ■

这个假设可以更加形式化的表示为：对于情感分类问题中微博 x 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：

$$x = \begin{cases} x_g = (w_1, \ldots, w_l) \\ x_c = (w_{l+1}, \ldots, w_v) \end{cases} \tag{4.2}$$

其中，x_g 是特征向量空间的通用部分，而 x_c 是领域依赖部分。

基于以上假设，情感分类问题可以形式化定义为：

定义 (基于特征空间划分的情感分类): 情感分类问题空间可以表示为 (X_g, X_c, Y)，其中:

- $X_g \subset R^d$ 和 $X_c \subset R^p$ 为通领域独立和领域依赖两部分输入特征空间，$d+p=n$，d 和 p 分别表示两部分特征空间的维度;
- 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，IID）微博实例集合 $D=\{(x_i^g, x_i^c, y_i)|i=1\cdots m\}$，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 x_i 领域独立部分特征，向量 x_i^c 表示领域依赖部分特征，y 表示实例的情感极性类别;

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 $D_g=\{(x_i^g, y_i)|i=1\cdots m\} \subset (X_g \times Y)^m$ 和 $D_c=\{(x_i^c, y_i)|i=1\cdots m\} \subset (X_c \times Y)^m$ 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 f_g 和微博情感分类器 f_c。

$$\begin{cases} f_g : D_g \rightarrow Y \\ f_c : D_C \rightarrow Y \end{cases} \tag{4.3}$$

■

当然基于部分特征空间的分类器性能上是否会降低是一个值得研究的问题，因此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，怎么区分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 自举式情感分类框架

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”，尤其是对于不常使用微博的人来说，对于一条微博中出现的一些词语可能不理解其表达的语义。但是整条微博的情感极性性却能够比较容易

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极性。直观上，这种现象可以通过特征空间划分假设来解释，也就是正规表达部分特征也能从一个不同的视角（view）来阐释整条微博的情感极性。而这些正规表达部分特征 x_g 是不依赖于具体微博语境的，对于任何人（常使用微博的或是很少使用微博的）都是容易理解的。

类似的，对微博情感分类，基于特征空间划分假设，可以认为一条微博的情感极性可以通过两部分特征分别能识别出来。也就是说，如果能够拥有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)。实际上有很多研究已经开始构建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[61, 176]、ANEW 词典[177]、AFINN 词典[178]、SentiWordnet[135]、HowNet 情感词典[104]，NTUSD 情感词典[105]、情感词汇本体词库[106] 以及我们构建的 SentiHowNet[155]。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极性性还是需要具体的上下文语境进行“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对微博情感分类研究是十分重要的。人类实际的语言知识库中这样语言现象是存在的，比如成语和谚语等就具有无歧义的情感极性性，无论在什么样的语境下都保持一致不变的情感极性，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节研究的重点。

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，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。这些成语的情感极性是固定不变的，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、情感极性性（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。由于成语的情感极性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：

假设 (条目情感极性)：每条成语条目跟描述的成语一样就具有领域独立的情感极性，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，我们可以根据现成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。从成语词典中条目中抽取的所有词语特征可以看作是领域独立部分的特征。

4.4.2 微博情感分类器

由于领域独立特征只是全特征空间的一部分，在识别情感极性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，比如特殊符号、缩写以及不规范用法等等。为了能够更好的识别出微博中的一些独特细致的情感，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所有词语特征中哪些是领域依赖的特征。随着微博数量持续增长以及用户在微博中语言使用的自主性，微博特有的观点或情感新的表达方式不断出现，即便使用通用词语，有时候在特定的微博语境下也会出现不同于原有的语义倾向。不断出现的新表达和旧词新用现象使得界定情感分类问题中领域依赖部分特征变得复杂，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主题表达某种情感极性时，除了描述主题所用词语外，只能够用很少的词语描述情感极性。因此可以认为，如果一条微博中含有某个成语，如果没有表示否定的词语，微博的情感极性可以看作是和成语或情感极性一致，并且微博语言的简洁使得我们可以将除成语外的其他词语特征可以视为领域依赖特征。第二个问题是如何能够获得标注数据来训练基于微博依赖特征的微博情感分类器。在微博情感分析研究中为了解决标注数据缺失问题常用的方法是使用远监督（Distant supervision）自动获得来标注数据[179, 180]，远监督方法主要是利用微博中具有明确情感极性的一些表达方式（如表情符）作为启发式规则挑选和标注数据，自动构造训练数据。微博情感分类器的训练数据构造也是基于这种思想，主要是利用成语资源作为微博情感极性判断的准则，找到包含成语的微博（过滤掉含有否定词微博）作为微博情感分类器的训练数据。

4.4.3 分类器的组合

本章提出的基于特征空间划分情感分类方法一个基本假设就是用户在表达一种主观观点时会使用不同的表达方式，一是可以使用不依赖于领域和语境的通用情感词语，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的特有的表达方式。因此我们可以将微博分类问题中情感极性表达特征空间划分为通用特征和领域依赖特征，主要目的是对同一条微博从相互补充的两种视角（View）来分析，使用特征空间的不同部分训练不同的情感分类器对同一条微博情感分类，达到更好分类效果。虽然两种分类器都能对微博进行情感分类，但是性能会受到训练数据的数量和质量的制约。从上面两小节的介绍可以看出，无论是成语的释义文本还是微博文本，都是比较短小的，而且微博不规范的语言特点，使得从这两种文本中抽取得到的特征向量会比较稀疏，对分类器的性能造成影响。为了能够应付这些问题，本小节提出一个自举式（Bootstrapping）学习框架将两种分类器组合在一起，相互补充，形成一个更强的情感分类器。

自举式学习框架如图 4.1所示，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 f_g 和微博情感分类器 f_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，组合性能提高。根据使用的分类器的不同，将分类器的输出用 $\{p_g, p_c\}$（例如 SVM 分类器的输出的判定距离或生成模型的判定概率值）来表示对微薄分类结果的可信度。对于每一条待分类微博，首先分别使用两个情感分类器对其单独进行分类，预测其情感极性性标签为 $c_i = \{c_g, c_c\}$，并输出可信度 $p_i = \{p_g, p_c\}$，然后将可信度按照公式～4.4进行加权组合：

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

公式中 λ 是控制不同分类器在同一条微博情感分类中的权重，对应于特征空间中领域依赖部分和领域独立部分在微博情感极性的不同决定作用。实际使用中首先将其初始化为 $\lambda = 0.5$ 以表示两部分特征的作用是等同的，对微博确定一个初始情感极性类别，然后随着迭代步数逐步增加，逐渐加大 λ 值以使得组合起来的分类器逐步适应微博的情感分类。如图中所示，该框架根据两个分类器对每个测试数据分别预测一个情感极性标签 $c_i\,(c_i \in \{1, -1\})$，根据预测标签将测试数据分为积极极性和消极极性两组，并分别在组内按照公式～4.4计算的可信度降序排列。从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微博情感分类器的训练数据加入到训练集中，以逐步提高组合分类器对微博数据的适应性。

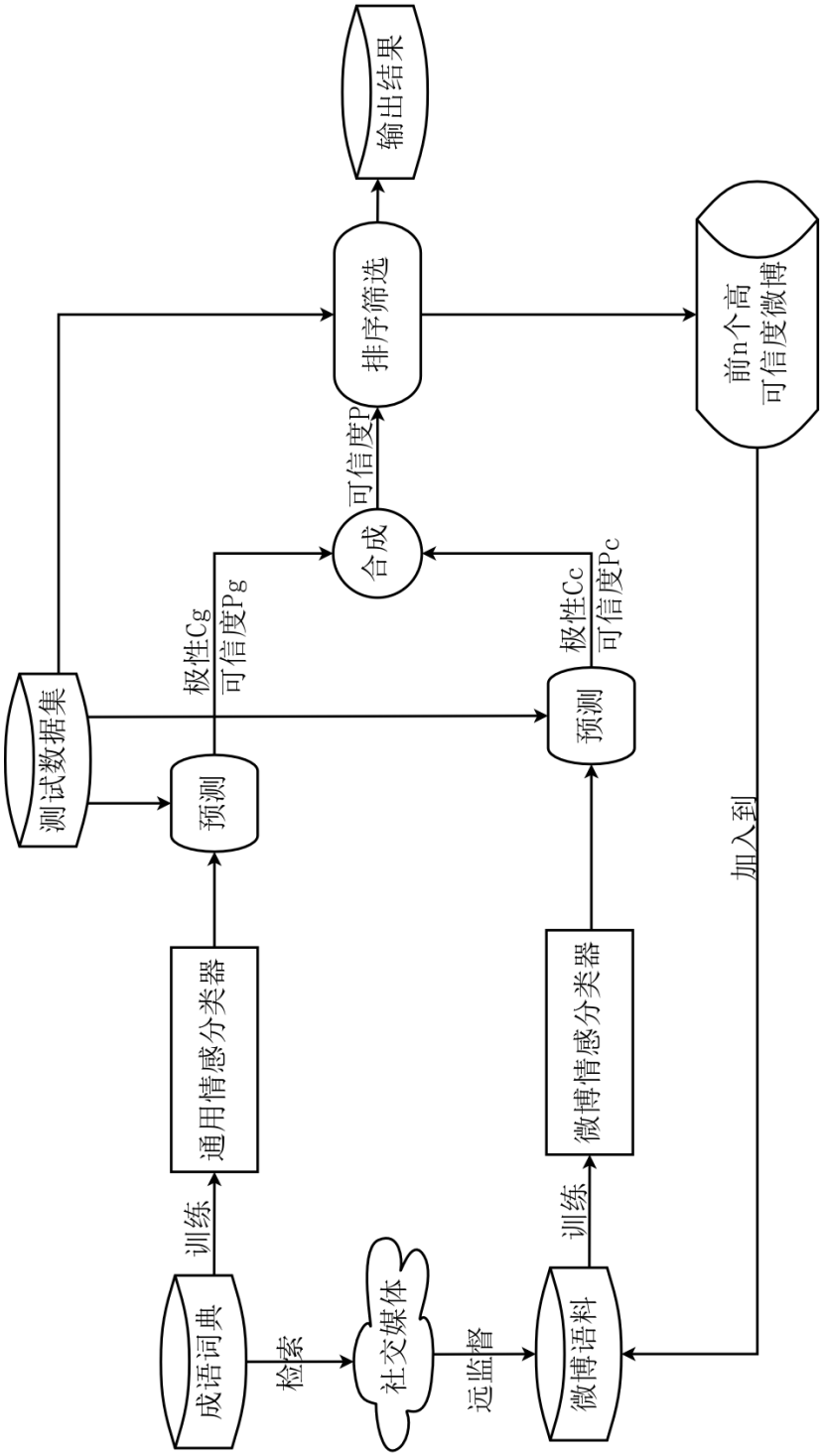

输出结果
排序筛选
可信度P
合成
前n个高
可信度微博
加入到
测试数据集
极性Cg
可信度Pg
预测
极性Cc
可信度Pc
预测
通用情感分类器
微博情感分类器
训练
训练
成语词典
检索
社交媒体
远监督
微博语料

图 4.1　自举式学习框架

这样的过程多次循环迭代，直至所有测试数据的情感分类可信度的变化因为小于某个阈值而收敛，最后得到这种自举式学习形成的情感分类器。

总体来说，整个框架可是一个协同训练（Co-training）[180] 的自举式（bootstrapping）机器学习算法，但是该框架并没有需要使用标注训练数据，而是从现成的成语词典资源作为训练的起始点，可以看作是一个无监督的学习算法，避免了标注微博数据所需时间和人力代价，对数量庞大的微博数据情感分类问题，该框架更加实用。

