分类号 TP391.1 学号 10069068

UDC 密级 公开

工学博士学位论文

社交媒体中的主观性信息挖掘与分析

博士生姓名 谢松县

学科专业 计算机科学与技术

研究方向 自然语言处理

指导教师 王挺 教授

国防科学技术大学研究生院

二〇一四年十月

Opinion Mining and Analysis in Social Media

Candidate: Xie Songxian

Supervisor: Professor Wang Ting

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha, Hunan, P. R. China

October 17, 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：社交媒体中的主观性信息挖掘与分析

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：社交媒体中的主观性信息挖掘与分析

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

摘 要 i

ABSTRACT iii

第一章 绪论 1

1.1 研究背景 1

1.1.1 社交媒体 1

1.1.2 主观性信息 4

1.2 研究问题 8

1.3 相关研究 9

1.3.1 Twitter 与自然语言处理 9

1.3.2 信息检索与机器学习 11

1.3.3 Twitter 中的传播分析 12

1.4 研究内容与方法 13

1.4.1 本文研究内容 13

1.4.2 本文研究方法 15

1.5 本文主要贡献 16

1.6 本文结构 17

第二章 应用语义关系自动构建情感词典 20

2.1 引言 20

2.2 词典资源简介 20

2.2.1 HowNet 语义词典 20

2.2.2 WordNet 语义词典 21

2.2.3 SentiWordNet 情感词典 21

2.3 基于语义关系的情感词典构建方法 21

2.3.1 词语抽取和义原抽取及语义分析 24

2.3.2 情感值的查询与计算 24

2.3.3 词语情感值计算 26

2.4 实验及结果 27

2.4.1 评价指标 27

2.4.2 性能评测结果 27

2.5 小结 28

第三章 基于语料资源的中文情感词典扩展 29
3.1 引言 29
3.2 问题描述 30
3.3 数据集及预处理 30
3.4 基于语言特征的情感词典扩展 31
3.4.1 连词选择 31
3.4.2 基于连词的极性计算 32
3.4.3 实验 32
3.5 基于统计特征的情感词典扩展 33
3.5.1 统计特征选择 34
3.5.2 基于上下文的情感词极性计算 34
3.5.3 实验 34
3.6 基于混合特征的情感词典扩展 36
3.6.1 基于混合特征的情感词极性计算 36
3.6.2 实验 37
3.7 小结 37
第四章 无监督的自举式情感分类 40
4.1 引言 40
4.2 相关工作 41
4.3 问题的形式化 42
4.4 无监督的情感分类框架 44
4.4.1 通用情感分类器 45
4.4.2 微博情感分类器 45
4.4.3 分类器的组合 46
4.4.4 分类器算法 48
4.5 实验 49
4.5.1 实验描述 50
4.5.2 实验结果 51
4.6 小结 52
第五章 Twitter 中信息传播者的发现 53
5.1 引言 53
5.2 相关工作 54
5.3 基于排序学习的 Twitter 信息传播者发现框架 54
5.3.1 Twitter 信息传播者发现排序学习框架 54
5.3.2 Twitter 信息传播者相关特征 55

5.4 转发历史特征 55
5.5 用户特征 56
5.6 用户活跃时间特征 57
5.7 用户兴趣特征 57
5.8 Twitter 信息传播者发现实验 57
5.8.1 Twitter 信息传播者发现实验数据 57
5.8.2 Twitter 信息传播者发现实验设置 58
5.8.3 Twitter 信息传播者发现基准系统（Baseline） 58
5.8.4 Twitter 信息传播者发现实验结果及分析 59
5.9 小结 61
第六章 总结与展望 62
6.1 工作总结 62
6.2 工作展望 63
致谢 65
参考文献 66
作者在学期间取得的学术成果 81

表 目 录

表 1.1 Alexa 统计访问量前十名网站 2
表 1.2 社交媒体的类型 2
表 2.1 T=0.05 时的性能对比 28
表 3.1 数据集及词典资源 30
表 3.2 各个领域性能评测结果 32
表 3.3 各个领域性能评测结果 37
表 4.1 结果对比表 51
表 5.1 Twitte 信息传播者发现实验数据统计信息 58
表 5.2 Twitter 信息传播者发现特征概况 59
表 5.3 基于不同特征组的 Twitter 信息传播者发现系统实验结果 60
表 5.4 基于不同特征的 Twitter 信息传播者检索系统实验结果 61

图 目 录

图 1.1 形形色色的中英文社交媒体 1
图 1.2 本文研究框架 14
图 1.3 论文整体结构图 18
图 2.1 HowNet 中概念的定义方式 21
图 2.2 基于语义关系的情感词典解决方案 23
图 2.3 词语和义原抽取处理流程 24
图 2.4 义原情感值计算过程 25
图 2.5 不同 T 值时的性能指标 28
图 3.1 语料预处理记录格式 31
图 3.2 Hotel 语料评测结果 35
图 3.3 Book 语料评测结果 36
图 3.4 NoteBook 语料评测结果 36
图 3.5 Hotel 语料评测结果综合比较 38
图 3.6 Book 语料评测结果综合比较 38
图 3.7 NoteBook 语料评测结果综合比较 38
图 4.1 自举式学习框架 47
图 4.2 λ 值的确定。 51

摘　要

现在的互联网上社交媒体随处可见，这给信息检索和传播分析工作带来了机遇与挑战。本文主要围绕在社交媒体中如何找到重要的信息以及信息是如何传播的展开。我们将 Twitter 作为研究对象，因为它是目前最著名的社交媒体之一，并且数据是公开的。这样从隐私的角度考虑，获取研究数据变得容易且能很好的为研究任务（如信息检索）服务。

信息检索的主要任务是在文档集合中，找到与给定话题相关的客观文本或主观文本。Twitter 是一个丰富的包含各种话题及其评论信息的资源库，本文将探讨如何在 Twitter 中找到相关的信息。但是 tweet 的短小化和非正式的文本特点，使得 Twitter 中的检索不同于以往的检索任务（如，网页检索）。本文将通过研究 tweet 文本特点和特有的 Twitter 社交媒体属性帮助 Twitter 检索。另外，Twitter 中信息的传播是一种普遍现象且与消息的质量相关（帮助 Twitter 中检索高质量的信息）。因此，我们从 tweet 本身和用户的角度，研究哪些因素影响了 tweet 的转发和人的转发行为。

我们的工作主要有四个部分：(1) 利用结构化信息的 Twitter 检索；(2) Twitter 观点检索；（3）Twitter 中传播观点的发现；（4）Twitter 中信息传播者的发现。四个工作具体如下：

利用结构化信息的 Twitter 检索： *Twitter 检索是在 Twitter 中找到与给定话题相关的 tweet 的任务*。绝大部分的 Twitter 检索系统在构造检索模型时一般都认为 tweet 是一个平面文本，但用户在编辑 tweet 时的一些习惯使得 tweet 文本呈现结构化的特点。这种结构化是通过一些不同的文本积木块组合而成，积木类型具体包括平面文本、主题词、链接、提及等。每一种积木都有自己独特的本质，一系列积木的排序组合又反映了一定的话语转换。以往的研究发现，通过开发文本的结构信息能够帮助结构化文本的检索（例如，网页检索）。本工作通过积木结构开发 tweet 的结构化信息，以此帮助 Twitter 检索。我们利用积木及其排列组合开发了一系列特征，并将其应用到排序学习的框架中。我们发现利用结构化 tweet 的方法进行检索能够达到目前最好的 Twitter 检索方法效果，将结构化 tweet 的方法和其他社交媒体特征一起使用能够进一步提高 Twitter 的检索效果。

Twitter 观点检索： *观点检索是在数据中找到对指定话题表达正面或反面观点的 tweet 的任务*。人们几乎在 Twitter 中表达了任何话题的观点，使其成为一个丰富的观点资源库。但是 Twitter 中也存在大量的垃圾信息和各种不同类型的文

本，使得 Twitter 中的观点检索充满挑战。我们提出了如何利用 tweet 的社交媒体信息和文本结构化信息的方法帮助 Twitter 的观点检索。特别的，基于排序学习，我们发现 tweet 的用户信息（如用户包含朋友的数目）、tweet 文本本身的结构信息和观点化程度影响着 tweet 的排序结果。实验结果表明社交媒体信息能够帮助 Twitter 的观点检索。基于无监督学习评价 tweet 观点化程度，并以此开发特征形成的检索方法能够到达手工标注 tweet 的有监督方法的检索效果，且这种方法能够帮助观点检索中话题依赖问题的解决。最后，我们在重新标注的 TREC Tweets2011 数据集上进一步验证了我们 Twitter 观点检索方法的有效性。

Twitter 中传播观点的发现： Twitter 已经变成人们收集观点做出决策的重要资源，但是数量众多且差异巨大的观点严重影响了人们使用这些资源的效果。本文我们考虑了*如何在 Twitter 中找到传播观点的任务—tweet 不仅表达了对某些话题的观点，且这个 tweet 在未来会被转发*。利用排序学习模型，我们开发了一系列特征，具体包括 tweet 的传播度特征、观点化特征和文本质量特征。实验结果证明了我们开发的特征对于 Twitter 中传播观点的发现是有效的，并且将所有特征整合的方法在发现效果上能够显著优于 BM25 方法和 Twitter 观点检索方法。最后，我们发现我们的方法在预测观点传播上可以达到人预测的水平。

Twitter 中信息传播者的发现： Twitter 和其它社交网络中一个重要的交流机制就是消息传播 —人们分享其他人创建的消息。虽然目前有许多工作研究了 Twitter 中的 tweet 是如何传播的（转发），但是一个未解决的问题是到底**谁**会转发给定的 tweet。这里我们考虑了*在 Twitter 中给定一条 tweet，发现作者的粉丝中谁会转发*。利用排序学习模型的框架，我们设计了一些特征，包括用户历史的转发信息，用户自身的社交媒体特征，用户使用 Twitter 的活跃时间，以及用户的个人兴趣。我们发现经常转发和提及作者的粉丝和与作者有相同兴趣爱好的人最有可能成为信息传播者。

通过以上四个问题的研究，我们发现 tweet 的文本信息和 Twitter 的社交媒体特征能够帮助 Twitter 信息检索和传播分析。

关键词: Twitter; **信息检索**; **观点检索**; **传播观点**; **信息传播者**

ABSTRACT

Social Media is now ubiquitous on the internet, generating both new possibilities and new challenges in information retrieval and propagation analysis. This thesis focus on finding important information and propagated information analysis in Social Media. We take Twitter as our research subject, since it is one of the most Social Media and public by default, which makes the data less problematic from a privacy standpoint, far easier to obtain and more amenable to target applications (such as information retrieval).

The main tasks in information retrieval are finding related objective or subjective documents about some topics in collection. Twitter is rich resource which contains information about various topics and opinions. Here we investigate how to find these information in Twitter. However, Twitter retrieval is different from traditional retrieval tasks (e.g, web search), since the text of tweet is short and informal. In this study we exploit textual features of tweet and the social media features to improve Twitter retrieval. Additionally, information dissemination is a prevalent phenomenon in Twitter and is related to the quality of message (which can help finding high quality information in Twitter). Therefore, from the point of view of tweets and users, we study the factors which affect tweet retweeting and users' retweeting behavior.

Our work can be divided into four parts: (1) improving Twitter retrieval by exploiting structural information, (2) opinion retrieval in Twitter, (3) finding propagated opinion in Twitter, (4) finding retweeters in Twitter. We introduce the four work in detail as follows:

Improving Twitter retrieval by Exploiting structural information. *Twitter retrieval deals with finding related tweets about some topics in Twitter.* Most Twitter search systems generally treat a tweet as a plain text when modeling relevance. However, a series of conventions allows users to tweet in structural ways using combination of different blocks of texts. These blocks include plain texts, hashtags, links, mentions, etc. Each block encodes a variety of communicative intent and sequence of these blocks captures changing discourse. Previous work shows that exploiting the structural information can improve the structured document (e.g., web pages) retrieval. In this study we utilize the structure of tweets, induced by these blocks, for Twitter retrieval. A set of features, derived from the blocks of text and their combinations, is used into a learning-to-rank scenario. We show that structuring tweets can achieve state-of-the-art performance. Our

approach does not rely upon social media features, but when we do add this additional information, performance improves significantly.

Opinion retrieval in Twitter. *Opinion retrieval deals with finding relevant documents that express either a negative or positive opinion about some topics*. Social Networks such as Twitter, where people routinely post opinions about almost any topic, are rich environments for opinions. However, spam and wildly varying documents makes opinion retrieval within Twitter challenging. Here we demonstrate how we can exploit social and structural textual information of tweets and improve Twitter-based opinion retrieval. In particular, within a learning-to-rank technique, we explore the question of whether aspects of an author (such as the number of friends they have), information derived from the body of tweets and opinionatedness ratings of tweets can improve performance. Experimental results show that social features can improve retrieval performance. Retrieval using a novel unsupervised opinionatedness feature achieves comparable performance with a supervised method using manually tagged Tweets. Topic-related specific structured Tweet sets are shown to help with query-dependent opinion retrieval. Finally, we further verify the effectiveness of our approach for opinion retrieval in re-tagged TREC Tweets2011 corpus.

Finding Propagated opinions in Twitter. Twitter has become an important source for people to collect opinions to make decisions. However the amount and the variety of opinions constitute the major challenge to using them effectively. Here we consider the problem of *finding propagated opinions – tweets that express an opinion about some topics, but will be retweeted*. Within a learning-to-rank framework, we explore a wide spectrum of features, such as retweetability, opinionatedness and textual quality of a tweet. The experimental results show the effectiveness of our features for this task. Moreover the best ranking model with all features can outperform a BM25 baseline and state-of-the-art for Twitter opinion retrieval approach. Finally, we show that our approach equals human performance on this task.

Finding retweeters in Twitter. An important aspect of communication in Twitter (and other Social Networks) is message propagation – people creating posts for others to share. Although there has been work on modelling how tweets in Twitter are propagated (retweeted), an untackled problem has been **who** will retweet a message. Here we consider the task of *finding who will retweet a message posted on Twitter*. Within a learning-to-rank framework, we explore a wide range of features, such as retweet history, followers

status, followers active time and followers interests. We find that followers who retweeted or mentioned the author's tweets frequently before and have common interests are more likely to be retweeters.

Based on the study of four work above, we find the textual information of tweet and social media features in Twitter can help Twitter retrieval and propagation analysis.

Key Words: Twitter; Information Retrieval; Opinion Retrieval; Propagated Opinion; Retweeter

第一章　绪论

1.1　研究背景

1.1.1　社交媒体

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

图 1.1　形形色色的中英文社交媒体

社交媒体中的互联网用户不但是单纯的信息接收者，也已经成为信息的制造者，人们通过社交媒体平台进行交流获取和产生信息。目前，中国拥有 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中可以看出，根据互联网网站流量信息公司 Alexa[2]统计，流量前十的互联网网站中社交媒体平台占了绝大部分。

那么什么是社交媒体呢？作为社交媒体的维基百科是这样定义的[3]：

[1] http://www.cnnic.cn/hlwfzyj/hlwfzzx/qwfb/201408/t20140825_47878.htm
[2] www.alexa.com
[3] http://en.wikipedia.org/wiki/Socialmedia/

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**
2	**Facebook.com**	7	**Amazon.com**
3	**Youtube.com**	8	**Twitter.com**
4	Yahoo.com	9	**Qq.com**
5	Baidu.com	10	**Taobao.com**

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.