4.4.4 分类器算法

对于框架中的两个分类器，我们采用跟 Pang 等 [75] 使用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 [75] 的研究验证，并证明支持向量机取得的性能是最好的（在电影评论数据集上的准确率达到 82.9%），本节对这三种算法进行简单介绍。

4.4.4.1 Naïve Bayes 分类器

Naïve Bayes 算法在文本分类任务中是最常用的方法。对与情感分类问题，为了确定一篇文档 d_i 的情感极性性类别 c_j，Naïve Bayes 算法需要计算后验概率 $P\left(c_j \mid d_i\right)$。根据 Bayes 法则和多项式分布，基于每一维特征概率独立性假设，可以得到：

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

公式中 r 表示不同的情感类别，w 表示文档中出现的词语特征。通过计算不同情感极性类别的后验概率，概率最大极性类别被视为文档 d_i 的情感极性类别。

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验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：

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

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。最大熵分类器使用训练数据集 D 训练学习过程就是解一个最优化问题过程：

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

4.4.4.3 SVM 分类器

支持向量机（SVM）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就是发现支持向量所确定的决策平面，该平面能够将训练数据在特征空间上分开为两类，然后使用支持向量确定测试数据的情感极性类别。训练过程是解一个受限的最优化问题：

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

情感分类问题通常使用线性支持向量机分类器。

4.5 实验

本节主要介绍针对腾讯微博[1]开展的情感分类实验。

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]抓取了在线成语词典，经过对词典数据整理和去除不具情感极性成语条目，最后得到了有 8,160 个条目的成语数据集，其中褒义（积极情感极性）的成语有 3,648 条，贬义（消极情感极性）的成语有 4,512 条，将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。在 2013 年 4 月 15 日开始到 5 月 15 日一个月的时间内，通过监控腾讯微博的实时数据流，查询并抓取了至少含有一条成语的微博，形成 120,346 条微博数据集。经过筛选和过滤，过滤掉过短微博和噪声，最后得到 91,268 条微博组成的数据集用于训练微博情感分类器。为了测试自举式情感分类器的性能，实验使用了第一届自然语言处理与中文计算会议（Natural Language Processing and Chinese Computing）微博情感分析与语义关系抽取评测（First Chinese tweet Sentiment Analysis and Semantic Relationship Extraction Evaluation）[3]发布的标注微博数据集作为测试数据。评测数据来自腾讯微博，评测数据全集包括 20 个话题，每个话题采集大约 1000 条微博（其中积极极性和消极极性各 500 条），共约 20000 条微博。

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

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

[3] 中国计算机学会（CCF）举办: http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html

4.5.1.2 实验配置

为了能够多角度评价分类器的性能，机器学习中有各种评测指标，由于本实验不是为了比较这些评价指标的不同，因此实验中选择了比较直观的准确率作为分类器性能的评价指标，准确率主要检验分类器对测试数据集分类的准确性。

对于分类器工具，实验选择了自然语言处理的工具箱 NLTK（Natural Language ToolKits）[181] 中的 Naïve Bayes 分类器和最大熵分类器，以及常用的 Libsvm[182] 工具包的支 SVM 分类器。以上分类器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

为了客观评价自举式分类器的性能，实验设置了三个评价基准用于对比评测。一个是 50% 的基本基准，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准确率可以达到 50% 准确率；第二个基准是用一个基于情感词典的情感分类器性能作为对比基准，实验使用的是前面两章构建的 SentiHowNet 情感词典，通过将每条微博中的包含的情感词的情感极性值叠加来计算微博综合情感极性值，然后根据综合情感极性值判断微博的情感极性；第三个基准是使用标注数据进行有监督训练的机器学习方法构建的情感分类器性能，实验按照 Pang 等 [75] 的方法使用测试数据通过 5 倍交叉验证方式训练了 Naïve Bayes 分类器、最大熵分类器以及 SVM 分类器。

4.5.1.4 数据预处理

中文文本信息不像英文那样自然单词结构，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。实验使用中科院 ICTCLAS 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 4.2中的最优的 λ 值，首先进行了从 0 到 1 对 λ 值遍历实验，实验中分别使用成语数据集和通过远监督方式获取微博数据集训练分类器，两个分类器对测试数据同时进行情感分类，分类输出的可信度值按照 λ 加权组合判断测试数据的极性类别，用准确率评价。每一次遍历 λ 值增加 0.1，实验结果如图 4.2所示。

从图中可以确定对于三种分类器最优的 λ 值为：对 Naïve Bayes 分类器，$\lambda = 0.4$；对最大熵分类器，$\lambda = 0.5$；对 SVM 分类器，$\lambda = 0.3$。

确定了 λ 后，使用图 4.2的自举式的学习框架进行训练后对测试数据进行情感分类，分类准确率对比结果如表 4.1所示。

根据表中结果，可以得出以下结论：

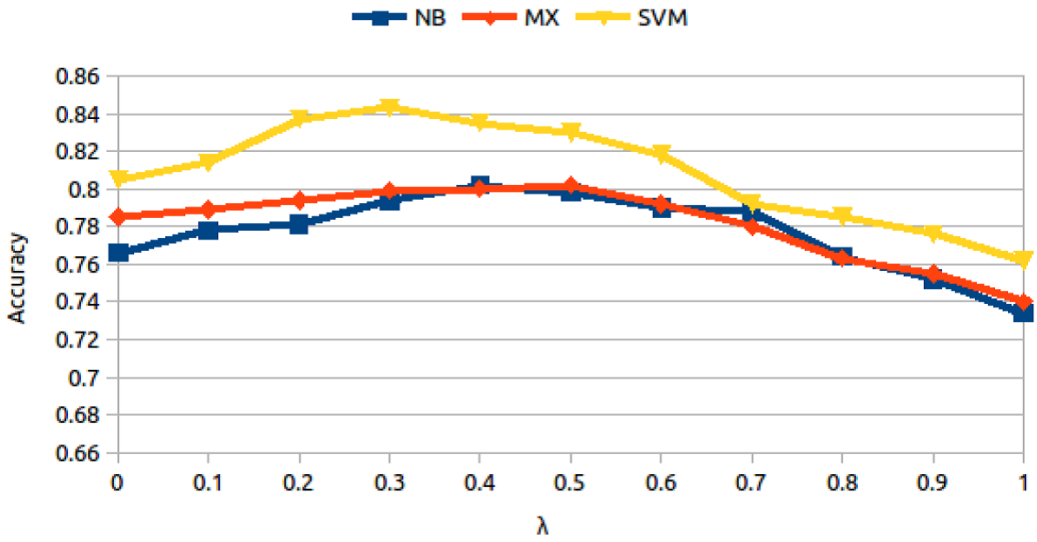


图 4.2　λ 值的遍历。

表 4.1　结果对比表

分类器	Naïve Bayes 分类器	最大熵分类器	SVM 分类器
基于词典方法	0.725	0.725	0.725
有监督学习方法	0.785	**0.806**	0.826
领域独立部分特征	0.734	0.740	0.762
领域依赖部分特征	0.766	0.785	0.805
自举式学习方法	**0.802**	0.802	**0.843**

- 首先，无论是基于领域独立部分特征的通用情感分类器还是基于领域依赖部分特征的微博情感分类器，准确率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，比随机猜测更准确。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，本章提出特征空间划分假设可以作为微博情感分类的一种有效的方法。

- 其次，通用三种情感分类器的准确率都比基于情感词典的分类器高一些，这是因为尽管通用情感分类器和基于词典的分类器潜在假设都是使用情感极性独立于领域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知识资源所抽取的特征空间训练得到的，而情感词典中情感知识是由单个词语的独立情感极性（值）组成，很多词语的情感极性往往具有一定的歧义；而对于微博情感分类器，准确率比基于词典的分类器和通用情感分类器都要高，因为它是使用微博依赖部分的特征训练得到的，更能适应微博语言环境，测试数据中出现的微博式的“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训练得到的组合分类器显示出了最好的准确率，因为它结合了通用分类器和微博分类器的情感分类性能，其准确率也超过了被 Pang 等 [75] 证明的准确率比较高的有监督分类器，这说明本章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极性，也能照顾到微博语言特有的情感表达发那个是，准确反映出微博细致的情感极性。

4.6 小结

本章首先分析了情感分类问题中面临的领域依赖问题，并针对微博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。根据微博情感表达特点，从通用情感表达和微博特有情感表达两个视角对情感分类问题进行了重新定义，通过将情感分类问题的特征空间进行划分，将整个词语特征空间分为领域独立特征的领域依赖特征两个部分。基于两部分特征划分的假设，提出了在两个特征空间分别训练通用情感分类器和微博情感分类器的解决方案，其中通用情感分类器使用现成的成语词典资源作为训练数据进行训练，微博情感分类器使用远监督方式获取训练数据。为了能够综合两部分特征空间的情感分类作用，设计了自举式的机器学习框架将两种分类器组合起来，形成分类效果更好的情感分类器。实验证明本章所提出的方法性能上超过了基于词典以及有监督机器学习方法，获得了良好的微博情感分类准确性。

第五章　用户主观性建模

5.1　引言

上一章介绍了如何对社交媒体中的单个文档进行情感分类以确定文档观点表达的情感极性，本章将在上一章的基础上介绍用户层面的观点分析，主要包括如何对用户发表的多个文档中的观点信息进行集成以及集成的观点如何进行表示。

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社交媒体上针对各种话题发表短的文本信息表达意见和观点。本章研究的用户主观性是就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。一方面，社交媒体的文本数据因为覆盖话题广泛和用户观点信息丰富而成为研究用户主观性的重要数据来源。另外一方面，使用社交媒体数据研究用户的主观性反过来也会有利于针对社交媒体的后续研究及应用，比如用户观点查询、观点追踪或者用户行为的预测等研究，在社会学、心理学、政治以及商业领域具有重要作用。社交媒体中用户产生的内容数量巨大，而用户产生的文本信息短小分散，以碎片化形式存在于海量的社交媒体数据中，因此用户的主观性信息是散布在“碎片化的信息”中，使得从这些数据中挖掘和分析用户各种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结果，此处结果查询日期为 2014 年 2 月 14 日），会返回大概 231,233 用户的 830,879 条微博（tweet，本文中统称为微博），意味着很多用户发表了不止一条微博来表达对“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“iphone”这一感兴趣话题的主要观点[183]。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用户主观性研究来说是很有意义的。

本章针对这一用户的主观性建模问题提出了主观性模型概念，使用一个框架将话题和观点结合起来。主观模型分为两部分，其中一部分是用户感兴趣话题分布，用于对用户对各种话题的兴趣度建模；另一部分是用户在每个话题上的观点分布，用于对用户对话题发表的多个观点信息集成建模。图 5.1展示了主观模型框架的总体结构，具体来讲，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（以 Twitter 平台为例，当然该框架也可以适用于其他社交媒体平台）中抽取出用户感

兴趣的话题；（2）使用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。

本章具体安排为：首先介绍相关工作，然后定义了用户层面的观点集成问题，接着给出主观模型定义以及构建方法，并以观点预测实际应用为例对主观模型进行了定性定量实验评测，最后是小节。

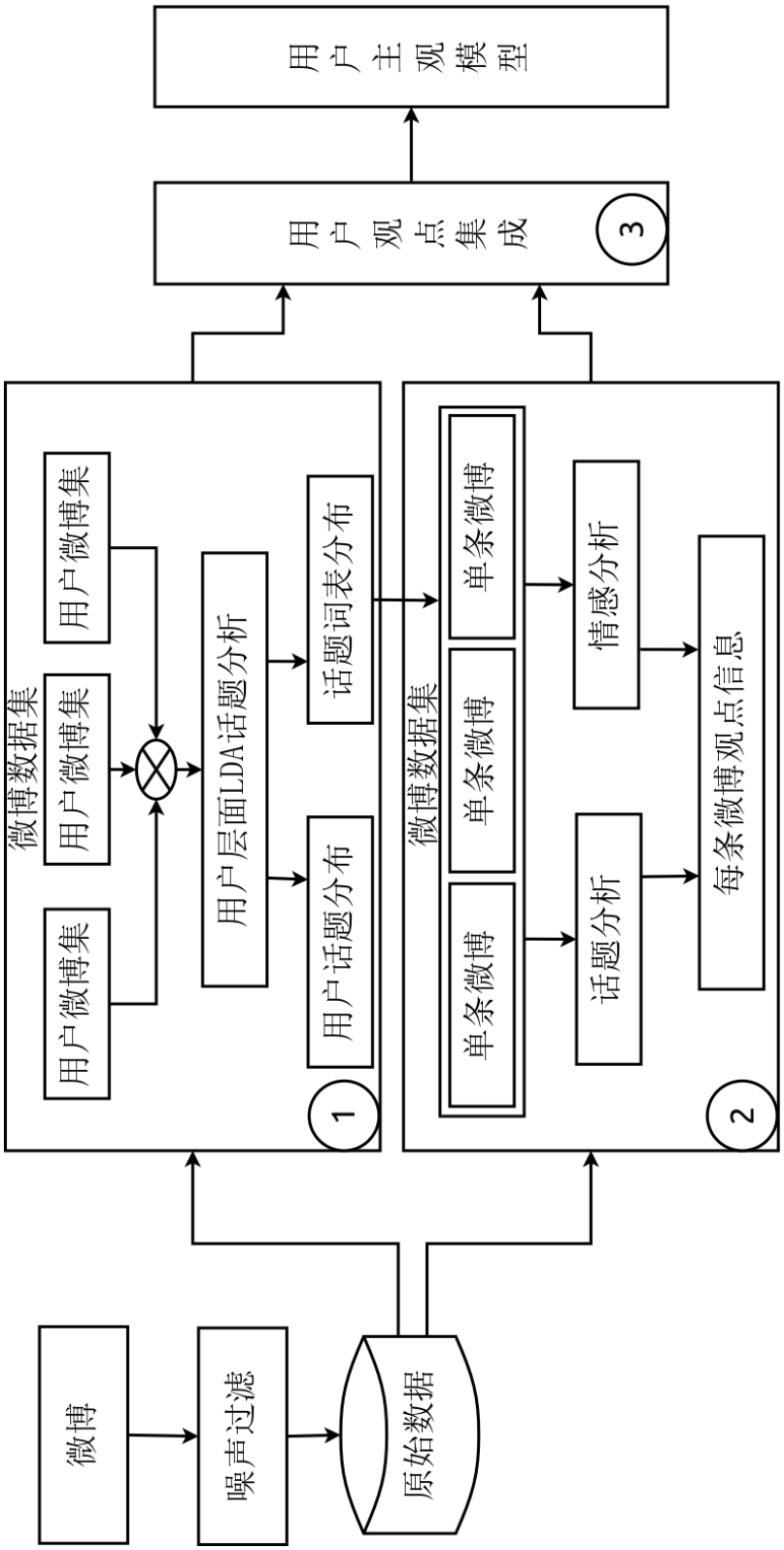


图 5.1　主观模型总体框架

5.2 相关工作

虽然观点挖掘（Opinion mining）研究最先是在商品评论（Review）和新闻评论（Comment）[23, 24] 兴起的，近年来越来越多的工作开始关注于 Twitter 等社交媒体短文本的观点信息，目前工作重点主要是针对单个短文本进行情感分析[127, 167, 168, 184, 185]，往往忽视了社交媒体的文本信息不是独立的，用户之间、数据之间以及用户与数据之间存在广泛联系。还有一些工作开始着眼于研究用户层面（User level）的主观性信息[186, 187]，研究还主要是识别用户发表观点针对的目标[188] 或是针对特定目标分析用户的情感极性[189]，而没有考虑到用户关心的多个话题、话题的各个方面（aspect）以及各方面观点的集成。自从 Blei 等[190] 对文本的话题分析引入潜语义话题模型（Latent dirichlet allocation，LDA），开始有各种基于 LDA 的扩展模型用于从大规模语料中抽取用户的话题[191, 192]，也有很多模型将情感分析与话题模型想结合设计话题情感模型（Topic-sentiment model），这种模型将情感极性与话题关联起来，代表性的主要有 Mei 等的 TSM 模型[193] 和 Lin 等的 JST 模型[194] 等，与本章提出的主观模型很接近，本章将与其进行定性和定量的分析对比。

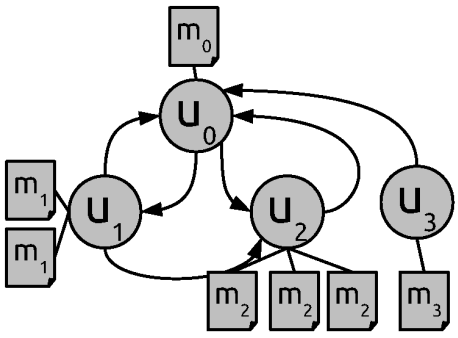
随着用户在社交媒体上发布信息的增多，研究者因此能够获得越来越多的数据对用户建模（User modeling），这些用户模型对于研究用户行为等研究具有促进作用。例如 Hannon 等[195] 首先提出使用 Twitter 的社交网络关系以及用户微博内容对 Twitter 用户进行建模用于分析用户的转发微博行为；Macskassy 和 Michelson[94] 使用 Wikipedia 作为外部知识库识别用户产生内容中的实体来对用户兴趣进行建模，并使用用户模型进行用户的分类研究；Ramage 等[196] 使用 4S（Substance，Status，Style 及 Social）维度利用话题模型对用户的微博及社交关系分析建模，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[197] 提出了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预测用户的发帖行为；Pennacchiotti 和 Popescu [198] 提出了一个综合各类信息对用户建模方法用于用户分类任务，确认了从用户产生内容中挖掘出的深层次特征的作用，方法反映了对用户及其网络结构的深入理解。

上述这些工作都证明了从用户自己发布的内容中挖掘关键信息的重要性，并且从四方面信息对用户进行建模，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”），但是少有工作关注于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，本章基于这一动机提出主观模型概念对用户的主观性建模。

5.3 观点集成问题

如在引言部分所介绍，用户在使用社交媒体平台的时候发布的信息是碎片化的信息，因为用户会在不同的时间就感兴趣的多个话题以及话题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信息，应该将他所有与特定话题相关的信息中的表达的观点进行综合才能确定用户的真正观点。为了满足对用户的主观性建模分析需求，在此提出观点集成问题（**Opinion Integration Problem** (OIP)）定义作为用户层面的观点分析的基础。如果站在信息消费者角度，信息使用者主要关注用户层面（User level）的观点信息，而不是单个微博层面（tweet-level）的观点信息，因为观点分析的最终目标是发现人的主观想法而不只是单条微博中的观点信息，对单条微博中观点信息分析只是对分析用户主观性的一个中间步骤。此外，很多情况下用户的单条微博中的观点因为受到长度限制以及上下文语境的缺失常常是不明确的，但是通过用户的所有微博就可以知道其明确的观点信息[185]。

本节所提出的话题相关的观点集成问题（OIP）可以用图 5.2进行说明。如图



	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。

中所示，假设 Twitter 上的一个异构网络（heterogeneous network）是由用户集合 $V = \{u_i\}$，用户之间的关系 $E = \{(u_i, u_j)|u_i, u_j \in V\}$ 以及每个用户 u_i 发表的微博集合 $M_i = \{m_i\}$ 构成，其中用户所关注的话题 $T = \{Topic_j\}$ 以及用户每条关于话题 $Topic_j$ 微博表达的观点 $o_{i,m,j}$ 可以从在网络中确定和抽取出来。于是观点集成问题可以定义为：

定义 (观点集成问题)：用户 u_i 对某一话题 $Topic_j$ 的观点 $O_{i,j}$ 不是他某条微博 m_i 表达的观点，应该是从他所有与话题 $Topic_j$ 相关的微博 $M_i = \{m_i\}$ 中通过某种方法 $f(o_{i,m,j})$ 集成得出的，即：

$$O_(i, j) = \sum_{m} f(o_{i,m,j})$$

观点集成问题有两个因素必须考虑：首先为了观点所针对目标话题的一致性，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）还是话题粒度都能够保持一致；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观点是需要考虑的重要问题。下一节将会提出一个主观模型的概念，可以很好满足以上两个要求。

5.4 主观模型

其实心理学已经对主观性进行了广泛的研究，并基于个人的历史行为和言论中定义其主观性，以表示独特个性[199]。在语言学上，语言中的主观性定义为作者在发布的文本中所表现出的自己立场、态度和情感[200]。社交媒体的出现为用户提供了能够针对感兴趣话题表达自己意见以展现自己独特主观性言论平台，因此在社交媒体平台上，用户的“**主观性**”可以定义为用户产生内容中涉及的话题和针对话题的表达的观点，因此主观性不但涉及到用户观点，也包含观点针对的目标。

本节首先给出主观模型的形式化定义以满足提出的用户层面观点分析需求。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极性分为“积极的(Positive)”或是“消极的（Negative)”。“积极的”情感表示该用户对话题支持或者喜欢，而“负面的”情感表示不支持或不喜欢。本节所提出的主观模型采用了更加通用（General）的“观点”定义，也就是用户针对某话题观点是一个在情感表示空间上的分布，该情感表示空间由可以表示情感强度的情感极性值构成。情感表达空间可以表示更细粒度的观点，因此可以更好的区分细致的观点差别，比如对话题持支持度为 8 的观点比支持度为 5 的观点更加具有“积极性”。其实对观点表示形式的定义还没有统一的标准，本节采用这种比较广义的定义是为了能使得主观模型能够更加通用。为了具体讨论问题，下面统一在 Twitter 平台对主观模型进行定义和分析，其实本章所提出的主观模型可以适用于其他的社交媒体平台。另外，之所以将模型命名为“**主观模型**”是因为是对社交媒体中用户产生内容中的主观性信息进行建模。

5.4.1 模型定义

以 $G = (V, E)$ 表示 Twitter 上一个异构社交网络，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（Following relationship)。对于每一个用户 $u \in V$，

对应的微博集合 M_u 表示其发布所有内容。假设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，以及一个情感表示空间 S 用于表示用户观点。对于用户 $u \in V$ 的“**主观性（subjectivity）**”，定义为用户所发表的所有微博 M_u 中所涉及的话题以及针对话题集成的观点。

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 $\{O_t\}$，观点用情感表示空间 S 上的情感分布表示。

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

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 对话题 t 的观点 O_t 指的是对话题所有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兴趣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值的分布表示，主要目标是为了研究用户层面的观点信息，获得用户兴趣和观点比较全面的理解。

5.4.2 主观模型的构建

根据主观模型的定义，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微博中经过计推理得出。然而对 Twitter 数据进行内容分析面临一些挑战：Twitter 上微博数量十分庞大，但是每条微博由于受限于 140 字的限制而相对短小，并且各种不规范的语言被广泛使用，缺乏大规模的标注数据等。这些都使得机器学习方法和自然语言处理技术很难以达到好的分析效果[201]。因此能够有效的对 Twitter 的微博内容进行建模分析需要一些能应对这些挑战并且尽量不使用需要标注数据的方法和技术。为了主观模型的通用性，在设计构建方法时主要考虑使用一些无监督的技术从用户微博中挖掘话题和观点信息，然后通过观点集成来构建用户的主观模型。因此提出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应对短文本微博造成的稀疏问题，并使用基于规则以及无监督方法解决无标注数据问题[202]。