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
Wiki	Wikipedia，Scholarpedia
Blogging	Blogger，LiveJournal，WordPress，博客
Social News	Digg, Mixx，Slashdot
Micro Blogging	Twitter，Google Buzz
Opinion & Reviews	ePinions，Yelp
Question Answering	Yahoo! Answers，百度知道
Media Sharing	Flickr，Youtube
Social Bookmarking	Delicious，CiteULike
Social Networking	Facebook，LinkedIn，MySpace

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格式的数据，包括文本、图像以及视频等。Kaplan 和 Haenlein[1] 从数据和信息流动角度讨论了社交媒体。首先从媒体部分（media），社交媒体中最突出的方面是它区别于电视，广播和报纸等传统媒体。在传统媒体中，信息的流动是从几个内容生产者到众多用户消费者。在社交媒体中，内容产生的权利转移到了传统媒体的消费者的手中，而且信息流动的方式更加不确定。内容消费者和生产者可以多次在瞬间改变自己的角色。另外，为什么我们称这种新媒体为社会化的（social）媒体。社会化意味着信息内容不只是由个体用户产生，更是与其他用户的协作产生。因此内容（content）变得更加多样化，因为社交媒体不只是用来产生和传播的内容，也为用户互相通信交流提供了便利。社交媒体中的用户产生内容（UGC，User-Generated Content）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是小的，但是数量巨大的用户群体以及社会化特性将用户的数据链接在一起产生了一种新形式的大数据。比如平均每天有超过 300 万条的微博（tweets）发布到 Twitter，每分钟有超过 3000 张照片上传到 Flickr，每年有超过 160 多万的博客（blogs）发表。

- **广泛链接（linked）：**社交媒体的社会化特性使得社交媒体的数据天生就是广泛链接的。比如用户产生的内容往往由于用户之间的各种社交关系链接在一起。这种链接的数据显然不是独立同分布的（IID，independent and identically distributed），与传统的数据挖掘和机器学习提出的基本假设相违背[3, 4]。

- **充满噪声（noisy）：**社交媒体中普通用户作为内容消费者和产生者常常使得社交媒体的数据质量参差不齐，充满噪声[?]。并且不仅于此，社交媒体中的网络结构也是充满噪声的，一是存在这一些传播虚假和无用信息的用户[5]，二是建立关系的方便性使得各种社会关系混杂在一起，比如好朋友和一般认识人[6]。

- **非结构化（Unstructured）：**社交媒体中用户产生的数据一般是高度非结构化的。尤其是随着移动互联方式的普及与来越多的用户使用移动设备更新 Facebook 的状态，发送微博，或者回复别人的帖子，这不但导致了文本内容短小，而且错误拼写频繁出现[7]。还有一些非自然语言的广泛使用，比如表情符（:)，:() 和缩写（h r u?）等[8]。

- **不完整性（Incomplete）：**为了用户的隐私保护，社交媒体平台一般允许用户将其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[9]。因此社交媒体的数据是极度不完整和稀疏的。

社交媒体的迅速普及与壮大，使得它在政治、经济、教育、社会等多方面发挥着越来越重要的作用。就如社会会随着网络而演化，无论对个人还是商业组织，社交媒体数据也逐渐变得越来越重要。Web 2.0 时代的到来，更是由于广泛的用户参与而使得用户产生内容（user-generated content (UGC)）成指数级的爆炸式增长，并且更庞大和复杂。目前常见的社交媒体数据应用有：一是基于用户个人信息、行为、位置、微博等数据而进行的个性化推荐、交叉推荐、品牌监测等营销类大数据应用，被互联网广告、电子商务、微博、视频、相亲等公司普遍采用。第二，公共服务类大数据应用，即不以盈利为目的、侧重于为社会公众提供服务的大数据

应用。典型案例如谷歌开发的流感、登革热等流行病预测应用能够比官方机构提前一周发现疫情爆发状况。国内也有搜索引擎公司提供诸如春运客流分析、失踪儿童搜寻的公益大数据服务。三是积极借助外部数据，主要是互联网数据，来实现相关应用。例如，金融机构通过收集互联网用户的微博数据、社交数据、历史交易数据来评估用户的信用等级；证券分析机构通过整合新闻、股票论坛、公司公告、行业研究报告、交易数据、行情数据、报单数据等，试图分析和挖掘各种事件和因素对股市和股票价格走向的影响；监管机构将社交数据、网络新闻数据、网页数据等与监管机构的数据库对接，通过比对结果进行风险提示，提醒监管机构及时采取行动；零售企业通过互联网用户数据分析商品销售趋势、用户偏好等等。

一些服务商拥有海量用户数据的大型互联网企业以自有社交媒体大数据资源为支撑，以 SaaS 形式为用户提供服务。典型的服务如谷歌和 Facebook 的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等。国内百度推出的大数据营销服务“司南”就属于此类。同时，政府也是社交媒体数据的积极使用者，2013 年曝光的棱镜门事件显示出美国国家安全部门在使用社交媒体数据应用的强大实力，其应用范围之广、水平之高、规模之大都远远超过人们的想象。2012-2013 年间美国国家安全局（NSA）、联邦调查局（FBI）及中央情报局（CIA）等联邦政府机构大量使用 Facebook，Google，Twitter 等公司的数据对全球进行监控引起全世界的关注。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的影响[4]。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。今年世界杯的主要赞助商之一可口可乐就首次尝试 iBeacon 在世界杯营销中的运用。为此，可口可乐挑选了粉丝在 Facebook 和 Twitter 上分享的照片，印制在足球场大小的旗帜上，并将在开幕式上展示。百威在圣保罗开设了社交媒体工作室。该工作室将从不同国家挑选“影响力人物”制作视频，并发布至网上。

1.1.2 主观性信息

语言学家 Lyons[10] 将语言功能分为描述（descriptive）的, 社交（social）的和表达（expressive）的。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主观性的信息（subjective information）。主观性信息，也可以称为观点，是人们在语言中表达对于谈论的目标事物的态度、情感或者看法[11]。观点常常简化为人对目标的同意或不同意（或者认为目标好或者坏，或持正面（positive）态度还是负面（negative）态度）[12]。

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

实际上，别人是怎么想的（或者大众观点）在各种决策过程中是必不可少的信息。比如在商业领域，消费者在选择产品的时候常常需要知道其他人的对这些产品的观点，而商家为了市场营销也常常需要知道这些观点。在政治领域，投票者受到其他人关于候选人的观点的影响，同时，大众的观点也会影响到政策制定者的决策。以往为了获取大众观点，需要进行问卷调查. 社会化网络的出现（Social Web），为人们提供了新的内容共享服务，使百万计连接到全球资讯网（World Wide Web）的人能够在时间和代价上更高效的方式创建和共享自己的内容，思想和观点。现在大家可以在各种社交媒体发表和表达自己观点：比如消费者可以在 Amazon[5], Yelp[6], 以及 TripAdvisor[7]上发表各种商品和服务的评论；用户可以在 Twitter[8]和 Facebook[9]上对最新话题表达自己的观点。社交媒体上巨大的用户群以及由他们产生的海量信息成为了发现人们对各种话题所持观点的宝贵资源。随着社交媒体用户产生内容的增多，人工地去发现和总结这些主观性信息是低效和难以全面的，因此需要计算机能够自动对这类主观信息进行分析和挖掘，于是文本的情感分析 (sentiment analysis，或者称为观点挖掘，opinion mining) 技术应运而生。情感分析技术[13] 是对带文本中有情感色彩的主观性信息进行分析、处理、归纳和推理的过程，其目标是自动发现和区分对某目标的情感和观点，目标可以是命名实体、也可以是话题或事件。近年来，情感分析（观点挖掘）研究逐渐发展成为介于自然语言处理（Natural Language Processing (NLP)）与自然语言理解（natural language understanding(NLU)）之间的一个独立领域。不像其他的自然语言处理任务（文摘或文本分类），观点挖掘主要处理与自然语言概念相关的语义信息和情感信息的推理而不需要对给定文本的深度分析[14]。

在进行情感分析时，Liu[15] 将观点形式化定义为观点五元组 $(e_i, a_ij, s_ijkl, h_k, t_l)$。其中，$e_i$ 是目标名称，a_ij 是目标的不同属性（方面，aspect），h_k 是持有观点的用户，t_l 表示时间，s_ijkl 是观点的情感值。Kim 和 Hovy 也对观点做了定义[16]：由四个元素组成：即主题 (Topic)、持有者 (Ho1der)、陈述 (Claim)、情感 (Sentiment)。

目前的研究通常把情感分成两类（即正面和负面）。其中正面类别是指文本表达积极的态度，而负面类别是指文本表达消极的态度。文本情感分类目前被广泛应用于电影评论、产品与服务评价、产品推荐、舆情分析、信息过滤、智能化搜索等方面。上一部分描述了社交媒体内容的特点，由于社交媒体语言的这些特性，

[5]www.amazon.com
[6]www.yelp.com
[7]www.tripadvisor.com
[8]www.twitter.com
[9]www.facebook.com

使得观点挖掘的研究遇到了一些挑战，通常需要用到模式识别、信息检索、机器学习、统计学、自然语言处理等多学科的交叉知识进行解决。

情感分析有别于传统的话题分类。话题分类关心的是文本所阐述的话题，如文档是属于教育类还是娱乐类的，而情感分析主要用来识别文档中表达的观点、喜好、立场和态度等主观信息，需要理解词汇的含义，词性，甚至句法和篇章结构等信息。在传统的话题分类中，主题词是最重要的特征，而在情感分析中，情感词是最重要的特征。情感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，情感分析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倾向，比如：“轻薄” 在通常意义下具有负面的情感倾向，如 “举止轻薄”，而在电脑领域，“轻薄” 却表示褒义。

2. **语境依赖性:** 某些词语具有多个词性，并且不同的词性常常呈现出不同的情感倾向。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倾向，前者是形容词，具有正面的情感倾向，后者是名词，不具有情感倾向。

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

情感词典，是人们在表达观点时常用的一些词语，是主观性信息的最明显的证据（比如 “好”)。无论对无监督还是监督方法，因此一部好的情感词典是从文本信息中发现主观信息的重要特征。近年来从网络数据中发现主观性信息变得越来越重要，能够获得大家对一些事物、人或事件的主观性态度比仅仅只有百科式描述更重要，比如对产品的问卷调查，政治选举的民调以及商业广告效果分析等。因此很多研究者开始注意到这种信息需求，并试图从网络数据中挖掘并分析主观性信息。然而大多数工作主要关注与于本中情感倾向的总体以及详细的分析，并且仅仅是对某些特定领域的文本数据，比如产品评论或博客，并不适用于其他类型数据。随着社交媒体的普及，用户可以更方便地发表与自己有关的信息，比如自己的生活和对周围食物的观点，这些越来越多的用户产生数据使得主观性信息的挖掘和分析变得更重要。受到这种趋势和研究需求的影响，TREC 评测[10]从 2006 年开始就有关于从网络信息中挖掘观点的评测，并且受到很多自然语言处理研究

[10]http://www.trec.state.tx.us/

人员的关注。判断一片文档是否有主观性信息最简单直观的方式就是看看是否含有主观性词语，这种基于词语的判断方式的基本假设就是文档中含有主观性词语通常是表达作者的主观观点，比如在商品评论中出现“喜欢”通常表示作者喜欢这个商品。因此很多研究者研究了一些通过人工或自动方式产生带有主观情感倾向的词典。主要有两种类型的网络上的信息（即，事实和观点)。然而，目前的搜索引擎（如谷歌)，都是为了表达同主题关键字的事实。Wiebie[17] 将带有个人心理视角的文本称为主观性信息，相对于客观地叙述事件或描述虚构世界的文本。Dave 等[18] 提出了观点挖掘作为从网络数据中发现主观性信息的研究。观点是所有人类活动的中心因为观点是我们行为的关键影响因素，我们对现实世界的感知和看法是建立在他人是怎么看这个世界的基础上的，当我们做决策之前通常会寻求其他人的观点。随着社交媒体的出现并流行（评论，论坛，博客，微博，评论，以及在社交网络中的帖子等)，带有主观性信息的数据爆炸式增长，因此我们对观点信息的获取不再需要靠传统的问卷调查等手段。然而发现并综合各种社交媒体中观点信息并不是意见容易的事，因为数据量非常大之外，人工阅读并发现主观性信息是不可能的，因此需要自动观点挖掘技术。近年来，在社会化媒体带有主观性信息的贴子在我们的现实生活中帮助重塑了企业，并左右公众的情绪和情感，这对我们的社会和政治制度深刻影响。这样的帖子对于鼓动群体运动引起政局变化具有很大作用，比如 2011 年发生在一些阿拉伯国家的阿拉伯之春运动。因此收集和研究在网络上的意见成为一种必然。情感分析应用已经普及到几乎所有可能的领域，从消费产品，服务，医疗保健和金融服务等社会活动和政治选举。除了现实生活中的应用，很多应用为导向的研究论文也已发表。例如，情感信息可以预测产品的销售量[19], 电影的票房[20–23]，股市走向[24, 25]，政治选举的结果[26–30] 等等。尽管语言学和自然语言处理已经有很长的研究历史，但是直到 2000 年才开始进行观点挖掘和情感分析等主观性信息研究。从那以后，该研究成为了非常活跃的研究领域。原因主要有：一是具有广泛的应用领域，二是提供了很多以前从未遇到的具有挑战性的问题，本文将系统的定义和讨论这些问题，并描述目前一些主流的解决方法；三是由于社交媒体的出现，我们第一次拥有了海量的具有主观信息的数据，如果没有这些数据，主观性信息研究是不可能的。毫不奇怪，情感分析的开始和快速增长与社会媒体是一致。事实上，情感分析正处在在社交媒体研究中心。因此，情感分析研究不仅对 NLP 有着重要的影响，而且还可能对管理学，政治学，经济学和社会科学产生了深远影响，因为它们都受到人的主观性的影响。

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，如何帮助人们利用平台更好地交流，并且在这些社交媒体中发现有意义的信息变得越来越重要。而信息检索技术是满足以上需求的重要技术手段。信息检索是在文档集合中，找到与给定话题相关的客观文本或主观文本。它能够帮助人们在海量的社交媒体信息中，快速找到相关内容，帮助有意义的信息发现，以此满足人们的需求，方便人际之间的交流。另外，以 Twitter 为代表的社交媒体的一个显著特点，就是信息的传播性。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。另外，Twitter 中已有的研究发现，转发的信息往往意味着高质量的信息[31]，这是基于人们在 Twitter 传播行为上的一个基本假设：当人们认为一个 tweet 非常重要且值得和大家分享此信息时，他们将通过转发传播这个 tweet。因此研究社交媒体中信息传播的内在规律变得十分重要，且可以帮助在 Twitter 中检索高质量的信息。

但是 Twitter 中的信息检索与传播分析任务也存在着挑战。由于 Twitter 客户端使用的多样性，如大量使用移动平台，以及 tweet 文本本身 140 个字符的限制，造成 tweet 文本与其他文本（如新闻）编辑质量与风格的巨大差异。同时移动平台的广泛使用，使得 Twitter 中信息传播速度更快，范围更广。再加上 Twitter 用户参与的低门槛性，使得信息在 Twitter 中的传播不像以往媒体（如报纸）的新闻传播，会对信息的正确性进行层层验证，这就造成了 Twitter 中信息传播的随意性，使得信息的质量难以保证。因此本文主要从两个科学问题来思考与研究 Twitter，以此帮助 Twitter 中的信息检索与传播分析：

1. **人们在 Twitter 中如何用自然语言描述话题和表达观点？**
2. **以 Twitter 为代表的社交媒体有何新特点？如何利用这些特点帮助获取信息和对信息进行传播分析？**

第一个问题的研究主要是从自然语言处理的角度分析人们在社交媒体上如何组织语言来描述客观话题和表达主观观点。显然 Twitter 参与的低门槛特点使得大量的用户参与其中，由于参与者编辑文本的水平参差不齐，编辑的内容与目的也多种多样，另外，再加上 tweet 本身的字符限制都使得 Twitter 中的文本呈现低质量、噪音大的短文本特点，这给传统的以正式文本（如新闻）为处理对象的自然语言处理技术带来了挑战。因此深入地研究 Twitter 文本的特点，对于解决 Twitter 中的信息检索与传播分析任务变得十分重要，并且分析 Twitter 中文本的特点也能够帮助其他以语言为基础的 Twitter 应用，如 Twitter 中的事件发现、观点挖掘等等。