5.4.2.1 话题分析

微博所涉及的话题一般是隐性的，需要从微博内容中经过推导得出。目前针对微博话题发现研究主要集中于定位关键词（Key words）[203]，抽取实体（Entity extraction）[204]，借助外部知识库（External knowledge categories）[94]，或者使用语义框架（Semantic framework）[205] 等方法。这些方法面临一个主要的问题是数据稀疏问题，因为在谈论同一个话题时，不同用户会使用各种不同的词汇来描述和表达。还有一个重要的方向是各种无监督话题模型，其中 LDA 话题模型[190] 及其各种扩展模型[206] 对微博中的话题分析更加有效[207]。LDA 模型的话题表示形式在概念上更加宽泛，每一个话题都表示为在所有词空间上的一个分布，因此可以有效应对话题表示的稀疏问题。主观模型的构建框架采用了用户层面 LDA 话题模型（User-level LDA）从用户所有的微博中发现隐性话题，对应于 LDA 话题模型，用户层面 LDA 话题模型生成过程可以使用图 5.3来表示。

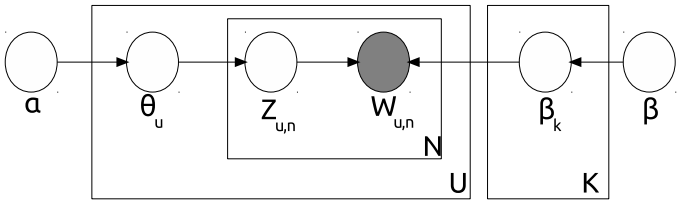


图 5.3　用户层面 LDA 话题模型

具体生成过程如下：

- 对每个用户 u，从先验中获取兴趣话题分布 $\theta_u \sim Dir(\alpha)$；
- 对用户微博中的每个词语 $w_{u,n}$，$n \in \{1, \cdots, N\}$：
  - 从用户兴趣话题分布中获取一个话题 $z_{u,n} \sim Multinomial(\theta_u)$；
  - 根据话题 $z_{u,n}$，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抽取出涉及的话题，用户产生的微博内容应该和 LDA 话题模型的文档对应起来。构建主观模型时，主要目标是为了了解用户感兴趣的话题而不是单条微博谈论的过细话题，所以我们将每一个用户的所有微博连接起来组成一篇微博文档作为 LDA 模型的输入文档。因此用户层面 LDA 模型中的一篇文档就对应于一个用户，一个用户感兴趣的话题可以使用在话题空间的一个多项式分布来表示，分布的权重可以和主观模型的话题权重相对应。形式化表示为：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）方法进行估计。构建框架中实际本章中我们使用的是基于变分推理的话题模型工具 Gensim[208]，该工具使用的是在线批处理模式的变分推理方法。

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，因此为了分析微博用户的主观性，需要了解用户每条微博表达的观点。前面三章内容详细介绍了社交媒体观点分析的各种方法，主要方法可以分为基于规则(词典）方法以及基于机器学习方法两类。如果要准确分析微博观点，基于机器学习方法训练过程需要大量标注数据，Twitter 的庞大数据量以及语言的动态性决定了很难对这样的数据进行标注，主观模型的构建优先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 的一些语言特点转化成为分析规则，因而更适用于 Twitter 的观点分析[169, 174]。

在基于规则的方法中，SentiStrength 是专门针对社交媒体中的不规范短文本数据进行情感分析的工具包[174]。SentiStrength 将对应于社交媒体语言特点的规则融合进了基于词典的方法，非常适用于主观模型对 Twitter 观点分析需求。SentiStrength 对每条微博情感分析后输出两个情感值：一个积极极性情感强度值 ($[1,5]$ 范围内) 和一个消极极性情感强度值 ($[-5,-1]$ 范围内)。SentiStrength 输出的情感值不是简单的积极和消极极性二值结果，而是细粒度的情感强度值，符合主观模型细粒度情感表示空间上分布需求。因此主观模型构建框架使用 SentiStrength 对所有微博进行情感分析。为了使用方便，将 SentiStrength 的两个输出结果映射为一个在 $[0,8]$ 的离散整数值表示情感强度，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 输出的积极极性情感强度值，n 代表消极极性情感强度值。与情感极性想对应，在 $[0,8]$ 情感表示空间中，强度值 4 和 5 表示中性(neutral）情感，强度值大于 5 表示积极极性，强度值小于 4 表示消极极性。使用 SentiStrength 对用户的每条微博进行情感分析的到一个在 $[0,8]$ 情感表示空间的情感值，将所有微博的情感值综合，就可以对用户的观点进行集成。

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。用户主观模型的构建过程可以分为三个步骤：

1. 在生成的话题模型中，参数 θ_u 可以直接对应到用户 u 在话题空间中的兴趣话题分布，可以确定用户感兴趣话题为：

$$Z_u = \{t|p(t|\theta_u(t)) > 0, t \in T\} \tag{5.3}$$

2. 将话题模型应用到用户 u 每条微博 m 确定涉及话题为：

$$Z_m = \{t|p(t|\theta, \beta, m) > 0, t \in T\} \tag{5.4}$$

3. 对用户 u 发布的所有涉及话题 t 的微博观点进行集成分析的到用户在话题上的观点：

$$O_t = \left\{ \frac{N_s}{\sum_{s \in S} N_s} | s \in [0, 8] \right\} \tag{5.5}$$

其中 $N_s = \sum_{m \in M_u} I(s_m)$（$s_m = s \& t \in Z_m$）表示用户发布的涉及话题 t 情感值为 s 微博数目。

最后综合形成用户主观模型：

$$P(u) = \left\{ \left(t, p(t|\theta_u), \left\{ \frac{N_s}{\sum_{s \in S} N_s} \right\} \right) | t \in Z_u, s \in S \right\} \tag{5.6}$$

对用户 u 构建主观模型 $P(u)$ 的详细过程如算法 5.1所示：

以上的构建方法中，由于微博话题的集中性，简单假设微博 m 的情感 s_m 是针对微博所涉及的所有话题 Z_m，并没有区分针对不同话题的不同情感。

5.4.3 与生成模型比较

在情感分析领域，一些研究提出了基于话题模型的话题情感模型，能够扩展基本的话题模型将文档中表达的情感与话题相结合统一建模[193, 194]。其中 TSM 模

算法 5.1 主观模型的构建过程

已知:

用户集合 V;

每个用户所发布的微博集合 M_u;

求:

为每个用户 u 构建的主观模型 $P(u)$;

使用用户层面的话题模型对用户微博内容分析获得模型 $P(\theta,\beta|M_u,V)$;
for all 用户每条微博 $m \in M_u$ **do**
    对 m 情感分析获得情感值 s_m;
end for
for 每个用户 $u \in V$ **do**
    用户感兴趣话题分布为参数 θ 对应的分布 θ_u;
    使用公式 5.3确定用户话题集合 Z_u;
end for
for 每条微博 $m \in M_u$ **do**
    使用公式 5.4对微博 m 话题分析，得到微博涉及话题 Z_m;
end for
for 用户每个兴趣话题 $t \in Z_u$ **do**
    for 情感表示空间的每个情感值 $s \in S$ **do**
        统计用户 u 发布的微博中情感值为 s 且涉及话题 t 的数目 N_s;
    end for
    使用公式 5.5计算用户 u 对话题 t 的集成的观点分布 O_t;
end for
构建用户 u 的主观模型 $P(u)$;
return $P(u)$

型（Topic Sentiment Mixture model）[193] 认为文档中表示情感等主观信息的词语与描述话题的词语是相互独立的，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建模，在文档的生成过程中任一词语的生成或者从话题语言模型中采样获得，或者从情感语言模型中采样获得，二者只能选择一个。JST 模型（Joint Sentiment/Topic model）[194] 提出了一种新的方式来分析文档中的情感信息，在话题模型抽取话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联合建模。这些模型在发现话题相关的情感极性时跟本章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主观性信息建模。

但是这类模型通常假设存在一个词语 -情感分布，需要通过学习获得这个通用词语 -情感分布来对文档中的情感知识建模，这对短小和不规范社交媒体语言，尤其是 Twitter 来说是很困难的。相对于话题的表达，情感、观点等主观信息更难识别，因为情感信息常常隐式的存在于一些微妙的语言表达方式中（比如反讽），并

且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语言的表达方式外，还有很多微博特有的语言来表达，比如表情符、字母大小写变化、不规范词语中字母重复强调以及惊叹号等标点符号的使用等等。微博上的这些语言特点表达出的情感很难用词语 -情感概率分布表示。但是基于规则的情感分析方法可以很容易通过设计规则反映这些特有语言特点，用规则获取微博语言中微妙的情感表达方式。因此，主观模型构建中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性的建模。

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应用到用户观点分析以及行为（转发、关注等行为）分析中。本节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。根据用户观点的一致性，假设用户不会就某一话题随机表达积极或消极极性观点，例如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[209]。因此得到了用户的主观模型，可以根据用户前期表现出的主观性预测用户就某一话题发表的微博所持观点。

首先将这种观点预测问题形式化为三元组 $< author, m, t >$，其中 $author$ 是微博 m 作者，微博涉及话题为 t。观点预测的目标就是通过计算得出用户 $author$ 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情感表达模式，然后利用这些模式来预测观点的极性。单条微博经常会由于缺乏上下文信息而使得观点具有模糊性，用户的主观模型是从用户所有的历史微博中构建的，因此有丰富的上下文信息，并且根据用户主观性的一致性假设，主观模型中对某一话题的观点比一条微博中的观点更加稳定一致。因此可以使用主观模型来提高用户未来发表微博中观点预测效果。具体来说，对微博 m，其作者的主观模型可以根据算法 5.1构建，假设作者新发布的微博为 m，m 的情感值通过某种方法比如 SentiStrength 分析得出为 s_m，并且 m 所涉及话题可以使用的公式 5.7推导得出：

$$\hat{t} = argmax(\hat{P}(t|\theta, \beta, Z_u)|t) \tag{5.7}$$

表 5.1　Twitter 数据集统计

项目	规模	项目	规模
总用户数	139,180	每个用户平均朋友数	14.8
总连接数	4,175,405	每个用户平均粉丝数	14.9
总微博数	76,409,820	每个用户平均微博数	549

用户 *author* 在话题 $\hat{t}$ 上的观点分布可以从主观模型 *P*(*author*) 中确定为 $O_{author,\hat{t}}$，是一个在情感表示空间 S 上的分布，因此可以计算出用户在话题 $\hat{t}$ 上归一化情感值：

$$\hat{s_m} = \sum_{i \in T} d_i * v_i \tag{5.8}$$

其中 v_i 表示情感值，d_i 表示情感值对应的分布。

于是可以使用微博情感分析得到的情感值 s_m 和主观模型计算出的情感值 $\hat{s_m}$ 联合进行观点极性 p 预测：

$$p = \begin{cases} positive & if \ \dfrac{\hat{s_m} + s_m}{2} > \dfrac{|S|}{2} + 1 \\ negative & if \ \dfrac{\hat{s_m} + s_m}{2} < \dfrac{|S|}{2} \\ neutral & otherwise \end{cases} \tag{5.9}$$