第二个问题的研究主要是分析 Twitter 作为新型媒体的特点，从中发现一些规律和有价值的信息帮助 Twitter 中信息检索与传播分析问题的解决。显然 Twitter 中无论是 tweet 本身还是 tweet 的用户都呈现了一些新的特点。比如，tweet 中包含 hashtag，以此可以作为 tweet 的内容主题。tweet 中也包含大量的链接，这些链接与 tweet 中描述链接的文本存在什么样的关系？另外，每个 tweet 都有作者，Twitter 的一个显著特点就是用户信息的公开化，包括用户的朋友关系，发布信息的历史，个人的属性信息等，如何发现这些用户信息的普遍规律与内在价值，以此帮助 Twitter 中的信息检索与传播分析任务是本文研究的主要问题，当然社交媒体的新特点研究同样也能帮助 Twitter 中其他任务的解决，如用户推荐等。

1.3 相关研究

无论是 Twitter 中的信息检索还是传播分析，对 tweet 文本的理解都是其中一个重要的环节。但是 tweet 文本的短小与大量的噪音文本（存在着许多缩写词、错别字等等）都造成自然语言处理技术在 Twitter 上的应用存在着新的挑战，我们在本节将介绍自然语言处理技术在 tweet 文本处理上的相关工作。另外，Twitter 上信息检索的研究也离不开以往传统信息检索技术的借鉴与应用，我们主要利用 tweet 的文本特征与社交媒体的特征帮助 Twitter 中的信息检索，而这些特征如何有效地整合到检索模型中是需要考虑的问题，因此本节将详细介绍基于机器学习的信息检索技术。最后本节还将介绍已有的 Twitter 中的传播分析工作，以此为本文具体的传播分析任务的解决提供帮助。

这里要强调的是，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。

我们从整体上介绍了三个相关研究，包括 Twitter 与自然语言处理（见 1.3.1），主要介绍目前已有的自然语言处理技术在 Twitter 中的研究成果，以此帮助本文的关于 tweet 的文本分析；信息检索与机器学习（见 1.3.2），主要介绍目前机器学习技术在信息检索中的应用，以此为解决本文的具体 Twitter 信息检索和传播分析任务提供问题解决框架；Twitter 中的传播分析（见 1.3.3），主要介绍目前已有的关于 Twitter 转发的研究成果，为本文 Twitter 传播分析的具体任务提供借鉴。

1.3.1 Twitter 与自然语言处理

随着以 Twitter 为代表的社交媒体的广泛使用，自然语言处理技术在 Twitter 的文本处理中得到广泛应用，但是研究者发现 Twitter 的文本明显区别于以往的很

多文本类型。Eisenstein 将这种文本类型称为**坏语言**（*bad language*）：文本“无视”我们以前期望的词汇、拼写和语法[2]。

研究人员发现最先进的自然语言处理技术在 Twitter 的文本应用中都显著差于其他文本。在自动地词性标注测试中，Stanford tagger 在 Wall Street Journal 语料上的正确率可以达到 97%[32]，而 tweet 的文本处理仅仅只有 85%[33, 34]。利用 CoNLL 数据训练的 Stanford 命名实体识别器，对 CoNLL 测试语料进行实体识别，F1 值可以达到 86%[35]，而在 Twitter 的文本中仅仅只有 44% 的 F1 值[36]。另外，Foster 等人也对语法分析效果进行分析，发现最先进的语法分析器在 Twitter 的文本应用中，正确率下降约 15%[37]。

为了解决自然语言处理技术在 Twitter 中所遇到的挑战，研究人员主要从两个方面进行了相关研究:

1. **文本的正常化（Normalization）**，即把坏语言变成好的语言，以其适合于传统文本的自然语言处理技术。Han 和 Baldwin 开发了一个分类器，能够识别“非正常（ill-formed)”的词，然后利用基于形态音位（morphophonemic）相似的方法将其转换为正确的词[38, 39]，Han 等人还提出了一种构造词典的方法，简单替换词的变形（例如 tomorrow 替换 tmrw)，这种方法结合词语的上下文评估各种变换的可能性[40]，但 Liu 等人提出一种没有明确分类的方法，进行词的正常化[41]。另外，Liu 等人提出了一种基于图模型的方法同时解决 tweet 中命名实体识别和 tweet 文本正常化的方法[42] 。Liu 等人设计一个正常化认知驱动系统解决 Twitter 中文本的正常化问题[43]，该系统整合人们对于“非正常（ill-formed)”词的各种认知角度，包括字符转换、视觉感知、字符串和语音相似等等。最近 Hany 和 Menezes 提出了一种无监督学习的方法对 Twitter 中的文本进行正常化，他们在大量 tweet 文本中构造 n 元词串，以此构造语境相似的二部图，然后利用 Random Walks 算法发现“非正常（ill-formed)”词与正常词的对应关系[44]。以上所有的方法都在一定程度上解决了 Twitter 中文本正常化的问题。

2. **领域化（Domain adaptation）** 与其让 Twitter 的文本适应以前的自然语言处理技术，不如改变这些技术适应 Twitter 文本。一系列的工作从领域化的角度出发进行了相关研究。这些工作包括适合 Twitter 文本的自动词性标注器[33, 34]，自动命名实体识别的方法[36, 42, 45–49]，语法分析器[37]，对话模型[50]，自动摘要[51–59] 等等。这些工作采用各种方法，使其能够很好地适应 Twitter 文本的特点，具体为:

预处理（preprocessing） 减少词语中某些重复的字符（经常有些词用重复的字符表达感情[60]），去掉 hashtag、链接、提交（@username）等等；

标注新数据（new labeled data） 根据任务在 Twitter 中标注部分数据[36, 45]，以此进行有监督学习；

自定义标注标准（new annotation schemes） 定义适合 Twitter 的标注标注，如词性标注中对 hashtag、链接、提交（@username）等定制特定的标注类型[33, 34]；

“远端”监督（distant supervision） 通过一定的规则，构造大量粗糙的训练数据帮助 Twitter 的文本机器学习模型训练，然后应用到具体的任务中[36]。

毫无疑问，tweet 文本的特殊性使得传统自然语言技术在 Twitter 上的应用充满挑战，我们将利用以上所涉及到的方法、思想或已有的开发工具，按照信息检索任务和传播分析任务的具体需求，设计对应的 tweet 文本自然语言处理方法，开发有效的文本特征，提高 Twitter 中的信息检索效果与传播分析的准确性。

1.3.2 信息检索与机器学习

Twitter 中的信息检索是本文的一个重要研究任务，由于 tweet 文本的特点和丰富的社交媒体属性使得 Twitter 中的信息检索不同于以往的信息检索任务（如图书馆文档检索)。在 Twitter 检索任务中需要考虑因素很多，如 tweet 用户的信息等。传统的检索模型在构造排序函数的时候往往只需要考虑不多的因素，如查询词在文档的频率、位置等，因此可以手工构造这些函数对文档排序，但是 Twitter 中的检索需要考虑的因素相当多，造成手工构造排序函数变得复杂，但是基于机器学习的排序模型可以通过训练数据自动构造排序函数，因此这里我们详细介绍基于机器学习的信息检索模型。

信息检索与机器学习领域有很多研究的重叠，上世纪 60 年代提出的相关反馈就是一个简单的机器学习算法，它构建一个分类器区分相关文档和非相关文档，以此作为用户关于初始排序中文档重要性的反馈[61]。到了 80、90 年代，研究人员开始使用机器学习方法来基于用户反馈学习排序算法。但是，许多机器学习算法在信息检索上的应用都受到训练数据规模较小的影响，如果系统要对每个查询构造分类器，基本上是不现实的。

但进入 21 世纪，随着网络搜索引擎的出现，从用户交互中积累了海量的查询日志，潜在的训练数据的规模非常庞大。借此基于点击流数据的排序学习算法被提出[62–69]。由于对于每个查询中文档的相关性判断非常稀疏，但是有一定数量文档在网络搜索引擎的检索返回结果中被用户点击浏览，这些行为可以隐性地认为

是用户对文档相关性的判定。例如，如果一个用户在一个查询的排序中点击了第三个文档而不是前面两个，那么可以假设第三个文档应该在下次排序中获得较高的排序位置。

在排序学习模型中，最著名的排序函数莫过于基于支持向量机（SVM）的方法，通常被称为 Ranking SVM。它的输入是针对一组查询的偏序排序信息的训练集合：

$$(q_1, r_1), (q_2, r_2), ..., (q_n, r_n)$$

其中 q_i 是一个查询，r_i 是所需排序的文档关于查询的部分排序信息或相关性级别。这意味着如果文档 d_a 应该比 d_b 排序更高，那么 $(d_a, d_b) \in r_i$。这些排序信息可以通过点击流数据获得，然后训练排序模型。相关的研究从排序学习的数据[70–73]、排序学习模型[74–77] 和评估学习效果[78–82] 三个方面展开。

本文将主要采取基于排序学习的机器学习算法，针对 Twitter 信息检索的具体问题，将 tweet 文本分析的结果和 Twitter 社交媒体的新特点转换成特征，整合到机器学习的模型框架中，帮助 Twitter 中各项检索任务的解决。

1.3.3 Twitter 中的传播分析

Twitter 中一个重要的机制就是转发，即重新发布其他人发布过的 tweet。这种简单的机制可以使得作者的全部粉丝看到转发的信息，使得信息迅速、广泛的传播。我们本节将介绍 Twitter 中已有的对于转发行为的研究，以此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。这些研究成果可以帮助本文具体的传播分析任务。

boyd 等人[11]研究了 Twitter 中转发的各种类型以及转发的原因，他们分析了不同用户，用户属性，用户交流方式对于转发的影响，同时也分析了人们在 Twitter 中喜欢转发的内容[83]。他们发现 18% 的转发 tweet 包含 hashtag，52% 的转发 tweet 包含链接，11% 的转发 tweet 包含连续的转发符号串（如，“RT @user1 RT @user2 ”），另外，9% 的转发 tweet 都包含回复原 tweet 作者的回复字符串（“@reply”）。这说明 tweet 文本中的 hashtag，链接、回复、提交和转发符号都与 tweet 的转发存在着一定的对应关系。

Yang 和 Counts 通过 Twitter 中的提及（“@username”）抽取了用户之间的关系，并在此基础上构造了用户关系的复杂网络。他们研究了信息在这个复杂网络上是如何传播的，包括信息传播的速度，规模，以及范围[84]。他们发现大约只有

[11]Danah Boyd 因为家庭的原因一般使用小写拼写姓名，这里并不是拼写错误。

25% 的 tweet 是被信息作者的朋友转发，大部分是被粉丝但非朋友转发。这说明 Twitter 中用户形成的复杂网络，影响着人们的转发行为，因此信息在传播路径上具有一定的规律可循。

Macskassy 和 Michelson 分析了一个月用户的 Twitter 数据，他们解释了各种信息传播的方式，尤其是转发的行为模式，他们发现 tweet 的内容是 tweet 被转发的决定因素，因此他们构建了基于内容的转发模型[85]。

Starbird 等人对具体事件在 Twitter 上的传播进行了深入研究，他们分析了 2011 年埃及的政治事件，演示了这个事件的相关信息在 Twitter 上是如何生成，发展，传播的[86]。

Comarela 等人研究了影响用户回复或转发的因素，他们发现以前是否回复，发布信息的频率，信息的时效性，tweet 的长度决定用户是否回复[87]。

除了以上的工作，最新的研究还从不同角度对 Twitter 中的转发行为进行了深入的研究[88–91]。

综上所述，我们发现影响人们转发行为的因素主要包括 tweet 文本的内容、tweet 文本的社交媒体属性（如，是否包含链接、hashtag、提及等）、tweet 作者的用户属性，tweet 作者的朋友圈子，当然以上的研究都是从宏观上大规模分析 Twitter 转发数据得出的研究结论。从微观的角度则可以考虑给定一个 tweet，未来这个 tweet 是否会被转发，我们将在 **??**介绍相关工作。

虽然已有的 Twitter 转发研究从许多不同的角度进行了考虑，但是仍然有许多问题与因素被忽视，例如 tweet 的转发预测针对的一般是普遍 tweet，并未细粒度的划分类型，本文我们将针对特定类型 tweet 进行转发研究。另外，目前的转发大多都是从 tweet 本身进行考虑，并未从受众的角度进行分析，本文将对 tweet，作者、受众三个方面在转发过程中的相互关系进行探讨。

1.4 研究内容与方法

1.4.1 本文研究内容

本文的研究内容主要是围绕在给定话题的情况下，如何在 Twitter 中找到与话题相关的主客观 tweet。主要利用 tweet 文本特点和社交媒体属性帮助 Twitter 中的信息检索。本文中 Twitter 的传播分析主要是从内容和受众两个方面进行考虑，如何在 Twitter 中发现会传播的观点和如何在 Twitter 中发现信息的传播者。同样也通过分析 tweet 文本特点和社交媒体特点帮助这两个问题的解决。参见图1.2本文研究框架。

具体四个研究内容的定义如下：

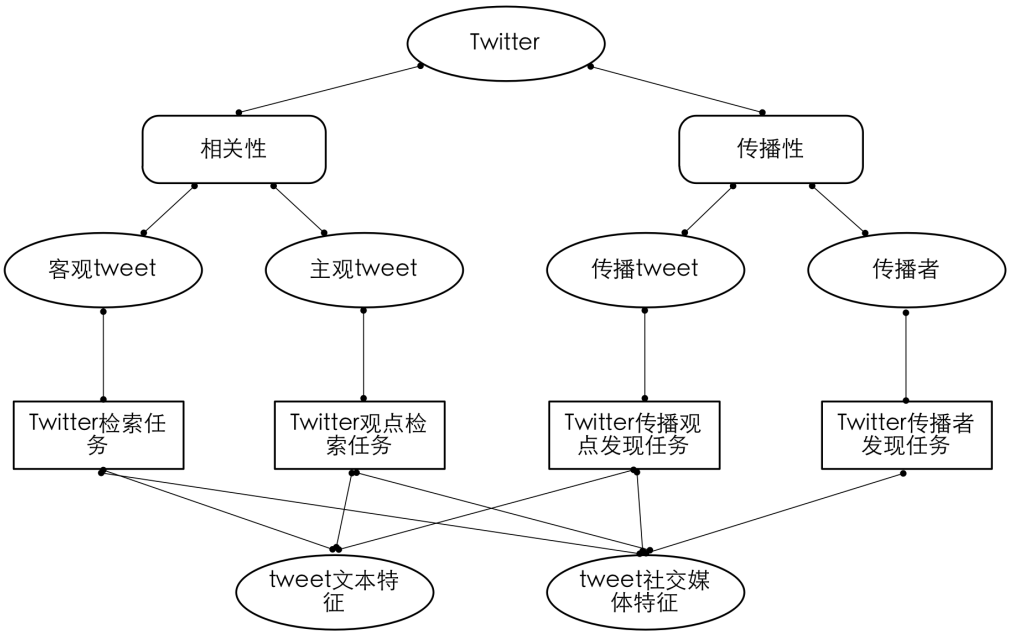


图 1.2　本文研究框架

1. **Twitter 检索**：给定关键词，在 Twitter 中找到话题相关的 tweet。通过文本的结构化信息（我们发现 tweet 文本呈现结构化特点）和社交媒体信息帮助提高 Twitter 检索，以此分析了人们在 Twitter 中是如何描述话题以及社交媒体特征与相关话题 tweet 的内在联系。

2. **Twitter 观点检索**：给定关键词，在 Twitter 中找到话题相关且带观点的 tweet。通过 tweet 观点化信息（基于结构化信息）和社交媒体信息帮助提高 Twitter 观点检索，以此分析了人们在 Twitter 中是如何表达观点以及社交媒体特征与 Twitter 中观点的内在联系。

3. **Twitter 传播观点发现**：给定关键词，在 Twitter 中找到话题相关且带观点的 tweet，并且这个 tweet 在未来会被转发。通过 tweet 观点化信息（基于结构化信息）和社交媒体信息帮助发现 Twitter 中传播观点，以此分析了人们在 Twitter 中是如何表达高质量的观点以及社交媒体特征与 Twitter 中传播观点的内在联系。