5.5　实验

5.5.1　数据集及设置

实验使用的数据集是通过 Twitter 公开 API 抓取的数据集[210]，数据集的具体规模如表 5.1所示。

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大量时空开销。根据社交网络的同质性（Homophily）[211]，也就是“物以类聚（Birds of a feather flock together)”的原则[212]，社交网络中连接紧密的用户更趋向于讨论相同话题，持有相似观点[213]。在 Twitter 上，用户之间的连接关系对应着用户间的认可或关注关系，还意味着拥有有相似话题或观点。因此可以利用社交网络的连接紧密用户形成的社区结构（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，这种利用社交网络结构特点化整为零的方法，可以降低构建主观模型的计算复杂度。实验使用 Igraph[1] 工具

[1] http://igraph.org/

包划分社区，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 模型进行话题分析效果比较差，因此实验中将这些社区和用户过滤，同时也过滤掉了发微博少于 5 的、每次发微博字数少于 5 个的还有微博内容只有 url 连接的用户总共 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。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，实验也将主观模型与两个话题情感生成模型 JST 和 TSM 进行了对比试验。所有模型的 Dirichlet 先验参数设为 $\alpha = 50/T$（T 为话题数目），$\beta = 0.01$。JST 的不对称情感先验参数 γ 依照经验设为 $(0.01, 1.8)$。JST 和 TSM 模型推导都经过了 2,000 次迭代 Gibbs 采样。

5.5.2 样例分析

为了定性展主观模型表达用户主管行的能力，在此给出了使用本章提出框架构建的一个用户的主观模型如图 5.4和 5.5所示。该用户发表了 533 条微博，所有微博可以用词云图[2] 5.4来展示。

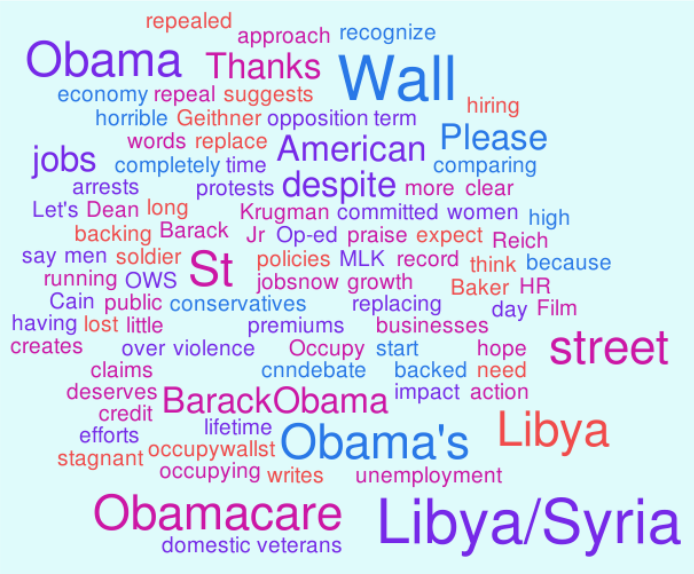


图 5.4　微博词云

图 5.5[3]是用户在 $[0, 100]$ 话题空间和 $[0, 8]$ 情感表示空间中可视化的主观模型。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”），图 5.5左侧展示了用户在每个话

[2]实验使用 TagCrowd (`http://tagcrowd.com/`) 生成词云图。

[3]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: $O_2 = (d_{u,2}(4) = 0.5, d_{u,2}(5) = 0.5)$, $O_{32} = (d_{u,32}(4) = 1.0)$, $O_{83} = (d_{u,83}(4) = 0.5, d_{u,83}(5) = 0.5)$。

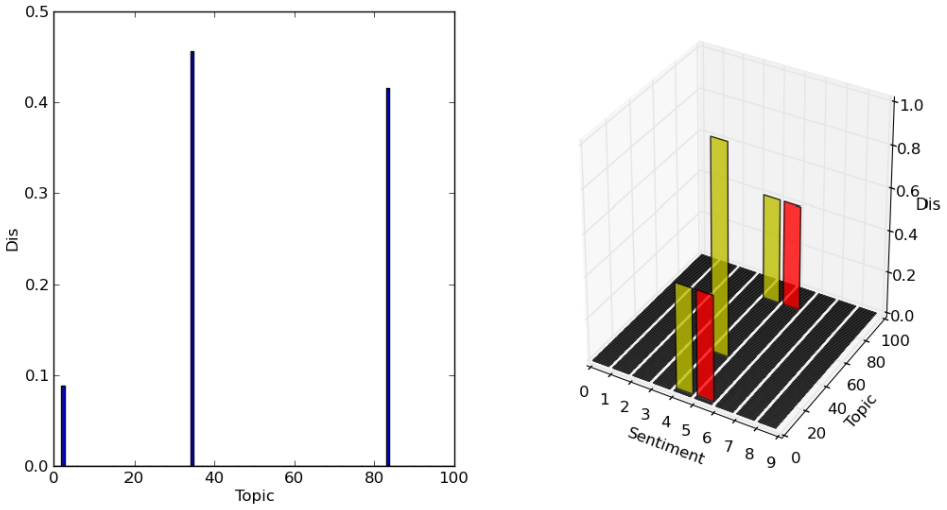


图 5.5　主观模型样例

题上的兴趣权重。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，属于中性偏积极极性。从这个样例可以看出，主观模型对用户的主观性进行了细致的建模，不但有用户的兴趣分布，也有细粒度的观点分布信息。

5.5.3　观点预测性能

为了定量评价主观模型的性能，主观模型在观点预测任务上与两个话题情感模型（TSM and JST）以及几个常用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，实验中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[176]。
- **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的在线工具。
- **STR:SentiStrength** 将微博中的一些语言特点转化成规则，并结合基于情感词典方法，专门针对微博等社交媒体短文本进行情感分析工具[174]。

实验从数据集中随机选择了 1,000 个至少有 80 条微博的用户，然后选择每个用户按照时间顺序发布的最后一条微博组成了 1,000 条测试数据集。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。

表 5.2　观点预测对比实验结果

情感分类方法	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%

相对于 OF 显著的性能提升使用 * 标记。

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情感分析工具，不适用于 Twitter 这种语言环境;
- 其次，两个无监督情感分析方法 Sentiment140 和 SentiStrength 因为是专门针对 Twitter 设计，准确率都明显好于 OpinionFinder（Sentiment140：70.45%，SentiStrength：69.98%）;
- 第三点，总体上两个话题情感生成模型 TSM 和 JST 的准确率都好于 OpinioFinder，并且准确率都比 Sentiment140 和 SentiStrength 方法稍好（但是不显著)，证明了将情感信息与话题分析相关联的重要性;
- 最后，主观模型（SUB）准确率在四种话题设置下准确率都显著地好于三个无监督情感分析方法，并且将主观模型计算出的情感值与 SentiStrength 对微博情感值想结合后，显著提高了 SentiStrength 准确率，与两个话题情感生成模型相比较，主观模型性能明显比 TSM 要好，稍好于 JST，这是因为主观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，针对用户层面的观点分析，定义并研究了社交媒体中用户的观点集成问题，提出了主观模型概念并进行可形式化定义，主观模型中定义了通用观点表示形式，使得主观模型可以在更细粒度的情感表示空间中对用户的观点进行集成，将集成的观点表示为情感表示空间的分布；提出了基于规则和无监督方法构建主观模型的框架，该框架采用新的算法从用户的历史微博中抽取话题和观点信息，得到的话题分布对用户的兴趣建模，将同一话题微博中的观点信息集成为一

个观点分布对用户在话题上的观点建模；使用真实 Twitter 数据对主观模型进行了定性和定量评测，实验结果证明，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 两个话题情感生成模型更适应 Twitter 上用户的主观性建模。

第六章　用户转发行为分析

6.1　引言

上一章介绍了如何从社交媒体用户产生内容中挖掘出话题和观点信息并在用户层面集成对用户的主观性进行建模，本章主要介绍如何应用主观模型对用户在信息传播中的转发行为进行分析。

信息传播通过逐步层叠式的信息扩散触发大量用户参与到信息的病毒式传播中，在市场营销、政治选举等应用场景中发挥着重要作用，引起了众多研究者，尤其是社交网络研究人员的广泛关注。社交网络研究为信息传播设计了一些通用的传播模型，可以进行模拟信息流动（Information flow）[214, 215] 以及探测信息瀑布（Information cascades）爆发[216]。但是这些模型都是将用户看作是网络中的一个简单节点，忽视了用户在信息传播过程中的行为自主性。作为社交媒体中的信息消费者和产生者，每个用户都可以在社交媒体上发帖和转贴以表达自己的兴趣和观点，能自主选择信息和传播信息。在社交媒体中一条信息能否得到广泛传播主要依赖于用户间的“口碑（Word of mouth）”效应，只有口碑效应好的信息才会引起广泛用户的兴趣对其进行传播，口碑效应取决于用户信息消费的主观意图，因此分析用户的主观意图可以对促进信息的传播研究。随着自然语言处理（Natural language processing）和数据挖掘（Data mining）技术的发展，社交媒体用户的主观意图可以使用用户自己产生的数据进行建模来分析。本章就是基于口碑效应机制问题，研究给定某个用户的一条微博，分析所有收到该微博用户中谁最有可能参与到该条微博的后续传播中。一个典型的场景是，在如图 6.1所示 Twitter 的一个异构网络中，用户 Tony 和关注他的所有朋友在以往的信息交流中讨论了两个话题:“苹果手机（Iphone）”以及电影“冰雪奇缘（Frozen）”，要研究的问题是:Tony 新发布了一条有关电影“冰雪奇缘”微博，如何判定他所有朋友中谁会转发传播这条微博?

不同社交网络平台上的信息传播行为是不一样的，本章主要分析 Twitter 用户的微博转发行为。庞大的用户群以及信息的指数级增长，使得 Twitter 在互联网的信息传播中扮演着重要角色。尽管 Twitter 微博在长度上受到限制，但是 Twitter 提供用户间转发机制为信息的快速传播提供了前所未有的便捷途径。据有关统计，Twitter 中有超过四分之一的微博是由用户间的相互转发形成的[93]，因此理解了用户的转发行为就能够很好的解释 Twitter 上的信息传播。

社交媒体用户是信息传播的参与者，同时也是个性化的信息消费和生产主体，用户很自然地会在信息的交互中表达出自己兴趣和观点，表现出主观性。在心理

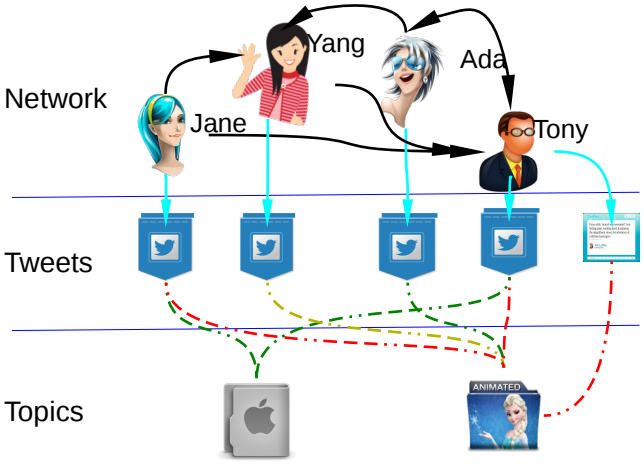


图 6.1　问题示意图

学的研究中证实了人的主观能动性（Subjective initiative）决定了人的主观性会影响自己的行为模式[217]，同样根据偏颇吸收（Biased Assimilation）理论，人总是趋向于选择跟自己偏执化观点（Biased opinions）相一致的信息进行传播[218]。因此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。转发行为分析研究提出了一些方法和模型来确定一些影响转发行为的因素[94, 219]。但是目前还没与相关工作关注到用户转发行为的主观动机。从 Twitter 的口碑效应来看，转发行为是一个连续的过程，包含着接收到微博，对微博内容评估，最后确定时候转发三个环节。三个环节中最重要的就是评估微博内容是否有价值的信息值得和朋友分享，也就是用户转发行为的主观动机。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。依据“物以类聚（Like attracts like）”原则，用户更容易转发那些能够迎合他个人口味的信息。以图 6.1中的例子来说明，每个用户对网络中谈论的两个话题所持观点可以从他们历史发布的微博中得出，假设 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 新发布的微博内容是对冰雪奇缘表达正面肯定的观点，应该可以确定 Jane 是最有可能转发这条微博的用户。从这种动机出发，本章主要研究用户的主观性是如何影响其转发行为。

为了研究主观性和转发行为的关系，需要回答两个问题：（1）怎样准确对用户的主观性进行建模？（2）怎么样从用户的主观性角度度量微博值得传播？上一章内容已经提出了使用观点集成方法来对社交媒体上用户主观性建模，并定义了社交媒体用户的主观模型及其构建方法。本章将继续使用主观模型概念，并针对用户的转发行为分析提出一个新的主观相似性计算方法来度量微博是否值得传播，针对影响转发行为的三个因素，即微博内容的吸引力、转发行为的社交需求以及

从众需求（Conformity needs）[220]，定义三个主观相似性度量三个因素对转发行为的影响。

6.2 相关工作

在微博转发行为分析方面，已经有大量工作在转发行为特征分析、提高微博转发性因素确定以及设计模型估计转发概率三个方面展开研究。例如 Suh 等 [221] 发现带有 Url 网络连接以及 hashtag 标记的微博更有可能被转发；Macskassy 和 Michelson [94] 发现从微博内容中推理得到的模型能够解释大多数的转发行为；Comarela 等[96] 发现与微博作者前期交互，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行为；Starbird 和 Palen [95] 特别针对危机发生时的微博信息转发机制进行了研究，发现有危机话题关键词的微博更有可能被转发；Osborne 和 Lavrenko [222] 通过引入一些特征，比如微博的新颖性和作者被加入朋友列表的次数，使用被动攻击算法（Passive aggressive algorithm）训练模型预测转发行为；Jenders 等 [98] 从微博及其作者的网络结构、信息内容以及情感信息分析了一些“显式”和“隐式”的影响转发的特征；Naveed 等 [223, 224] 引入了微博的趣味性指标，并使用表情符、情感以及话题等特征对趣味性指标进行量化来预测微博被转发的可能性；Feng 和 Wang [225] 构建了一个图模型，并将微博以及用户的所有信息组合到图的节点和边的信息里面，并提出了一个因子分解模型（Factorization model）对微博依据被转发的概率进行排序；Pfitzner 等 [226] 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的机会被转发；Luo 等[227] 设计了包括转发历史特征、用户特征、用户活跃时间特征以及用户兴趣特征四组特征集合对用户转发微博行为进行分析，并根据转发可能性对用户排序。

总体来说，上述所有工作主要是回答“哪些微博会否被什么样的用户转发”这样一个问题，但是忽略了用户在转发时的主观动机，也就是“站在用户角度，某条微博是否值得用户转发”这样的问题，本章将结合上一章提出的主观模型从用户用户的主观动机角度来分析转发行为。

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，也就是弄清楚用户喜欢什么和不喜欢什么（即用户感兴趣的话题和用户对话题所持的观点），这就是上一章为用户所建立主观模型的用途，通过用户主观模型可以清楚了解用户的主观性，为分析用户的转发行为分析提供信息基础。从技术角度来讲，上一章提出的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

历史数据中同时获得用户兴趣（对应的话题分布）和全面的观点（对应的观点分布）信息，为后续的一些应用提供信息支持。之所以主观模型是通用的，因为它不但将用户的兴趣和观点结合进一个整体框架，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表示空间上的概率分布。这个情感表示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。这种观点的表示形式一方面可以在细粒度的情感值空间区别不同观点，另外一方面可以以概率分布计算不同观点之间的相似性，能够准确区分观点和判断观点相似性是对用户主观动机分析的基础。本节主要定义主观相似性为转发行为的主观动机分析提供有效的度量手段。

6.3.1 主观相似性

构建主观模型框架是将话题分析和观点分析分开进行的。具体来讲，首先使用用户层面（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 $TM = (\theta, \beta)$, 其中 θ 表示用户在话题空间 T 的兴趣度分布，β 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，因此可以通过计算微博 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目进行归一化后获得用户在话题上的兴趣度：

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

至于观点的分布式表示，正如在图 6.1中的例子，虽然 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。情感分析研究主要是将观点表示为单一值，尤其是正负极性二值为主，并不区分针对话题的观点在不同方面的具体观点，也无法计算观点的大小顺序，比如那个用户更喜欢电影。在主观模型中，观点被定义为在情感表示空间 S 的概率分布，可以更精确的表示和区分观点。假设微博 t 通过

情感分析得出情感值为 s_t，用户在某一话题 k 上的观点分布可以将所有谈论该话题微博在每一个情感值上的数量归一化后获得：

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

为了量化“物以类据（like attracts like）”这样的效应，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上的相似性。首先定义在同一话题上两种观点的相似性计算方法。

6.3.1.1 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，分布的每一维都代表着在对应情感值上的观点权重。为了区分观点，可以定义情感表示空间中的情感值不独立，情感值之间按照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。这种情况下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。例如表 6.1所示的三个观点分布，代表在一个 $S = [0, 8]$ 情感表示空间中的三个观点：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

观点	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

使用常规的分布相似性计算方法（余弦相似性或 KL 距离），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现象，这是与实际不相符的，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，或者只是情感表示空间的一个距离值。为了准确计算观点之间的相似性，需要将观点在情感表示空间的距离和分布上的相似性结合起来，在此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} 维的情感值上的分布，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：

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

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

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

计算结果达到了与三个观点之间相似性的直觉理解一致的效果。

6.3.1.2 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，是两个用户之间感兴趣话题的交集；$\theta_u(k)$ 代表用户 u 在话题 k 上的兴趣度权重。

值得注意的是，当测量用户 u 在主观性上与用户 v 有多相似性时，话题权重使用的是用户 u 的话题权重，因此这个主观相似性度量方法是不对称的。之所以这样设计，是因为考虑到用户的主观上的相似性是个人主观判断，因此度量目标用户与自己主观想法上有多相似是根据个人的话题兴趣度以及观点相似性来确定的，不需要目标用户也作对称性的考量。因此在度量两个用户的主观相似性时，$Sim(SM_u, SM_v) \neq Sim(SM_v, SM_u)$。

6.3.2 转发行为分析

用户的转发行为受到多种因素的影响，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布的，因此用户的转发行为是因为社交需求；
3. 微薄内容是突发新闻或有趣段子，具有流行性，因此用户的转发行为是趋同需求（或称为从众需求，conformity needs）[220] 的结果。

这三种因素是用户产生转发行为的不同原因，从主观动机角度分析，可以使用三个主观相似性来量化这三个因素，从而对转发行为进行分析。

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 的所有关注者，当作者 u_a 发布微博 t 后，所有用户 $f \in F$ 都会看到微博 t，至于哪个用户会转发该微博，需要分析用户的主观动机。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，需要通过分析进行预测。

6.3.2.1 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% 的单值分布。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：

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

6.3.2.2 社交性度量

这种情形下，转发行为是基于用户的社交需求。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 t 的内容。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 并不是对关系亲密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”的微博感兴趣，但是对他的关于苹果手机微博不感兴趣。因此需要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，该权重由反应朋友类型和亲密程度的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 在微博接收的朋友圈中相对专家指数，专家指数越高的用户就会对其他用户有更多的影响力。

在此只是简单地根据用户 u_a 的发帖数量在朋友圈中所有用户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） $w_L(u_a)$ ：此处简单将用户的领导力影响定义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。该因子可以通过对 $Interation_{u_a,f}$ 使用用用户 u_a 和 f 所有的微博数目归一化计算获得：

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

综上所述，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 $\sum_{i=1}^{4} \lambda_i = 1$。本章中将其均衡设为 $\lambda_i = 0.25$。

6.3.2.3 流行性度量

用户在使用 Twitter 时，如果发现一条微博 t 是非常流行的（具有突发性、新颖性或传染性），在趋同效应或从众心理的作用下，用户很有可能会对其进行转发。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：

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

[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，需要对 $Sim(u_a, t)$ 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

实验使用了 Luo 等[227] 研究工作中使用的 Twitter 数据集[2]，在构建数据集时，作者使用 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 条历史微博。最后得到的数据集总共有 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影响，从数据集中采样抽取了 5,214 个没有转发目标为博的用户作为反例，与转发者一起构成平衡测试数据集。数据集的具体统计如表 6.2所示：

表 6.2 数据集统计

项目	数量
监控的目标微博数	500
每条目标微博的平均粉丝数	89
收集到的所有用户数	45,531
所有历史微博数	6,277,736
目标微博的所有转发者	5,214
目标为博所有未转发者	40,317

在构建主观模型过程与上一章一样，使用了 Gensim[208] 进行话题模型训练，话题数目设为 50,100,150 和 200；使用 SentiStrength[174] 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，采用了 Nielsen 等[228] 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时候与转发行为之间存在相关性，实验采用了 ANOVA（Analysis of Variance）[229] 假设性检验方法对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**”这一零假设（null

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

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

相似性指标		$Sim(f,t)$	$Sim(f,u_a)$	$Sim(u_a,t)$
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。后续实验将话题数目固定为 100 来进行讨论。

6.4.3　样例分析

为了定性说明主观模型在转发行为分析中的作用，首先用一个实际样例来进行阐述。在此从 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 6.2所示。

图 6.3显示了 14th 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 14th 号话题，话题是关于“love between people”，且作者对该话题的观点偏中性，可以看出这与图 6.2中微博的主观模型以及图 6.3中 14th 号话题词云图是一致的；微博作者有 188 条微博，主要集中在 14th 号话