4. **Twitter 传播观者发现**：给定 tweet，发现 tweet 的粉丝中，谁会在未来传播这个 tweet。通过社交媒体信息帮助发现 Twitter 中的信息传播者，以此分析了社交媒体特征与 Twitter 信息传播者的内在联系。

总的来说，针对 1.2的第一个问题，本文通过分析 tweet 文本的结构化信息和观点表达方式，找出规律与特点，将其利用到 Twitter 检索、观点检索、传播观点

发现任务。针对 1.2的第二个问题，本文通过分析社交媒体的用户属性、社会网络结构、文本属性，发现用户之间、用户与 tweet 文本（主客观信息)、用户与传播行为、tweet 文本与传播行为之间的内在联系，通过开发社交媒体特征，帮助解决 Twitter 检索、观点检索、传播观点发现和信息传播者发现任务。

1.4.2 本文研究方法

针对以上研究内容，本文基于自然语言处理技术和机器学习技术，深入分析 Twitter 中 tweet 的文本特点和社交媒体属性，解决 Twitter 中若干检索与传播分析问题。我们希望通过对 Twitter 中若干检索与传播分析问题的研究达到如下几个主要目标:

1. 认识以 Twitter 为代表的社交媒体的新特点，包括文本表现形式，用户属性，Twitter 中信息的传播行为等等。

2. 传统的信息检索技术如何在新型的社交媒体中使用，重点研究基于机器学习的信息检索技术在 Twitter 中的应用。

3. 深入研究 Twitter 中观点检索问题，寻找人们在 Twitter 中表达观点的方式，以及其它相关因素。

4. 针对 Twitter 中 tweet 文本质量较低，以及质量评价问题，帮助人们进一步理解 Twitter 中高质量文本的评价问题。

5. 通过研究特定用户查询问题，找到 Twitter 中 tweet、作者和粉丝之间的关系，帮助 Twitter 中传播分析的研究。

根据各个具体的研究内容，我们的具体研究方法为:

1. 针对 Twitter 中的信息检索问题，我们深入分析 tweet 中的文本特点，找到文本特定结构与社会属性之间的关系，开发文本结构特征，然后结合 tweet 的社交媒体特征（用户属性等)，将其整合到机器学习的框架中，通过排序学习，提高 Twitter 中信息检索的效果。

2. 针对 Twitter 中的观点检索问题，首先对 Twitter 中的观点检索问题进行定义，构造测试数据集，然后分析 Twitter 中用户表达观点的文本特点以及 Twitter 中观点所对应的潜在用户属性，开发特征，利用排序学习，解决 Twitter 中如何找到观点的问题。

3. 针对 Twitter 观点检索中大量返回低质量观点的问题，从发现传播观点的角度提出了 Twitter 中高质量观点的客观评价指标，这个指标利用 Twitter 中高质量信息大量传播的特点，分析了 Twitter 传播观点发现的问题，利用 tweet 中如何判定是否转发的方法，tweet 中文本是否包含观点以及 tweet 文本本身的语言质量帮助相关任务的解决。

4. 针对 Twitter 中信息传播者发现的问题，我们首先进行问题定义，构造数据集，分析信息传播者的特点，找到信息传播者与转发 tweet、转发 tweet 作者之间的联系，设计相关特征，将其利用到机器学习框架中，解决信息传播者发现的问题。

1.5 本文主要贡献

本文主要围绕分析 Twitter 文本的特点与 Twitter 社交媒体属性展开，通过 Twitter 中的信息检索和传播分析任务，发现哪些因素能够帮助或影响检索效果的提高与传播分析的准确性。

在 tweet 文本分析方面，我们发现，虽然 tweet 是短文本，但是它具有结构化的特点。不同的 tweet 文本结构对应不同的属性和文本质量，通过挖掘 tweet 的文本结构信息能够帮助 Twitter 的信息检索。另外由于某些特定结构的 tweet 具有某种属性（如主观化），因此可以利用结构化的 tweet 收集大量的相关文本，构造情感词典帮助 tweet 主观化判定，提高观点检索的效果。

在 Twitter 社交媒体属性的分析方面，我们通过对 tweet 中是否包含链接、hashtag、提及等的研究，确定这些符号串或功能与 Twitter 信息检索的对应关系，因此帮助该任务的解决。我们还分析了 tweet 作者的属性，包括作者的粉丝数目、朋友数目、分组数目、兴趣爱好、圈子、活跃时间等等，试图发现这些因素与 Twitter 信息传播之间的内在联系。

在具体的 Twitter 信息检索任务中，我们从给定关键词找到主客观相关 tweet 的两个方面进行研究。获取客观 tweet 方面，我们开发了 tweet 文本的结构化特征和社交媒体特征，将其整合到基于排序学习的模型中，实验结果验证了我们的方法是有效的。获取主观 tweet 即 Twitter 中的观点检索是一个全新的工作，我们定义了这个问题，发布了关于研究这个问题的实验数据[12]，截止到 2013 年 9 月这个实验数据已经有 40 多个国家和地区超过 200 多个研究单位和个人下载，并且这个数据还被 ICWSM 会议推荐为官方的社交媒体研究数据[13]，针对 Twitter 中的观点

[12]下载地址：http://sourceforge.net/p/ortwitter/wiki/Home/

[13]http://www.icwsm.org/2013/datasets/datasets/

检索我们也提出了我们的方法，主要是开发 tweet 的文本特征与社交媒体特征结合排序学习框架进行解决，同时我们也提出了一种基于无监督学习的 tweet 观点化评价的方法，目前有许多其它研究单位的工作围绕我们的 Twitter 观点检索工作展开[92–95]。

在具体的 Twitter 传播分析任务中，我们从传播的内容和受众两个方面进行考虑，提出了 Twitter 中传播观点发现的新任务和信息传播者发现的新任务。Twitter 中传播观点发现可以帮助我们解决 tweet 质量评价主观化的问题，由于以往的研究主要是围绕给定 tweet，预测该 tweet 在未来是否会被转发，我们对这个问题进行了细粒度的研究，从观点化的 tweet 能否被转发的角度进行了探讨，通过开发 tweet 的文本特征和社交媒体特征解决传播观点发现的问题。Twitter 信息传播者发现的问题，针对的是以往研究忽视“谁”会转发的问题，我们定义了这个问题，提出了解决这个问题的方法，发现兴趣与转发的历史信息是决定信息传播者的重要因素，同样我们也公布了研究这个问题的数据集[14]，供以后科研人员继续使用。

以上所有的工作都通过论文的形式公开发表[96–99]。

1.6 本文结构

本文的研究工作主要围绕社交媒体中检索与传播分析任务展开，我们可以将这两方面的工作分为以下几个主要部分：在 Twitter 检索方面，我们首先分析了 tweet 的文本信息与社交媒体信息，以此帮助 Twitter 中传统的信息检索任务；然后以此为基础，进一步探讨 Twitter 中观点检索问题，给定关键词，检索到话题相关且带观点的 tweet；在 Twitter 的观点检索任务中，我们发现检索结果存在大量的低质量观点，结合 Twitter 中的传播分析，我们从传播的内容角度考虑，转发的 tweet 一般是高质量的信息，因此我们再进一步研究了在 Twitter 中如何发现传播观点的问题；Twitter 信息的传播分析不仅可以从 tweet 的本身进行研究，也可以从受众的角度进行分析，因此最后我们讨论了在 Twiter 中如何需找信息传播者的问题。上述工作共分为六个章节, 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:

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 Twitter 的一些基础知识，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 Twitter 中的信息检索，首先介绍了 Twitter 信息检索的研究背景，然后提出了以往 Twitter 信息检索方法忽视 tweet 文本结构化特点以及存在大量社交

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

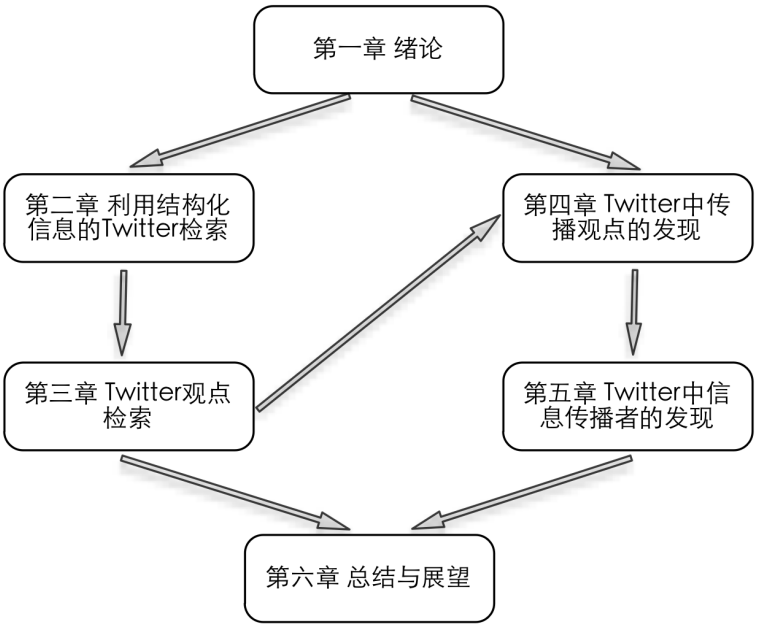


图 1.3　论文整体结构图

媒体信息的问题，以此设计了一种标注 tweet 文本结构的自动标注器，最后利用自动标注器标注 tweet 文本开发结构特征，结合社交媒体特征帮助 Twitter 信息检索，实验验证了方法的有效性。这一章回答了 Twitter 中人们是如何用自然语言描述客观话题，并且社交媒体特征与 tweet 话题相关性存在怎样的联系。

第三章是 Twitter 中的观点检索，本章开头定义了 Twitter 中的观点检索问题，分析了 Twitter 观点检索与以往观点检索的不同特点，接着提出了一种自动获取主观 tweet 与客观 tweet 的方法自动生成情感词典，依靠词典对 tweet 的文本进行主观化判定，结合 tweet 的用户属性信息和文本信息，利用排序学习算法，实现观点检索。实验部分，我们构造了自己的 Twitter 观点检索语料，并发布了语料供以后的研究者使用，实验结果证明了我们的观点检索方法有效。这一章回答了 Twitter 中人们是如何用自然语言表达主观观点，并且社交媒体特征与 tweet 观点相关性存在怎样的联系。

第四章中我们针对 Twitter 观点检索存在大量低质量观点的问题，依据转发 tweet 一般是高质量文本的既有研究成果，提出了在 Twitter 中发现传播观点的问题。我们首先定义了问题，然后构造了数据集，接着开发了 tweet 传播度特征、观点化特征和文本质量特征，将其整合到排序学习的机器学习模型框架中。实验结果说明了这些特征对于 Twitter 中发现传播观点是有帮助的，另外我们的方法可以达到人判定传播观点的效果。这一章回答了 Twitter 中人们是如何用自然语言表达传播性的观点，并且社交媒体特征与传播性的观点存在怎样的联系。

第五章中我们探讨了 Twitter 中发现信息传播者的问题，给定一个 tweet，发现 tweet 的作者粉丝中谁会转发该消息。我们开发了转发历史特征、用户特征、用户活跃时间特征和用户兴趣特征，并依然将其应用到排序学习的框架中构造模型进行排序。实验部分，我们构造了自己的测试数据与基准系统，我们公布了数据，实验结果显示了我们的方法能够成功找到 tweet 转发者。这一章回答了社交媒体特征与信息传播者存在怎样的联系。

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，并对未来社交媒体检索与传播分析的一些问题和方法进行了尝试性地思考。

第二章　应用语义关系自动构建情感词典

2.1　引言

随着互联网的发展, 尤其是社交网络的发展，各种社交媒体的用户发布内容中出现了海量含有用户主观情感色彩的文本数据。针对网络文本的信息处理开始由获得关键词[100]、事件 [101]、话题[102] 等事实信息, 开始向情感观点等主观信息深入，情感分析便是近年来迅速发展的信息处理技术[103]。从数据中提炼出用户的主观信息对于商业情报、舆情分析等具有重要意义。情感分析技术就是对带有情感色彩的主观性文本进行自动推理、分析、归纳的过程，涉及自然语言处理、机器学习、认知科学以及社会心理学等方面的研究[104]。语言的情感表达往往使用具有明确情感色彩的词汇，因此构建带有情感色彩的词典资源是进行情感分析研究的基础。情感分析研究在英文上发展迅速，积累了许多情感词典资源，比如：General Inquirer（GI）[105]，OpinionFinder（OF）[106]，Appraisal Lexicon（AL）[107]，SentiWordNet[108] 以及 Q-WordNet[109]。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[110–112]。目前研究使用主要有 HowNet 情感词典[113]，NTUSD 情感词典[114] 以及大连理工大学的情感词汇本体词库[115]。这些词典主要是以手工或半自动方式编辑而成，可靠性和领域适应性受到限制，并且情感词以主要褒贬二值区分，缺少情感强度值的细粒度划分。能够将资源丰富的英文词典跨语言向资源相对贫乏的语言进行适应性的转化，以产生其相应情感分析词典，既可以省去耗费大量人力的人工标注过程，又可以克服自动或半自动方法的可靠性问题。

本章提出基于可靠的英文情感词典资源的中文情感词典的构建方法，可以根据语义关系将英文词及其情感值转化得到中文词语的情感强度值，并且完全是自动的，可靠性和适应性更高。

2.2　词典资源简介

2.2.1　HowNet 语义词典

HowNet 是一个以中英文词语所代表的概念为描述对象，揭示概念与概念之间以及概念的属性与属性之间的关系的知识库。义原是 HowNet 最小语义单元，用于定义和描述概念的属性和概念间的相互关系，义原通过一个树状的层次结构组织构成上下位关系。概念是对词汇语义的一种描述，每一个词可以表达为几个概念[116]。如图 2.1所示，HowNet 采用 KDML（Knowledge Dictionary Mark-up

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

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

图 2.1　HowNet 中概念的定义方式

2.2.2　WordNet 语义词典

WordNet 是由 Princeton 大学的心理学家，语言学家和计算机工程师联合设计的一种基于认知语言学的英文词典[117]。WordNet 是根据词义而不是词形来组织词汇信息。WordNet 使用同义词集合（Synset）代表概念，词汇关系在词语之间体现，语义关系在概念之间体现。WordNet 将英语的名词、动词、形容词和副词组织为 Synsets，每一个 Synset 表示一个基本的词汇概念，并在这些概念之间建立了包括同义关系（synonymy）、反义关系（antonymy）等多种语义关系。其中，WordNet 最重要的关系就是词的同义反义关系。

2.2.3　SentiWordNet 情感词典

SentimentWordNet 是 Baccianella[108] 等在语义词典 WordNet 基础上使用随机游走的图算法得到的情感词典。词典的每条记录都是一个 WordNet 的 Synset，并且每个 Synset 都计算出了褒义、贬义情感强度值，本文就是利用 SentimentWordNet 的情感强度值以及 HowNet 概念的语义关系进行计算得到中文词语的情感值。SentimentWordNet 共有 117,000 多 Synsets，192,493 单词。

2.3　基于语义关系的情感词典构建方法

将英文情感词典的研究成果转化为文资源，可以利用语言之间的语义对应关系减少词典的歧义，使情感词典更加可靠，还可以直接将英文中对情感强度的计算直接转化为中文词语的情感强度计算，减少了计算开支。本研究正是基于这种