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

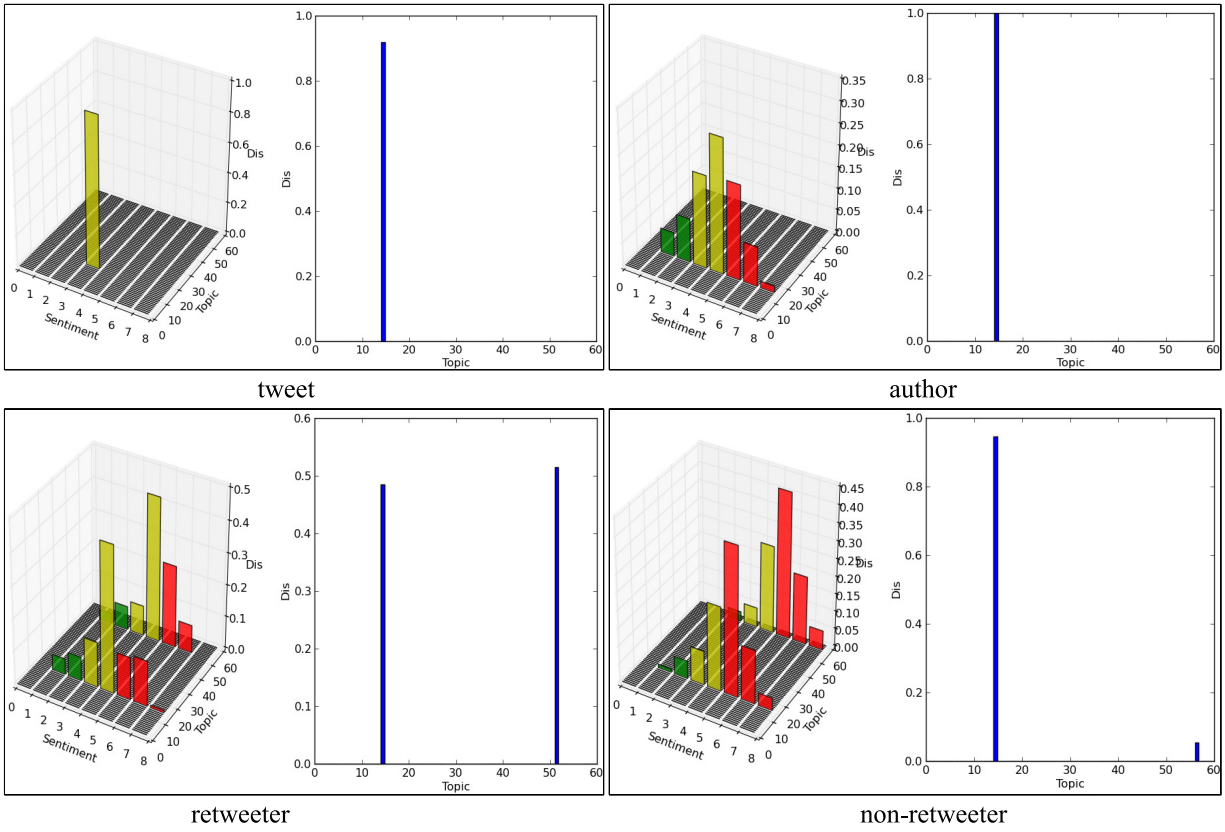


图 6.2　主观模型示意图

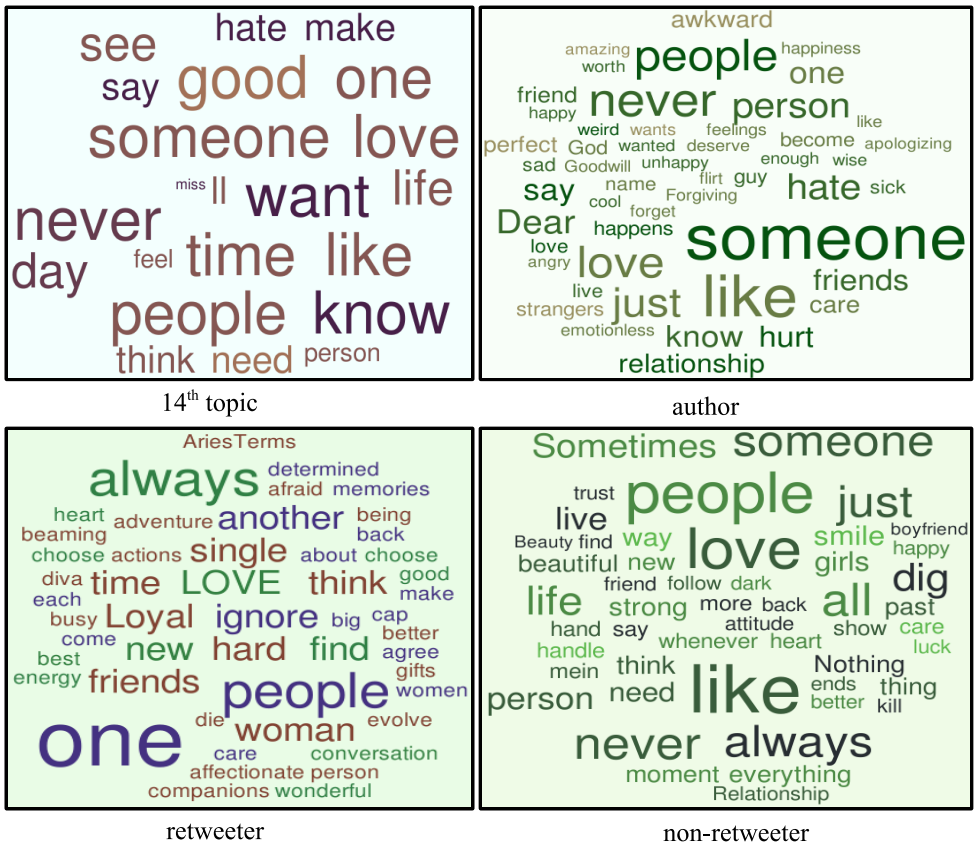
题，观点分布为 $O_{u_a}^{14} = (0, 0.04, 0.05, 0.25, 0.35, 0.25, 0.05, 0.01)$，偏中性；至于两个关注者，转发者有 196 条微博，主要涉及 14^{th} 号话题和 52^{nd} 号话题，话题分布比较均匀（其中 $w_{u_r}(14) = 0.48$），对 14^{th} 号话题，观点分布为 $O_{u_r}^{14} = (0, 0.02, 0.04, 0.15, 0.50, 0.13, 0.15, 0.01)$，偏中性；未转发者有 156 条微博，涉及到 14^{th} 号话题和 56^{th} 号话题，主要谈论了 14^{th} 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考虑三个引起转发因素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（度量微博流行性）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

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

从以上分析中看得出，对于微博的两个关注者，仅就感话题兴趣度来说，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



14th topic　　author

retweeter　　non-retweeter

图 6.3　14^{th} 号话题、微博作者与两个关注者词云图

$w_{u_a}(14) = 1.0$，未转发者 $w_{u_n}(14) = 0.98$，转发者 $w_{u_r}(14) = 0.48$），但是考虑到主观相似性，转发者因为与微博以及作者有更相似的观点分布因而主观相似性更高（与微博主观相似性（度量吸引力）$Sim(f,t)$:0.854 vs 0.805，与作者主观相似性（度量社交性）$Sim(f,u_a)$：0.967 vs 0.919），因此主观相似性的不同引发了他们不同的转发行为，从这个例子可以看出主观模型结合三个因素的主观相似性度量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，分三阶段进行了转发预测实验。

首先，将本章模型与其他基于话题的模型进行对比实验。对比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[204]，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，将本章模型与两个话题情感生成模型（generative topic-sentiment models）TSM 模型[193] 以及 JST 模型[194] 进行了对比实验。虽然 TSM 模型和 JST

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

模型也能同时对话题和话题相关的情感建模，但是他们的情感输出为正负二值极性。在训练这两个模型时，同样也是将用户所有微博作为一篇文档输入，并且使用本章定义的主观相似性计算方法 6.5来计算三个主观相似性，但是在实际使用中是将三个相似性值组合起来作为特征同时加入到分类器中评价模型的预测性能。

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 Twitter 使用习惯等元数据，也将本章模型和综合考虑其他因素的方法进行了对比。在此主要是和 Luo[227] 的工作（标记为“LUO”）进行了对比，表 6.5是 Luo[227] 的模型使用的一些特征。其中对用户兴趣建模部分，LUO 方法使用的是简单的词袋模型，为了验证本章提出的方法在预测转发行为时的作用，实验将 LUO 方法中的用户兴趣特征替换为本章的三个主观相似性指标作为特征进行组合实验（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。对于基准（baseline）设置，采用了一个基本的基准，该基准假设如果一个粉丝曾经转发目标微博作者的的微博，那么他很有可能会继续转发，因此直接将其预测为目标微博的转发者。准确率结果如表 6.6所示。

表 6.6　准确率评测结果

特征	准确率 (%)	特征	准确率 (%)
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。

从表中可以看出：

- 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%，准确率低的主要原因是微博中 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起来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%）准确率不如本章提出的模型，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%）。

综上所述，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

本章在第五章提出的主观模型基础上从主观动机角度对用户的转发行为的进行了分析，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微博吸引力、朋友间的社交性以及微博的流行性三个不同因素，对用户的转发行为进行分析，并将三个因素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，可以很好的解释和预测用户的转发行为，对于理解用户的转发行为的主观动机有重要作用。

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以指数级速度增长，本文主要围绕社交媒体中观点信息的挖掘、分析以及在转发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，本文充分利用了社交媒体作为媒体所产生文本的语言特点和社交媒体用户之间广泛连接的社交功能来帮助解决这些问题。

对社交媒体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。为了确定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，本文特别针对社交媒体的文本进行了情感知识词典的构建和对社交媒体非规范化文本的情感分类研究，因此可以从社交媒体中挖掘分析观点信息。在获得文本中的观点信息后，可以利用这些观点信息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。得到的主观模型可以对于理解用户的在线行为提供帮助。

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可靠性问题，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值，本文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将英文情感词典 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识别文本的观点信息会受到词典覆盖面以及领域适应性的限制，而且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词并扩展情感词典的方法，本文通过研究中文的连词语言规则和上下文统计特征，以实验验证了三种从语料中抽取词语并计算情感值的情感词典扩展方法，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。

第三，从社交媒体文本一般是不规范的短文本，从中确定观点信息需要对这种不规范短文本的情感倾向性进行分类，本文通过将情感分类问题形式化为特殊的文本分类问题，根据词语在表达情感极性时的不同作用，提出了特征空间划分

假设，将情感分类的词语特征空间划分为领域独立和领域依赖两部分，并使用现成的无须标注成语资源和远监督方式在不同的特征空间训练通用的分类器和微博情感分类器，将两个分类器用一个自举式机器学习框架组合在一起形成一个更强情感分类器，本文提出的方法在缺少大规模的标注文本而无法训练分类器的情况下，使用无监督方式达到了有监督机器学习方法的性能。

第四，用户在使用社交媒体时发表的文本一般是短小的、碎片化的，因此用户的观点信息散布在这些碎片化的文本信息中，目前情感分析研究主要是针对文本片段分析抽取其中的观点信息，无法完整呈现出一个用户整体的观点，因此本文提出了用户观点集成问题，并就这一问题提出了主观模型的概念，主观模型可以将用户在社交媒体中感兴趣的话题以及针对这些话题的发表的观点进行集成，并使用一种通用的细粒度的形式表示观点，将观点表示为在可扩展情感表示空间的一种概率分布，主观模型可以对用户在社交媒体中的整体观点信息集成表示，解决了用户信息的碎片化而造成的观点表示不全面不准确问题。

最后，针对信息传播研究中被忽略的用户的传播主观动机问题，结合主观模型对用户的主观性建模分析，本文主要分析了 Twitter 中用户在信息传播中的转发行为，将用户转发行为的主观动机量化为用户之间以及用户与微博之间的主观相似性，通过分析影响用户转发行为的三个因素，也就是微博内容上的吸引力，朋友间的社交需求以及微博信息的流行性，将其转化为三个主观相似性度量值，并分析研究了它们与用户转发行为之间的关系，在真实 Twitter 数据集上的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应用还有很多工作需要完成。在此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信息的实时性，目前虽然有一些研究工作，但主要都是围绕在 Twitter 中发现实时客观信息展开，包括新事件发现[230–236]、实时灾害报道（如地震、疾病、火灾等）[237–241] 等，在 TREC 评测中的 Twitter 检索[242–248] 也将实时性作为一个重要指标。本文的研究中，并未对观点信息挖掘与分析受到实时性的影响进行讨论研究，社交媒体语言的实时性特点需要后续工作中考虑到相关研究[249, 250]。

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]、

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

Flickr[3]等等。这些社交媒体肯定都有自己独特的特点，在这些类型社交媒体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，这就需要研究者在充分理解各种社交媒体的特点和用户对各种社交媒体不同使用习惯上，提出方法解决问题。

3. 目前将观点分析研究结合与其他应用和任务相结合是一个新的研究方向，主要是原因是越来越多的应用和任务需要以社交媒体用户观点信息作为有用的特征使用，比如在股市指数的预测中，用户的观点指标会影响人的投资意愿[18, 22, 251]，未来工作重点将结合更多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。

[3]https://www.flickr.com/

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] Tang J, Chang Y, Liu H. Mining Social Media with Social Theories: A Survey [J/OL]. SIGKDD Explor. Newsl. 2014, 15 (2): 20–29. http://doi.acm.org/10.1145/2641190.2641195.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

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

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

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

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

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

[46] 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.