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

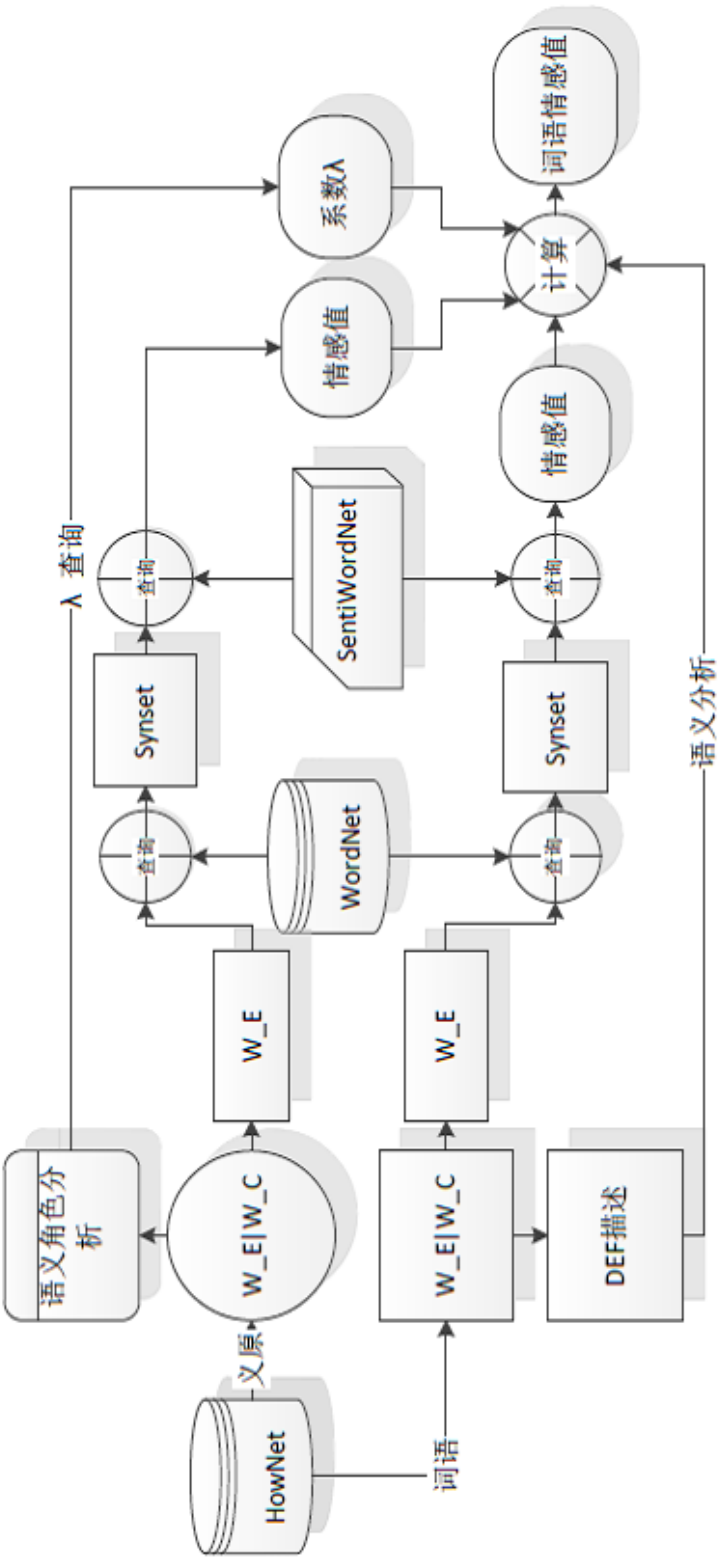


图 2.2　基于语义关系的情感词典解决方案

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

2.3.1 词语抽取和义原抽取及语义分析

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

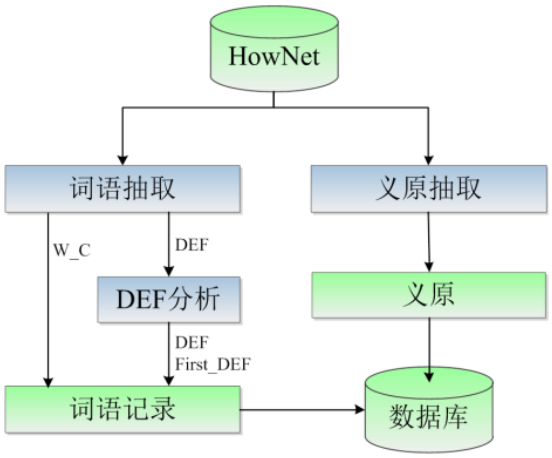


图 2.3　词语和义原抽取处理流程

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

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

2.3.2 情感值的查询与计算

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

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

2.3.2.1 词语倾向性值查询与计算

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

2.3.2.2 义原倾向性值查询与计算

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

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

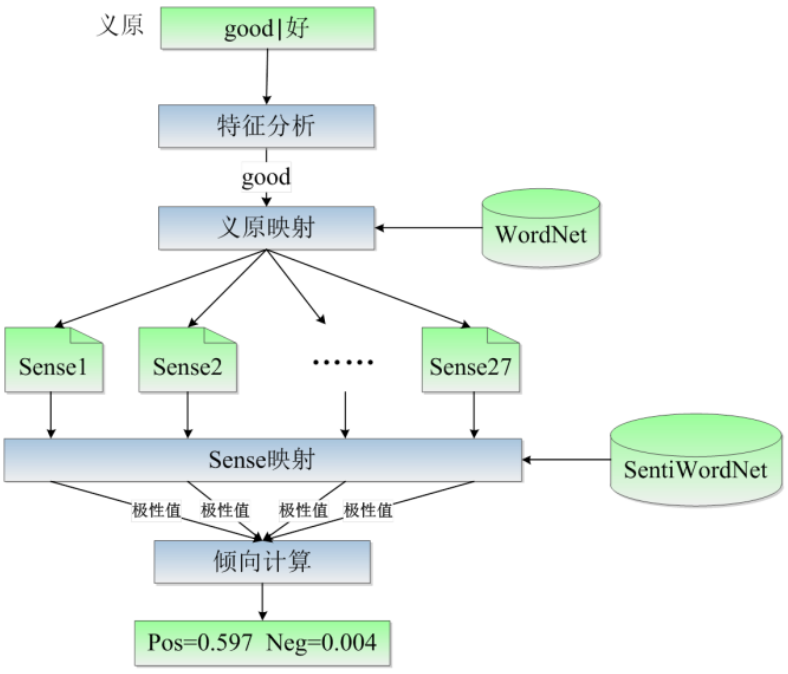


图 2.4 义原情感值计算过程

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

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

2.3.3 词语情感值计算

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

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

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

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

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

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

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

其中：$\sum_{p \in P} \varphi(s,p) = 1$。

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

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

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

2.4 实验及结果

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

2.4.1 评价指标

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

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

召回率计算公式为：

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

F 值计算公式为：

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

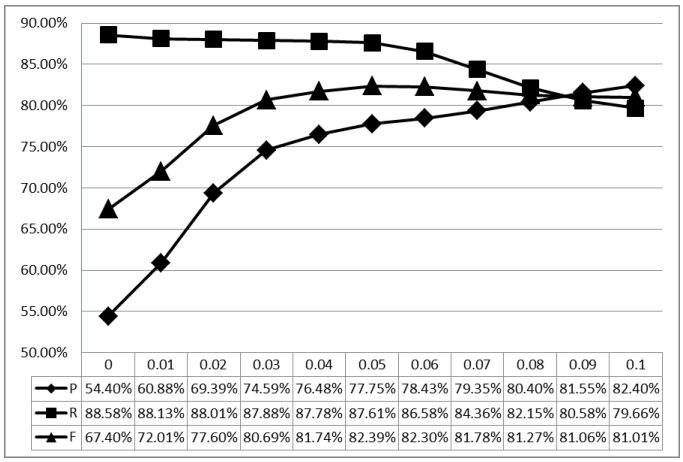
。

2.4.2 性能评测结果

2.4.2.1 阈值 T 的设定

由于基准词是褒贬二值标注的，因此需要将生成的情感词典连续情感值转换为离散褒贬值。将褒义和贬义情感值相减得到词语的倾向值来判断词语的极性，

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



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

图 2.5　不同 T 值时的性能指标

2.4.2.2　与其他词典性能对比

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

表 2.1　T=0.05 时的性能对比

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

2.5　小结

本章对中文情感词典构建相关研究进行了分析，以英文情感词典为基础，设计了基于语义关系的情感词典自动构建方法。方法以 HowNet、WordNet 语义词典和 SentiWordNet 情感词典为基础，借鉴英文情感词典进行中文情感词典的构建，并且与现有的常用情感词典进行了实验对比。实验结果表明，本文设计的方法取得了较好的评测性能。

第三章　基于语料资源的中文情感词典扩展

3.1　引言

自然语言中，一个词语的语义极性（semantic polarity）表示它对于其语义组（semantic group）或词汇场（lexical field）范式的偏离方向[118]。在自然语言处理领域，情感分析（Sentiment Analysis）能够使用计算手段自动从自然语言中发现观点和情感等主观信息[13, 15]，通常会使用一些标注了极性（积极或消极）的词汇构成的情感词典资源。研究如何能够通过计算方式获得词语的语义极性，自动构建情感词典一直得到计算语言学和自然语言处理研究人员关注。在英文情感词典构建中，Wilson 等[119, 120] 对一些单词进行了人工极性类别的标注形成了 OpinionFinder 词典；Bradley 等[121] 标注了并发布了情感范式的英文词典 ANEW，后来 Nielsen 等[122] 在 Twitter 语言上应用并自动扩展了 ANEW，形成 AFINN 词典。Esuli 和 Sebastiani[123] 以及后来 Baccianella 等[108] 在著名的语义词典 Wordnet 基础上采用自动计算的方式开发出了情感词典 SentiWordnet。Thelwall 等[124] 设计实现了能对词语的情感强度进行估计的方法。情感也可以通过创建情绪词典来进行计算，Plutchik 情绪轮提出了四对对立的情绪状态：joy-trust，sadness-anger，surprise-fear 和 anticipation-disgust[125]。Mohammad 和 Turney[126] 根据 Plutchik 情绪轮分类方法使用情绪分值标注了一些词语形成 NRC 情绪词典。在 2013 和 2014 年举办的 SemEval（Semantic Evaluation）评测中，NRC-Canada 队利用 NRC 词典并扩展出两种新的词典，取得了最好成绩[127, 128]。为了克服以上语法层面建立的词典的上下文语境以及领域适应性问题，一些学者提出了基于概念（concept-based）构建情感词典[129]，其中 SenticNet 是使用常识知识库建立的公开可用的基于概念的情感词典[130]。

中文情感分析研究起步较晚，想对于丰富的英文情感词典资源，缺乏普遍认可的可靠的中文情感词典。目前研究使用主要有 HowNet 情感词典[113]，NTUSD 情感词典[114] 以及大连理工大学的情感词汇本体词库[115]。这些词典主要是以手工或半自动方式编辑而成。我们的前期工作提出了根据语义词典 HowNet 语义关系将英文情感词典跨语言转换为中文情感词典的方法，并构建了带有极性标示以及极性强度值的的情感词典 SentiHowNet[131]。基于语义词典的情感词典构建方法是一种常用的情感词典构建方法。采用这种方法的优势在于可以比较容易获取情感词语，基于词语的语义关系也易于进行情感极性计算。但是，基于语义词典的情感词典构建方法受限于语义词典的规模和语义关系的定义，而且对于专业领域中不断涌现的新词语，对情感词典的覆盖度提出了严峻的挑战。随着互联网应用，

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

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

3.2 问题描述

基于语料资源的情感词语选择与极性计算，在英文中相关研究通常有两种实现思路：一是基于语言特征的方法，例如，Hatzivassiloglou[118] 等人采用并列或转折连词来判断新的情感词并计算其极性。二是基于统计特征的方法，例如，Turney 等[132] 采用点互信息统计学方法从语料中发现共现度高的情感词并计算其极性。基于以上情感词语选择与极性计算方法的分析，本文将基于中文语料资源扩展情感词典时需要解决的问题描述如下：

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

3.3 数据集及预处理

本文使用的数据资源如表 3.1所示，选取的语料资源是谭松波博士提供的酒店、书籍和电子商品评论三个领域的语料文本各 4000 篇[133]。

表 3.1 数据集及词典资源

词典	SentiHowNet	基于前期工作[131]
语料	Hotel	4000 篇
	Book	4000 篇
	NoteBook	4000 篇

其中对语料进行预处理需要将中文文本进行分词并进行词性标注。中文分词处理是对语料进行进一步处理的基础，采用的是中科院设计实现的 ICTCLAS 分词软件[134]；然后将词性标注为形容词（ADJ）和副词（ADV）的，在 SentiHowNet[131] 中出现的进行极性和极性值标注；生成的结构化语料预处理记录格式如图 3.1所

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

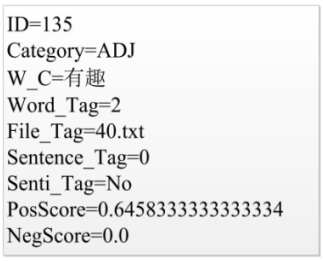


图 3.1　语料预处理记录格式

3.4　基于语言特征的情感词典扩展

早期对于英文语言特征一些研究[118] 发现，由连词（如 and 或 but）连接的两个形容词的极性往往存在一定的关联性，如“and”连接的形容词（如“nice and good”）极性相同，而“but”连接的形容词（如“nice but unnatural”）极性相反。而对于中文来说，基于语言特征的中文情感词是否会遵循想通的规律，非常值得进行研究。

3.4.1　连词选择

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

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

3.4.2 基于连词的极性计算

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

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

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

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

具体计算过程如算法3.1所示。

3.4.3 实验

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

表3.2 各个领域性能评测结果

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

算法 3.1 基于连词的极性计算

已知:

待标注词语集，$\{w_1\}$;

连词集合，$\{c\}$;

极性已知词语集合，$\{w_2\}$;

for 每一待标注词语 $w_1 \in \{w_1\}$ **do**

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

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

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

    else

      if c 为并列或递进连词 **then**

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

      end if

      if c 为转折连词 **then**

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

      end if

    end if

  end for

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

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

  将 w_1 加入到集合 $\{w_2\}$;

end for

3.5 基于统计特征的情感词典扩展

词语的上下文是词语在实际应用中的语言环境，它在自然语言处理中的价值体现在两个方面：一方面，在自然语言知识获取的过程中，上下文是知识获取的来源；另一方面，在自然语言处理的应用问题解决过程中，上下文扮演着解决所需信息和资源提供者的重要角色。特别是在语料库语言学中，各种机器学习方法的引入使词语的上下文成为计算语言学知识获取和问题求解过程中最为重要的资源，在无监督学习方法中更是如此[135]。本文设计实现的基于统计特征的情感词典扩展方法主要是采用基于上下文的方法进行情感词语极性计算，因为出现在相似上下文环境中的词语具有相似的极性。

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

3.5.1 统计特征选择

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

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

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

3.5.2 基于上下文的情感词极性计算

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

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

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

3.5.3 实验

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

算法 3.2 基于统计特征的极性计算

已知:

待标注词语集，$\{w_1\}$;

极性已知词语集合，$\{w_2\}$;