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

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

[65] 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．

[66] 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.

[67] 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.

[68] 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.

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

[70] 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.

[71] 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.

[72] 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.

[73] 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.

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

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] Tsytsarau M, Palpanas T. Survey on mining subjective data on the web [J]. Data Mining and Knowledge Discovery. 2012, 24 (3): 478–514.

[84] 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.

[85] 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.

[86] 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.

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

[88] 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.

[89] 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.

[90] 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.

[91] 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．

[92] 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.

[93] 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.

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

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 知网 HowNet 评价词词典. 2013.

[105] 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.

[106] 情感词汇本体库. 2013.

[107] Choi Y, Cardie C. Adapting a Polarity Lexicon Using Integer Linear Programming for Domain-specific Sentiment Classification [C/OL]. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2. Stroudsburg, PA, USA, 2009: 590–598. http://dl.acm.org/citation.cfm?id=1699571.1699590.

[108] Du W, Tan S, Cheng X, et al. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon [C]. In Proceedings of the third ACM international conference on Web search and data mining. 2010: 111–120.

[109] Klenner M, Fahrni A, Petrakis S. PolArt: A robust tool for sentiment analysis [C]. In Proceedings of the 17th Nordic Conference of Computational Linguistics. 2009: 235–238.

[110] Carmen Banea R M，Wiebe J．A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources [C] // Nicoletta Calzolari (Conference Chair) B M J M J O S P D T, Khalid Choukri．In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)．Marrakech, Morocco，may 2008．http://www.lrec-conf.org/proceedings/lrec2008/．

[111] Gyamfi Y, Wiebe J, Mihalcea R, et al. Integrating Knowledge for Subjectivity Sense Labeling [C/OL]. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA, 2009: 10–18. http://dl.acm.org/citation.cfm?id=1620754.1620757.

[112] Building a fine-grained subjectivity lexicon from a web corpus [C/OL]. European Language Resources Association (ELRA), 2012.

[113] Wiebe J. Learning Subjective Adjectives from Corpora [C]. 2000: 735–740.

[114] Wiebe J, Mihalcea R. Word sense and subjectivity [C]. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. 2006: 1065–1072.

[115] Godbole N, Srinivasaiah M, Skiena S. Large-Scale Sentiment Analysis for News and Blogs. [J]. ICWSM. 2007, 7.

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

[117] Hu M, Liu B. Mining and Summarizing Customer Reviews [C/OL]. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2004: 168–177. http://doi.acm.org/10.1145/1014052.1014073.

[118] Rao D, Ravichandran D. Semi-supervised Polarity Lexicon Induction [C/OL]. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA, 2009: 675–682. http://dl.acm.org/citation.cfm?id=1609067.1609142.

[119] Yessenalina A, Cardie C. Compositional Matrix-space Models for Sentiment Analysis [C/OL]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA, 2011: 172–182. http://dl.acm.org/citation.cfm?id=2145432.2145452.

[120] Blair-goldensohn S, Neylon T, Hannan K, et al. Building a sentiment summarizer for local service reviews [C]. In In NLP in the Information Explosion Era. 2008.

[121] Velikovich L, Blair-Goldensohn S, Hannan K, et al. The Viability of Web-derived Polarity Lexicons [C/OL]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA, 2010: 777–785. http://dl.acm.org/citation.cfm?id=1857999.1858118.

[122] Remus R, Quasthoff U, Heyer G. SentiWS – a Publicly Available German-language Resource for Sentiment Analysis [C]. In Proceedings of the 7th International Language Resources and Evaluation (LREC'10). 2010: 1168–1171.

[123] 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.

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

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

[126] Garcia D, Schweitzer F. Emotions in Product Reviews–Empirics and Models [C]. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. Oct 2011: 483–488.

[127] 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.

[128] Strapparava C, Mihalcea R. Learning to Identify Emotions in Text [C/OL]. In Proceedings of the 2008 ACM Symposium on Applied Computing. New York, NY, USA, 2008: 1556–1560. http://doi.acm.org/10.1145/1363686.1364052.

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

[130] Miller G A. WordNet: A Lexical Database for English [J/OL]. Commun. ACM. 1995, 38 (11): 39–41. http://doi.acm.org/10.1145/219717.219748.

[131] Kim S-M, Hovy E. Identifying and Analyzing Judgment Opinions [C/OL]. In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg, PA, USA, 2006: 200–207. http://dx.doi.org/10.3115/1220835.1220861.

[132] Ahsaee M, Naghibzadeh M, Yasrebi S. Using WordNet to determine semantic similarity of words [C]. In Telecommunications (IST), 2010 5th International Symposium on. Dec 2010: 1019–1027.

[133] Ide N. Making Senses: Bootstrapping Sense-Tagged Lists of Semantically-Related Words [M] // Gelbukh A. Computational Linguistics and Intelligent Text Processing: 第 3878 卷. Springer Berlin Heidelberg, 2006: 2006: 13–27.

[134] Budanitsky A, Hirst G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures [C]. In IN WORKSHOP ON WORDNET AND OTHER LEXICAL RESOURCES, SECOND MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. 2001.

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

[136] Valitutti A, Strapparava C, Stock O. Developing Affective Lexical Resources [J]. PSYCHNOLOGY JOURNAL. 2004: 61–83.

[137] Takamura H, Inui T, Okumura M. Extracting semantic orientations of words using spin model [C]. In ACL '05. Stroudsburg, PA, USA, 2005: 133–140.

[138] Andreevskaia A, Bergler S. Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses [C]. In EACL'06. 2006: –1–1.

[139] Takamura H, Inui T, Okumura M. Extracting Semantic Orientations of Phrases from Dictionary [C/OL]. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. Rochester, New York, April 2007: 292–299. http://www.aclweb.org/anthology/N/N07/N07-1037.

[140] Firth J. A Synopsis of Linguistic Theory, 1930-1955 [J]. Studies in Linguistic Analysis. 1957: 1–32. Cited by 0956.

[141] Baron F, Hirst G. Collocations as Cues to Semantic Orientation. 2003.

[142] Stepinski A, Mittal V. A Fact/Opinion Classifier for News Articles [C/OL]. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA, 2007: 807–808. http://doi.acm.org/10.1145/1277741.1277919.

[143] Ding X, Liu B, Yu P S. A holistic lexicon-based approach to opinion mining [C]. In Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008: 231–240.

[144] Kanayama H, Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis [C]. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. 2006: 355–363.

[145] Kaji N, Kitsuregawa M. Automatic construction of polarity-tagged corpus from HTML documents [C]. In Proceedings of the COLING/ACL on Main conference poster sessions. 2006: 452–459.

[146] Kaji N, Kitsuregawa M. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. [C]. In EMNLP-CoNLL. 2007: 1075–1083.

[147] Qiu G, Liu B, Bu J, et al. Expanding Domain Sentiment Lexicon Through Double Propagation [C/OL]. In Proceedings of the 21st International Jont Conference on Artifical Intelligence. San Francisco, CA, USA, 2009: 1199–1204. http://dl.acm.org/citation.cfm?id=1661445.1661637.

[148] Learning Multilingual Subjective Language via Cross-Lingual Projections [C/OL]. Association for Computational Linguistics, 2007.

[149] Hoang L , Lee J-T , Song Y-I , 等 . Combining Local and Global Resources for Constructing an Error-Minimized Opinion Word Dictionary [M] // Ho T-B , Zhou Z-H . PRICAI 2008: Trends in Artificial Intelligence : 第 5351 卷 . Springer Berlin Heidelberg , 2008 : 2008: 688–697 .

[150] Lu Y, Castellanos M, Dayal U, et al. Automatic Construction of a Context-aware Sentiment Lexicon: An Optimization Approach [C/OL]. In Proceedings of the 20th International Conference on World Wide Web. New York, NY, USA, 2011: 347–356. http://doi.acm.org/10.1145/1963405.1963456.

[151] 杜飞龙. 知网辟蹊径共享新天地—董振东先生谈知网与知识共享 [J]. 微电脑世界. 2000.

[152] 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.

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

[154] 徐琳宏, 林鸿飞, 潘宇, et al. 情感词汇本体的构造 [J]. 情报学报. 2008, 27 (2): 180–185.

[155] 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.

[156] When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging [C/OL]. Association for Computational Linguistics, 2008.

[157] Jijkoun V, de Rijke M, Weerkamp W. Generating focused topic-specific sentiment lexicons [C]. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 585–594.

[158] 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.

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

[160] 一种情感词倾向性的分析方法. July 21 2010. http://www.google.com/patents/CN101782898A?cl=zh. CN Patent App. CN 201,010,133,149.

[161] 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.

[162] Yarowsky D. One sense per collocation [C]. In In Proceedings of the ARPA Human Language Technology Workshop. 1993: 266–271.

[163] Martin W, Al B, Van Sterkenburg P. On the processing of a text corpus: From textual data to lexicographical information [J]. 1983.

[164] Sinclair J. Corpus, concordance, collocation [M]. Oxford University Press, 1991.

[165] 张猛, 彭一凡, 樊扬, et al. 中文倾向性分析的研究 [J]. 第一届中文倾向性分析评测研讨会. 2008: 38–45.

[166] 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.

[167] 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.

[168] 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.

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

[170] 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.

[171] 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.

[172] 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.

[173] 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.

[174] 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.

[175] 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.

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

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

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

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

[180] 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.

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

[182] 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.

[183] 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.

[184] 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.

[185] 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.

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

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

[188] 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.

[189] 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.

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

[191] 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.

[192] 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.

[193] 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.

[194] 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.

[195] 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.

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

[197] 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.

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

[199] 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.

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

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

[202] 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.

[203] 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.

[204] 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.

[205] 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.

[206] 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.

[207] Hong L, Davison B D. Empirical study of topic modeling in twitter [C]. In Proc. of the First Workshop on Social Media Analytics. 2010: 80–88.

[208] 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.

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

[210] 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.

[211] 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:.

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

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

[214] 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.

[215] 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.

[216] 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.

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

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

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

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

[221] 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.

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

[223] 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.

[224] 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.

[225] 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.

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

[227] 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.

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

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

[230] 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.

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

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

[233] 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.

[234] 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.

[235] 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.

[236] 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.

[237] 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.

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

[239] 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.

[240] 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.

[241] 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.

[242] 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.

[243] 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.

[244] 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.

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

[246] 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.

[247] 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.

[248] 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.

[249] Das A, Gollapudi S, Munagala K. Modeling opinion dynamics in social networks [C]. In Proceedings of the 7th ACM international conference on Web search and data mining. 2014: 403–412.

[250] Guerra P C, Meira W, Jr, Cardie C. Sentiment Analysis on Evolving Social Streams: How Self-report Imbalances Can Help [C/OL]. In Proc. of the 7th WSDM. New York, NY, USA, 2014: 443–452. http://doi.acm.org/10.1145/2556195.2556261.

[251] Zhang X, Fuehres H, Gloor P A. Predicting stock market indicators through twitter "I hope it is not as bad as I fear" [J]. Procedia-Social and Behavioral Sciences. 2011, 26: 55–62.

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.（**EI 检索**）

[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.（**社交媒体领域重要会议，EI 检索**）

[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.（**EI 检索**）