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

for 每一待标注词语 $w_1 \in \{w_1\}$ **do**

  for w_1 每一特征向量 $V(w_1)$ **do**

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

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

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

      else

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

      end if

    end for

    对各个特征向量下的情感值累加

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

  end for

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

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

  将 w_1 加入到集合 $\{w_2\}$;

end for

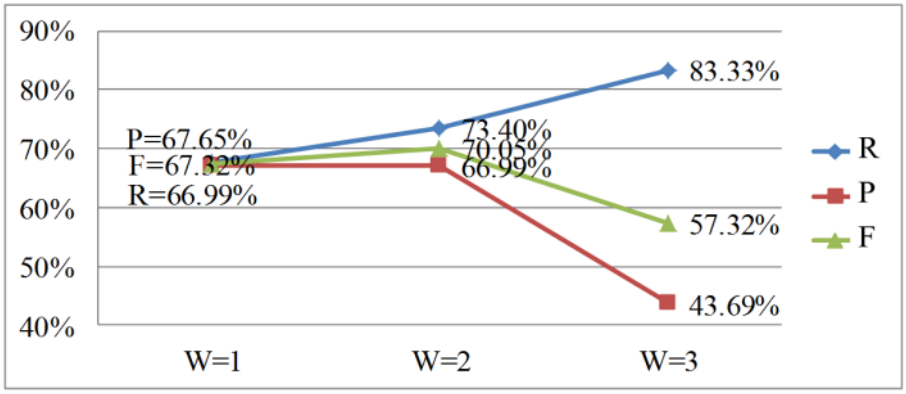


图 3.2　Hotel 语料评测结果

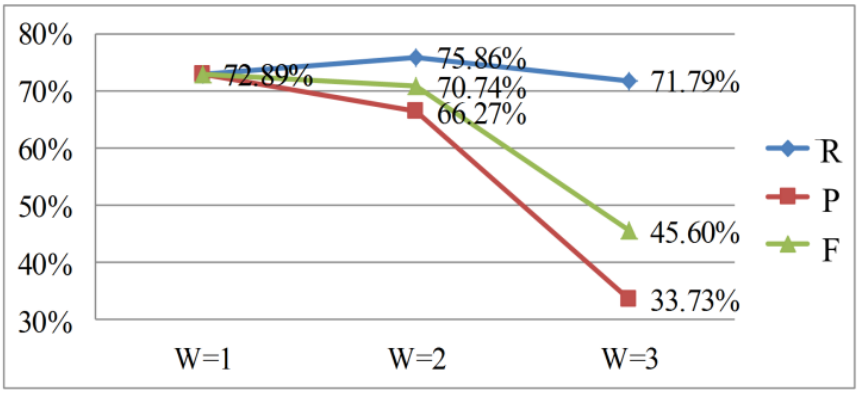


图 3.3　Book 语料评测结果

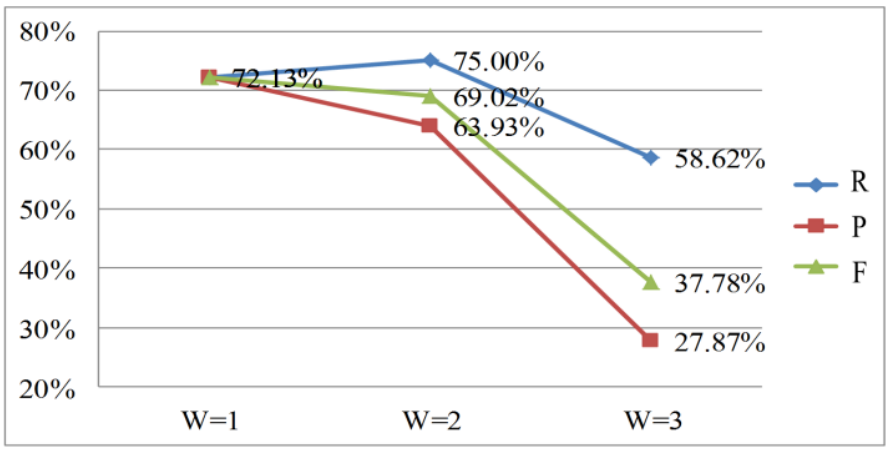


图 3.4　NoteBook 语料评测结果

3.6　基于混合特征的情感词典扩展

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

3.6.1　基于混合特征的情感词极性计算

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

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

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

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

算法 3.3 基于混合特征的极性计算

已知:

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

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

3.6.2 实验

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

表 3.3 各个领域性能评测结果

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

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

3.7 小结

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

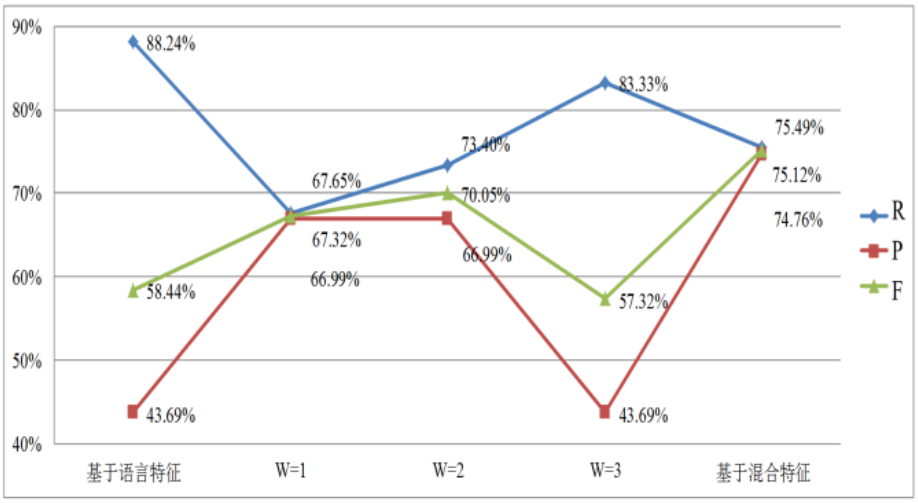


图 3.5　Hotel 语料评测结果综合比较

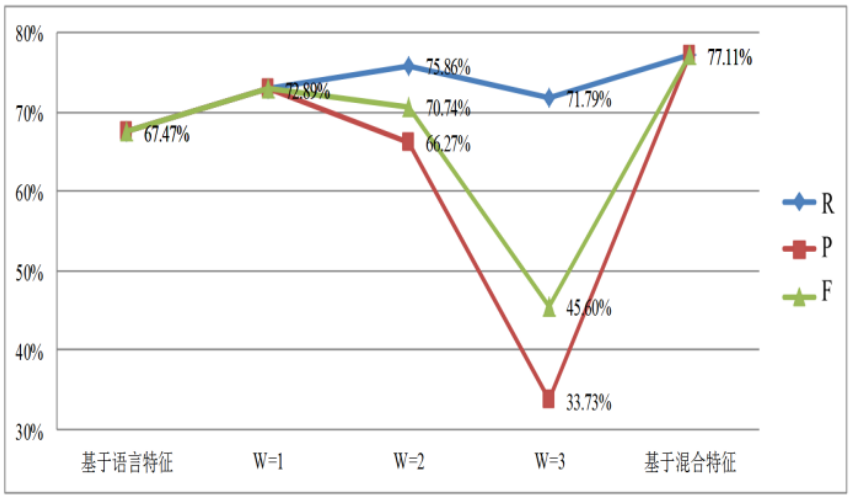


图 3.6　Book 语料评测结果综合比较

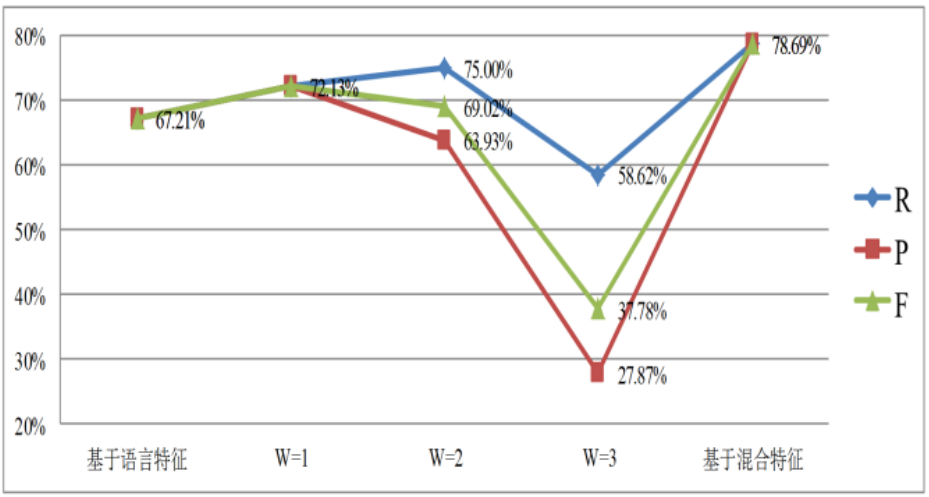


图 3.7　NoteBook 语料评测结果综合比较

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

第四章　无监督的自举式情感分类

4.1　引言

文本的情感分析是挖掘文本中主观性信息学的主要手段，是研究文本中观点、态度、情绪和立场等主观性信息是如何表达的。情感分析技术可以从数量庞大的文本数据中抽取并总结主观性信息，为后续的一些应用（商业智能（Business Intelligence），舆情分析（Public Opinion Analysis）或选举预测（Election Prediction）等）提供技术和工具支持[15]。在社交媒体中，一些基于文本的平台比如微博产生了大量针对各种话题或实体的带有主观性信息的数据。这对这些数据的分析，也就是情感分析正逐渐受到各个研究领域（比如推荐系统和搜索引擎）的重视。情感分析中一个主要的方法就是应用各种分类技术，也就是根据作者的主观态度将文本进行分类，一般将这种研究称为情感分类研究。情感分类一般是从一些标注过的训练数据中通过学习得到一个分类模型，学习得到能够将一种情感类型区别于其他类型的一些特征[136]。这种模型的性能主要依赖于其能够学习到的数据中出现的情感的文本表达模式，一般是文本中出现的词语、短语或者词语的各种组合。情感分类也就是情感分析的分类形式，虽然可以被视作文本分类的特殊形式，实际上情感分类是比文本分类更具挑战的任务，因为文本中情感的表达方式严重依赖领域和上下文环境[137?]。

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

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

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

4.2 相关工作

情感分类在观点挖掘研究中越来越受到重视，前期工作主要研究针对评论(商品或电影）进行情感分类。经常使用的方法可以分为基于规则的方法和基于学习的方法，其中基于机器学习方法性能一般比较好因而常被用来作为对表的标准[138]。

现在研究人员逐渐开始注意到微博中用户的主观性信息，并开始结合微博的语言特点进行对微博进行情感分类研究。一些研究显示可以将微博的一些独特的特征结合进情感分类方法中。比如，Barbosa 和 Feng [139] 提出了两阶段支持向量机分类器（Support Vector Machine (SVM) classifier）对 tweet 进行情感分类，证明了该分类其能对 tweet 的类别偏置（biased）和噪声具有很好的鲁棒性；Hu 等 [140] 将社交媒体数据中的情感表达解成情感指征（emotion indication）和情感关联（emotion correlation）两种信号，通过对两类情感信息进行联合建模方式实现了对微博的无监督情感分类；Jiang 等 [137] 主要关注依赖于特定目标的微博情感分类，提出了通过将目标依赖特征（target-dependent features）和相关微博同时进行考虑的监督学习方法，并证明了可以提升情感分类性能；Wang 等 [141] 针对 hashtag 级别的情感分类进行了研究，并提出了一个全新的图模型，然后使用提升（boosting）式分类方法进一步提高了模型的性能；Amir 等 [142] 针对单条微博的情感分类提出了一个分层分类器框架，框架通过抽取对特定目标的微博，将微博按情感类型分开以及分离正负情感类型微博三个层次进行有监督的分类学习；Hu 等 [143] 基于社交理论抽取微博之间的情感关系，提出了一种全新的社会学方法使用这些情感关系以促进情感分类性能，并有效解决了数据中的噪声问题；同样受到社会学理论的启发，Guerra 等 [144] 依据人类通常会持有一致的带有偏执的观点，提出了全新的迁

移学习（transfer learning）方法解决微博基于话题的实时情感分类问题；Thelwall 等 [124, 145] 设计了 SentiStrength 情感分析工具，用于对微博等社交媒体中非正式语言中的情感分析，该工具是基于规则的方法，使用了人工编辑的词典并结合了微博语言中的句法和拼写特点抽取微博中的情感强度，该工具获得了广泛的应用。

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

4.3 问题的形式化

简单来说，情感分类主要目标就是将文本分类为预先定义的情感极性类别(一般是积极的，positive 或消极的，negative)。形式化上，对于给定的文档语料库 $D = \{d_1, \ldots, d_n\}$，预定义的情感类别 $Y = \{1, -1 \mid \text{positive} = 1, \text{negative} = -1\}$，情感分类的任务就是对每一个文档 d_i 预测一个类别标签 y_i。为了与文本分类问题一致，每个文档可以表示为一个特征向量 $x = R^n$，n 表示特征空间的大小对于情感分类问题来说，对于每一个特征通常将其权重定为二值的，1 表示特征在文档中出现，0 表示没有出现[138]。对于有监督的机器学习，给定训练集 $D = \{x_1, \ldots, x_m\}$，可以学习到分类器：

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

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

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

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

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

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

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

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

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

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

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

基于以上假设，情感分类问题可以进一步形式化定义为：

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

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

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

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

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

■

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

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

4.4 无监督的情感分类框架

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

相似的，对于微博的自动情感分类，基于我们特征空间分割假设，可以认为一条微博的情感倾向性可以通过两部分特征都识别出来。也就是说，如果能够对与一些通用的情感表达知识，在某种程度上也能识别出一条微博的情感极性（根据微博中正规表达方式的比例不同，比例越大就越容易识别)。实际上有很多研究者已经开始研究如何建立各种情感词汇表来对这种通用的情感知识进行建模了，比如我们前面章节的工作中提到的 OpinionFinder 词典[119, 120]、ANEW 词典[121]、AFINN 词典[122]、SentiWordnet[123]、HowNet 情感词典[113]，NTUSD 情感词典[114]、情感词汇本体词库[115] 以及我们的 SentiHowNet[131]。虽然这些词典在尝试着建立

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

4.4.1 通用情感分类器

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

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

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

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

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

4.4.2 微博情感分类器

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

为了对微博情感特征的领域依赖部分进行建模，有两个问题必须考虑。首先是如何界定微博中抽取的特征中那些是领域依赖的特征。随着用户发布微博数量的急剧膨胀以及用户在语言使用上的自主性，一些微博特有的主管观点或情感的新的表达方式不断出现，并且同样使用一些通用词语，在特定的语境下也会出现不同于其固有的语义信息。不断涌现的新词和词语在为上的语义变异使得界定领域依赖部分特征变得非常困难。但是众所周知，微博文本属于短文本，每条微博都有字数上的限制（一般是要求 140 字以内），因此用户在一条微博中表达就某件事情表达某种情感倾向时，除了描述事情所用词语外，只能够用很少的词语描述情感倾向。因此我们可以假设，如果一条微博中含有某个成语或谚语，如果没有否定词，整条微博的情感倾向可以看作是和成语或谚语的情感倾向一致，并且除成语外的其他词语形成的情感特征可以视为领域依赖特征。第二个问题是如何找到足够的标注数据来训练得到依赖于微博特征的微博情感分类器。之前有些研究提出了远监督（distant supervision）方法来解决微博标注数据缺失的问题[146, 147]，主要是基于微博中含有明显的情感倾向的一些表达方式（表如表情符）为基准来发现一些含有噪声的微博作为训练数据。我们也是利用这样的思想，但是我们利用成语资源作为我们的情感表达基准，找到包含成语的微博（过滤掉含有否定词的部分）作为微博依赖的情感分类器的训练数据。

4.4.3 分类器的组合

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

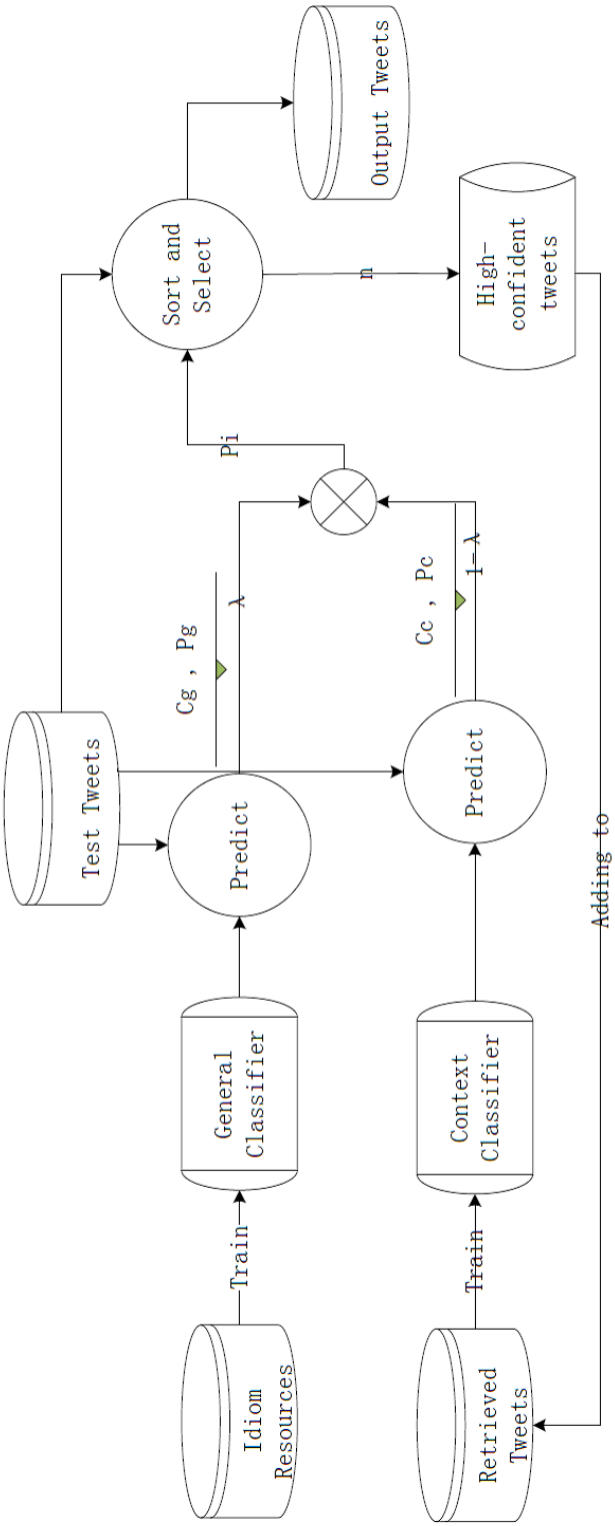

Output Tweets
Sort and Select
n
High-confident tweets
Pi
Cg , Pg
λ
Cc , Pc
1−λ
Test Tweets
Predict
Predict
General Classifier
Context Classifier
Train
Train
Idiom Resources
Retrieved Tweets
Adding to

图 4.1　自举式学习框架

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

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

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

总体来说，整个框架可以被视作是一个自举式（bootstrapping）共同训练（Co-training）[147] 机器学习算法过程，所不同的是该框架并没有使用标注好的训练数据，而是从现成的成语词典资源作为训练的起始点，是一个无监督的学习框架，因此节省了人工或自动标注微博数据的过程，对于数量庞大的微博数据来说，该框架更加实用。

4.4.4 分类器算法

对于两个分类器，我们采用跟 Pang 等 [138] 文章中一样的三种机器学习算法：朴素贝页斯（Naïve Bayes）算法，最大熵（Maximum Entropy）算法以及支持向量机（Support Vector Machine）算法。这三种算法的有效性已经得到 Pang 等 [138] 的验证，其中支持向量机取得的性能是最好的（准确率达到 82.9%）。

4.4.4.1 Naïve Bayes 分类器.

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

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

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

4.4.4.2 最大熵分类器.

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

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

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

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

4.4.4.3 支持向量机分类器.

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

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

情感分类问题通常使用线性支持向量机分类器。

4.5 实验

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

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

4.5.1 实验描述

4.5.1.1 数据集

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

4.5.1.2 实验配置

为了能够多角度测量分类器的性能，有各种评测指标，但是我们的实验不是为了比较这些评价指标的不同，因此我们选择了简单直观的准确率作为分类器性能的评价指标。对于分类器，我们选择了自然语言处理的工具 NLTK（Natural Language ToolKits）[148] 中的 Naïve Bayes 分类器和最大熵分类器，以及常用的 Libsvm[149] 工具包的支持向量机分类器。分类中所有的参数设置都经过交叉验证进行了优化。

4.5.1.3 评价基准

为了客观评价我们的方法的性能，我们设置了三个评价基准用于对比评价。一个是 50% 的基准，因为我们所用的测试数据集是平衡数据集，所以即便是随机判断的分类器的准确率可以达到这样的准确率；第二个是用一个基于情感词典的情感分类器的准确率作为基准，我们使用的是前面两章构建的 SentiHowNet 情感词典，通过计算每条微博中的包含的情感词语情感值叠加来计算综合情感值，然后判断微博的情感倾向性；第三个基准是有监督的机器学习方法构建的情感分类器，我们按照 Pang 等 [138] 文章中的方法使用测试数据通过 5 倍交叉验证方式训练了朴素贝页斯（Naïve Bayes）、最大熵（Maximum Entropy）以及支持向量机（Support Vector Machine）分类器，把三种分类器的在测试数据集上的准确性作为基准。

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

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

4.5.1.4 数据预处理

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

4.5.2 实验结果

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

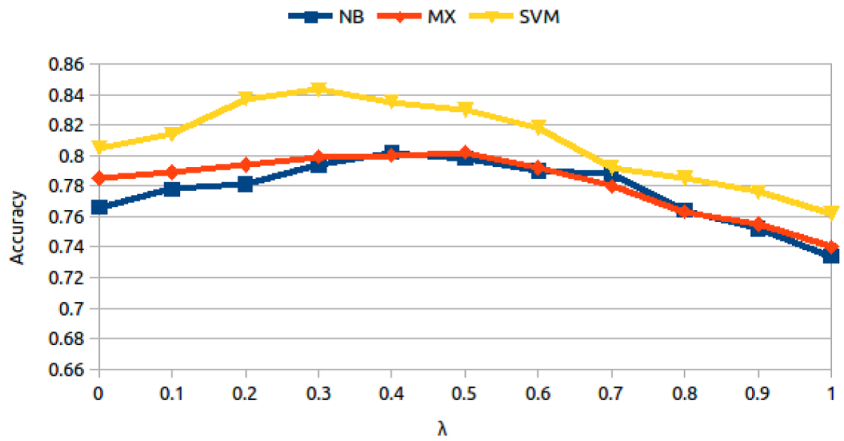


图 4.2 λ 值的确定。

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

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

表 4.1 结果对比表

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

从表中可以看出，首先无论是通用情感分类器还是微博情感分类器，性能上都超过了随机基准的 50% 准确率，这证明了无论是从那种视角进行分类，两种分

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

类器都是有效的，胜过随机猜测。因此在没有任何标注数据来训练有监督或半监督分类器的情况下，我们的特征分割假设可以作为情感分类的一种有效的方法。

其次，通用情感分类器的准确率比基于情感词典的分类器要稍微好些，这是因为尽管通用情感分类器和基于词典的分类器都使用领域独立的词语来对情感知识进行建模，但是通用情感分类器是经过成语等知识资源所抽取的特征空间进行训练的，而情感词典中的词语都是独立使用并且还是具有一定的歧义；而对于微博情感分类器，准确率都比基于词典的分类器和通用情感分类器要好，因为它是使用微博依赖部分的特征训练得到的，更能使用微博语言环境，测试数据中出现的微博的“火星语言”越多越能体现出微博情感分类器的性能优越性。

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

4.6 小结

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

第五章　Twitter 中信息传播者的发现

5.1　引言

Twitter 的快速发展使得信息的交流变得越来越方便和快捷。人们每天在 Twitter 上不仅接受资讯，同时也发表自己的观点。在 Twitter 中一个很重要的机制就是转发（retweeting）：重复发送其他用户发布的 tweet。这种机制是 Twitter 中最常用的信息传播手段，当一个 tweet 被某人转发时，该用户的所有粉丝都将看到此信息，同时这种转发行为也反映了转发者对 tweet 原作者所持观点的一种肯定的态度。

现有对转发的研究主要集中在给定一条 tweet，预测未来是否会被转发上[31, 150, 151]；另外一部分研究则把重点放在研究转发的行为模式上[83, 84, 152]。但是以上的研究都忽略了一个重要的问题：到底“谁”会转发给定的 tweet。我们将这个“谁”称为信息传播者。上一章，我们讨论了 Twitter 中传播观点发现的问题，这是从 Twitter 传播分析的 tweet 角度进行考虑，本章中我们将从传播的受众角度对 Twitter 进行传播分析的研究，即如何在 Twitter 中找到信息传播者。

由于用户参与 Twitter 的低门槛特点，因此许多人在 Twitter 上发布低质量甚至虚假的信息，其中对社会危害最大的就是谣言。这种信息加上转发机制，可以在 Twitter 上迅速传播，造成大众的恐慌，危害极大。目前在 Twitter 上减低谣言的危害，除了自动识别谣言信息以外，我们认为对谣言传播路径的监督与控制也是有效的手段之一。如果当一条 tweet 确定为一条谣言以后，我们通过信息传播者分析，能够预测哪些用户会传播这些谣言，对于维护社会和谐稳定是具有积极的意义。另外，研究信息传播者可以帮助商业或娱乐公司减少广告成本，因为这个研究可以找到性价比高的用户发布相关用户感兴趣的信息，并且迅速传播，扩大商业公司或娱乐公司所推广商品的影响力。

研究在 Twitter 中如何找到信息传播者十分有意义，它能够帮助我们加深理解信息是如何在这个社交媒体中传播的。这些影响信息传播的因素可能与消息的本身，消息的作者，以及消息的接受者有关。三种因素之间相互影响，决定信息的传播。我们将 Twitter 中找到信息传播者的问题看做一个排序问题，即给定一个 tweet，在作者的粉丝中发现“谁”将转发该消息。我们利用机器学习的方法，为用户设计一系列特征构建排序学习模型。这些特征包括了用户历史的转发信息，用户自身的社交媒体特征，用户使用 Twitter 的活跃时间，以及用户的个人兴趣。

我们构造了自己的实验数据来验证 Twitter 信息传播者发现方法的有效性，实验结果表明我们的方法显著优于随机排序的方法和基于用户以往转发原作者 tweet

数量的排序方法。另外，我们发现用户历史的转发信息、用户的兴趣、以及用户的活跃时间是决定转发者的重要因素。

本章的主要工作如下:

1. 我们定义了 Twitter 中信息传播者发现的新任务，帮助理解信息在 Twitter 中是如何传播的。

2. 我们设计了基于用户转发历史信息、用户属性、用户活跃时间和用户兴趣的 Twitter 信息传播者发现排序方法，并在真实的数据集上验证了方法的有效性。

3. 实验结果验证了我们的方法对于信息传播者发现是有效的，并且系统显著优于随机系统和基于用户历史转发记录的排序系统。

4. 另外，我们还发现用户历史转发信息，兴趣和活跃时间是决定信息传播者的重要因素。

5.2 相关工作

由于 Twitter 的流行性，数据的公开性，以及独特的转发属性使得 Twitter 的研究变得异常活跃。我们将相关工作分成两个部分：tweet 转发预测和 Twitter 转发行为分析。具体参见 **??**和1.3.3。但是要强调的是不同于 tweet 的转发预测，本章我们的工作集中在预测给定的 tweet 到底“谁”会转发，而 Twitter 转发行为的分析可以帮助提高发现信息传播者的效果。这里需要强调的是，据我们所知目前 Twitter 的研究中还没有对信息传播者发现的相关研究，因此我们自己构造了基准系统进行对比实验。

5.3 基于排序学习的 Twitter 信息传播者发现框架

给定一个 tweet t，它的作者为 user u，user u 的粉丝集合为 $Followers(u)$，我们的目的就是学习一个排序模型来评估每个粉丝 f_i ($f_i \in Follower(u)$) 未来转发 t 的可能性。不失一般性，我们将这个问题当做排序问题，而没有看成分类问题。另外，由于粉丝集合 $Followers(u)$ 中可能存在任意多个转发者，因此我们从 $Followers(u)$ 中选取 top-k 个评估分数最高的粉丝作为信息传播者。

5.3.1 Twitter 信息传播者发现排序学习框架

为了生成一个排序函数 F 能够根据粉丝是否转发 tweet 对其进行排序，我们将设计一些特征，并将其引入到基于排序学习的模型中。正如前几章介绍的，排

序学习是一种整合特征以数据驱动的机器学习模型。这里在训练数据中，每个粉丝 f_i 都被标注是否转发 t，一系列与 Twitter 信息传播者相关的特征都从 u 与 f_i 的关系和各自的属性中抽取，并且特征的有效性，可以通过特定的特征组合，在测试数据的排序表现中体现出来。

5.3.2 Twitter 信息传播者相关特征

一个 tweet 能否被转发，是 tweet 本身、tweet 的作者、作者的粉丝相互作用的结果。因此在 Twitter 信息传播者发现任务中，我们需要考虑三者的属性以及相互的关系。我们首先考虑 tweet 的作者与作者的粉丝以往历史的转发信息，以此捕获粉丝在转发 tweet 上是否具有偏向性。其次，粉丝本身的特征也会影响其转发行为，比如，直觉上对于同一个 tweet 转发的概率，非名人就比名人的转发概率高，因为名人更加注重声誉，对信息质量要求更高。再次，时间因素也是需要考虑的，如果转发者与信息发布者具有相同的使用 Twitter 时间，那么信息在两人之间传播的可能性就比较大。最后，tweet 的内容会影响人的转发行为，感兴趣的信息比不感兴趣的信息对人的转发行为影响更大。因此，为了在 Twitter 中找到信息传播者，我们设计了转发历史特征，用户特征，用户活跃时间特征，用户兴趣特征：

1. **转发历史特征（Retweet History Feature-RH）**：主要涉及用户以前转发 tweet 的信息，同时也包括自身 tweet 被转发的情况。

2. **用户特征（Follower Status Feature-FS）**：主要涉及用户的社交媒体属性。

3. **用户活跃时间特征（Follower Active Time Feature-FAT）**：主要涉及用户发布 tweet 的活跃时间。

4. **用户兴趣特征（Follower Interests Feature-FI）**：主要通过 tweet 的内容发现用户的兴趣爱好。

接下来，我们将具体介绍有关的特征。

5.4 转发历史特征

直觉上，如果粉丝 f_i 以前经常转发或提及 user u 的 tweet，那么粉丝 f_i 很有可能再次转发 user u 的 tweet。因此我们设计了两个特征来获取转发信息：

1. **用户转发数目（Num_fRu）**：以往历史记录中，粉丝 f_i 转发 user u 的 tweet 数目。

2. **用户提及数目（Num_fMu）**：以往历史记录中，粉丝 f_i 提及 user u 的 tweet 数目。

用户的交流是相互的，如果一方经常转发或提及对方，那么对方也很有可能转发或提及对方，因此我们设计了另外两个特征：

1. **用户被转发数目（Num_uRf）**：以往历史记录中，粉丝 f_i 的 tweet 中被 user u 转发的 tweet 数目。

2. **用户被提及数目（Num_uMf）**：以往历史记录中，粉丝 f_i 的 tweet 中被 user u 提及的 tweet 数目。

最后，我们发现有些用户（例如，垃圾用户）只转发其他人的信息，不撰写原始的 tweet，我们设计了两个特征来对这些情况进行建模：

1. **用户转发比例（Ratio_retweet）**：以往历史记录中，粉丝 f_i 的 tweet 中转发 tweet 的比例。

2. **用户提及比例（Ratio_mention）**：以往历史记录中，粉丝 f_i 的 tweet 中提及 tweet 的比例。

5.5 用户特征

信息的传播一般是从社会地位高的用户（例如，名人）流向地位低的用户，这主要跟用户的社会属性相关[153]。

我们随机抽取了 10 万条转发 tweet 进行了详细的研究分析，发现只有 38.8% 的转发是从发布 tweet 较少的用户到发布 tweet 较多的用户；仅仅 23.8% 的转发是从粉丝较少的用户到粉丝较多的用户；0.04% 的转发是从非官方验证的用户到官方验证的用户。这些统计结果充分说明了在 Twitter 中不同社会地位的用户存在不同的转发行为。

因此，我们设计了一些特征来对用户的社会属性进行描述：

1. **发布 tweet 数目（Posts）**：用户发布 tweet 的数目。

2. **粉丝数目（Followers）**：用户的粉丝数目。

3. **朋友数目（Friends）**：用户的朋友数目。

4. **分组数目（Listed）**：用户被分组的数目。

5. **验证用户（Verified）**：用户是否被官方验证。

5.6 用户活跃时间特征

Twitter 用户一般不太可能在很晚的时间与其他人进行交流，而且如果一条 tweet 在很晚发布，那么该 tweet 作者的粉丝很可能第二天忽视了这条消息，因为它很有可能淹没在大量的新消息中，而大部分 Twitter 用户浏览 tweet 的模式是从最新的信息开始。

我们随机抽取了 1 万条回复 tweet，并对其发布时间进行了分析，发现只有 12.4% 回复 tweet 发生在 00:00 到 06:00 之间，这说明用户的活跃时间有一定的规律。

因此，为了捕获这些用户交流的时间信息，我们设计了两个特征：

1. **时区时间（Timezone）**：粉丝 f_i 是否与 user u 在同一个时区。
2. **用户活跃时间（PostTimeConsis）**：以往历史记录中，粉丝 f_i 发布 tweet 不同时间的数目比例与 tweet t 发布时间的关系，选取比例值为特征值。

我们用以上两个特征来反映用户与粉丝在 Twitter 上活跃时间的一致性。

5.7 用户兴趣特征

当一个粉丝转发一条 tweet 时，通常意味着粉丝与 tweet 作者存在一些共性，例如，他们有共同的兴趣（“we both love cats”），这是一种不太明显的关系，不像 Twitter 中的朋友或粉丝关系。

这种潜在的共性可能可以帮助发现信息传播者，我们利用相似性模型（Similarity Model）来计算 tweet t 与粉丝 f_i 之前发布 tweet 的兴趣相似程度。我们将 tweet t 和粉丝 f_i 之前发布的 tweet 都表示成词向量，然后用 $tf\text{-}idf$ 表示词的权重，用余弦夹角计算两者的相似性，我们将这个特征称为“**相似兴趣（SimInterest）**”。

在计算这个特征的时候，对于每一对 tweet t 和粉丝 f_i 以往发布的 tweet，我们首先过滤到在 6 百万 tweet 中词频最高的 100 个词和词频小于 5 的词来表示词向量，具体的余弦夹角的计算利用了向量空间模型（Vector Space Model）[154]。

5.8 Twitter 信息传播者发现实验

5.8.1 Twitter 信息传播者发现实验数据

到目前为止还没有关于在 Twitter 中发现信息传播者的数据，因此我们自己构造了相关数据[1]。

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

我们从 Twitter Streaming API 中随机选取了 500 条 tweet，每个 tweet 至少被其作者的粉丝转发过一次。这些数据时间分布在 2012 年 9 月 14 日到 2012 年 10 月 1 日之间。Kwak 等人发现一半以上的转发行为发生在原始 tweet 发布一个小时以内，75% 发生在一天以内[155]，因此我们在一天以后重新抓取了 500 个 tweet，检测最后到底有哪些粉丝转发了对应的 tweet。另外，由于 Twitter API 的限制和有些受欢迎的用户粉丝很多，我们不太可能全部抓取所有的粉丝，所以对每一个 tweet，我们只研究了 tweet 作者最新的 100 个粉丝，并获取了每个粉丝最新的 200 个 tweet。

像前面介绍的，我们对每个粉丝进行了标注，转发者标注为 1，非转发者为 0。这里要强调的是，有些用户由于隐私问题未使其数据公开，更重要的是我们未对每个 tweet 抓取所有的粉丝，造成有些 tweet 并没有转发者，这就使得最后的评测结果数值偏低。

表 5.1给出了数据的一些基本情况：

表 5.1　Twitte 信息传播者发现实验数据统计信息

被转发 tweet 总数目	500
平均每个 tweet 粉丝数目	81.15
转发者总数目	257
非转发者总数目	40317

5.8.2　Twitter 信息传播者发现实验设置

我们使用 SVM Rank 来实现排序学习算法并构造排序模型[2]。排序学习依然使用线性核函数，所有的模型参数都调到最优。为了避免数据的过拟合，我们使用了 10 次交叉验证。评测指标继续使用平均准确率（Mean Average Precision-MAP）。

5.8.3　Twitter 信息传播者发现基准系统（Baseline）

Twitter 信息传播者发现是一个新的工作，目前还没有其他方法能够进行直接的比较。因此，我们自己选择了两个基准系统：

1. **Random**：对所有粉丝随机排序。

2. **RPT**：如果一个粉丝在历史上经常转发某用户的 tweet，那么在未来他很有可能继续转发，因此我们根据粉丝历史数据中（每个粉丝 200 个最新的 tweet）转发用户 tweet 的数目进行排序。

[2]下载地址：http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 5.2　Twitter 信息传播者发现特征概况

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

Random 系统是一个“弱”的基准系统，与它的比较能够反映哪些因素能够帮助 Twitter 中信息传播者发现。**RPT** 系统是一个“强”的基准系统，与它的比较能够说明我们最好的 Twitter 信息传播者发现方法的能力。

5.8.4　Twitter 信息传播者发现实验结果及分析

我们利用不同特征集合：转发历史特征、用户特征、用户活跃时间特征、用户兴趣特征，分别构造了排序系统 RH，FS，FAT 和 FI。表 5.2 简要概述了 Twitter 信息传播者发现的相关特征。

表 5.3给出了不同系统发现信息传播者的结果。我们可以看到基于用户兴趣特征的 FI 排序系统效果最好，而基于转发历史特征的 RH 系统和基于用户特征的 FS 其次，基于用户活跃时间特征的 FAT 系统没有明显效果。这说明粉丝的兴趣爱

好，粉丝的转发历史，以及粉丝的社会地位可以帮助我们发现 Twitter 中的信息传播者。

我们将所有特征整合到一个排序系统中，称为 All。这个系统达到最高的 MAP 值，高出 Random 301.4%，高出 PRT 25.6%。

表 5.3　基于不同特征组的 Twitter 信息传播者发现系统实验结果

	MAP
Random	2.17
PRT	6.93
RH	6.27*
FS	3.66*
FAT	2.91
FI	8.12*
All	8.71*†

* 和 † 分别表示排序结果显著高于 Random 信息传播者发现系统和 PRT 信息传播者发现系统（$p < 0.05$）。

接着我们分析了具体的特征对于信息传播者发现的影响。我们根据不同的特征进行单独的数据训练与测试。表 5.4给出了各个特征的排序表现，这里 PRT 的 MAP 值与 Num_fRu 的 MAP 值不同的原因是因为前者根据数值直接排序，后者作为特征基于排序学习算法进行排序。

我们可以看到转发历史特征集合中的各个特征都能显著提高检索信息传播者的效果，另外，我们还发现用户活跃时间（PostTimeConsis）对于发现信息传播者也是有效的，最后利用特征相似兴趣（SimInterest）取得的最好的排序结果充分说明粉丝转发 tweet 是根据自己的兴趣与 tweet 的内容是否匹配来进行的。

这里是我们数据中的一个例子，反映粉丝转发的历史对于检索信息传播者的有效性：

We are having a bake sale today in the Student Union from 11-2! Come buy a midday snack from the Pretty Poodles!

有个该 tweet 作者的粉丝在这条信息发布之前已经转发此作者的信息 30 多次，而且该粉丝继续转发了这条 tweet。我们的检索模型 RH 成功地将该粉丝排在了第一位。

这是另一个验证粉丝兴趣对于检索信息传播者有效的例子：

Excited to announce our debut London show. Full details here - http://t.co/P60Wc3Lj

表 5.4　基于不同特征的 Twitter 信息传播者检索系统实验结果

	MAP
Random	2.17
PRT	6.93
Num_fRu	6.83*
Num_fMu	7.08*
Num_uRf	6.20*
Num_uMf	7.62*
Retweet_Ratio	4.45*
Mention_Ratio	3.05*
Posts	3.79*
Followers	2.37
Friends	2.03
Listed	2.17
Verified	2.34
Timezone	2.37
PostTimeConsis	2.86*
SimInterest	8.12*

* 和 † 分别表示排序结果显著高于 Random 信息传播者检索和 PRT 信息传播者检索（$p < 0.05$）。

有一个该作者的粉丝转发了这个 tweet，并且该粉丝在以前的 tweet 中经常发布一些与音乐和演唱会有关的信息。我们的 FI 检索系统也成功地将其找到。

5.9　小结

在 Twitter 中寻找信息传播者可以帮助我们更有效地向其他用户传送信息，我们对于信息传播者检索的工作能够帮助其他研究者更好地了解社交媒体中信息是如何传播的。本章中我们发现粉丝转发的历史记录，粉丝的社会地位，粉丝个人的兴趣爱好对于信息传播者检索是有效的。

未来我们将设计更多的特征帮助发现信息传播者。例如，是否亲密的朋友会经常转发用户的 tweet，地理位置信息是否有所帮助等等。

第六章　总结与展望

社交媒体是一个新兴领域，本文主要围绕 Twitter 中文本特点和社交媒体特征展开研究。通过 Twitter 中的信息检索和传播分析任务，我们发现 Twitter 中的文本结构化信息和 tweet 的社交媒体信息可以帮助这些问题的解决。

Twitter 中的检索研究能够从 Twitter 的海量数据中快速找到有意义的信息，对于 Twitter 中的其他研究具有重要的意义。以往的信息检索研究主要是对图书馆文档或网页进行处理，我们针对 Twitter 数据，具体涉及了 Twitter 中的传统信息检索问题研究和 Twitter 中观点检索研究，以此解决如何在 Twitter 中找到主客观 tweet 的问题。

Twitter 中的传播分析问题，我们主要从 tweet 本身的传播和传播的受众角度进行分析，提出了 Twitter 中传播观点发现与传播者发现的问题。通过任务的定义与方法的研究，最后通过实验验证，找到了一些 tweet 文本特征和社交媒体特征与 Twitter 中信息传播的内在联系。

6.1　工作总结

本文的主要工作可以从以下四个方面来总结：

首先，针对现有 Twitter 信息检索工作忽视 tweet 文本结构信息对 tweet 排序重要性的问题，我们对 tweet 文本进行了结构化研究，以此帮助 Twitter 中的信息检索。这个工作的动机是基于普通文本和网页结构信息能够帮助传统信息检索的已有研究结论。虽然 tweet 文本短小，但是也存在结构化属性的特点。我们定义了 tweet 文本中的几个结构化模块，称之为 Twitter 积木。然后构造自动标注器，对 tweet 文本进行积木的标注。任何一个 tweet 文本都是由若干积木块排列组合而成，而 tweet 文本特定的积木组合又对应了文本特殊的属性。我们通过这种积木结构开发特征，然后结合 tweet 的社交媒体特征，将其应用到基于排序学习的 Twitter 信息检索任务中。实验结果发现我们的 tweet 文本结构化信息能够帮助 Twitter 信息检索。

其次，针对目前政府、企业、个人都通过 Twitter 来收集大量的观点帮助决策，但是并未对观点收集的基础工作观点检索展开系统研究，我们第一次提出了 Twitter 中观点检索的新问题。我们发布了 Twitter 观点检索的新语料，该语料已经作为 ICWSM 会议的常用语料供后续研究者研究使用[1]。另外，我们根据 Twitter 中观点检索与博客观点检索的不同点，利用社交媒体特征与 tweet 观点化特征，提

[1] http://www.icwsm.org/2013/datasets/datasets/

出了 Twitter 观点检索的方法，该方法在实验结果上显著优于优化的 BM25 基准系统和基于向量空间模型的基准系统。再者，我们还提出了一种基于社交媒体特征与 tweet 文本结构化信息收集近似主观化 tweet 和近似客观化 tweet 构造主观化词典的方法，该方法能够有效构造适合 Twitter 的情感词典并以此评价 tweet 的观点化程度。最后我们重新标注了 TREC Tweets2011 数据，证明了我们的 Twitter 观点检索方法在 TREC 数据上依然有效。

再次，针对 Twitter 观点检索中时常包含大量的低质量观点，而以往的研究认为转发的 tweet 通常是高质量文本，我们提出了 Twitter 中传播观点发现的新任务。我们根据新任务的特点开发了一系列特征以此提高传播观点发现的效果，这些特征包括了 tweet 的传播度特征、tweet 的观点化特征、tweet 的文本质量特征。我们在真实的数据集上进行了测试，结果验证了我们设计的特征对于传播观点发现是有效的，并且我们的方法显著优于 BM25 方法和我们的观点检索方法。另外，令我们鼓舞的是我们的方法能够在 Twitter 中预测观点是否会被转发到达人预测的水平。

最后，针对以往 tweet 转发预测研究中忽视“谁”转发的问题的重要性，我们研究了在 Twitter 中发现信息传播者的问题。我们定义了 Twitter 中信息传播者发现的新任务，以此帮助理解 Twitter 中信息是如何传播的。我们同样开发了一系列特征，并将其应用到排序学习的机器学习框架中，具体的特征包括用户历史的转发信息，用户自身的社交媒体特征，用户使用 Twitter 的活跃时间，以及用户的个人兴趣。由于以往没有相同的工作，因此我们自己构造了数据，并发布了数据供以后的研究者继续使用。实验结果证明了我们方法对于 Twitter 中信息传播者发现是有效的，方法优于随机系统和基于用户历史转发记录的排序系统。最终我们发现用户历史转发信息，兴趣和活跃时间是决定信息传播者的重要因素。

6.2 工作展望

展望未来，社交媒体中的信息检索和传播分析研究及其相关方向还有很多工作需要完成。这里总结以下亟待探索的研究方向和路线:

1. 以 Twitter 为代表的社交媒体一个重要特点就是消息的实时性，许多研究工作都围绕在 Twitter 中发现实时信息展开，包括新事件发现[156–162]、实时灾害报道（如地震、疾病、火灾等）[163–167]，另外，TREC 的 Twitter 检索[168–174]也将实时性作为一个重要指标。本文的研究中，我们并未对话题检索和观点检索深入讨论实时性对检索效果的影响。这个问题的关键是找到与话题相关的时间点，如何找到这个相关时间点是我们未来研究的一个重点。

2. 本文的社交媒体研究仅仅以 Twitter 为代表展开，实际上流行的社交媒体还有很多，如 Facebook[2]、YouTube[3]、Flickr[4]等等。这些社交媒体肯定有自己独特的特点，在其数据上进行检索任务和传播分析需要研究其特殊性；另外，未来一个可能的需求就是多种社交媒体综合检索和跨媒体的信息传播，这就需要研究者在充分理解各种社交媒体的特点和人们对各种社交媒体不同的需求上，提出方法解决问题。

3. 目前跨媒体之间的研究是一个新的研究方向，主要是基于不同媒体之间的差异，利用各自的优点，解决其他媒体存在的问题。例如，有的研究利用维基百科的知识，帮助扩展 tweet 文本的语义，以此克服 tweet 文本短小，信息缺失的缺点[175, 176]；有些研究利用维基百科的访问信息帮助 Twitter 中的事件发现[177]；还有些研究各种媒体之间的联系以此帮助其他任务的解决[178–183]等等。未来我们将利用其他社交媒体的优点，帮助 Twitter 中已有的信息检索和传播分析任务进一步提高效果。

总之，社交媒体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。

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

致　谢

时间应该“浪费”在美好的事物上，我很庆幸自己有机会从事这些课题的研究。

本课题承蒙国家自然科学基金 -面上项目“融合网络特征的文本观点挖掘”(项目号：61170156）和国家自然科学基金 -青年科学基金项目“结合社会网络的网络信息传播分析研究”(项目号：61202337) 的资助，另外感谢国家留学基金委对本人在英国爱丁堡大学留学访问的经费资助。经济基础决定上层建筑，没有钱一切白搭。

感谢我的导师王挺教授对本人的精心指导和悉心培养，您严谨治学的态度让我受益终生。感谢在我英国爱丁堡大学留学期间指导我的 Miles Osbonre 博士，您对科研的敏锐洞察力以及极强的逻辑思维能力让我明白科研可以轻松地“玩”。

感谢国防科技大学自然语言处理组的张晓艳、刘伍颖、唐晋韬、魏登萍、周云、李岩、麻大顺、谢松县、刘培磊、岳大鹏、刘海池、汝承森、张文文、姜仁会、胡长龙、李欣奕，和你们一同探索自然语言处理的未知领域让人回味；感谢英国爱丁堡大学信息学院 348 办公室的 Saša Petrovic、Desmond Elliott、Diego Frassinelli、Eva Hasler、Michael Auli 和 Luke Shrimpton，和你们仔细讨论我的课题细节以及英文的论述让我十分收益。

感谢戴波、雷鸣、林正帅、毛先领、张湘莉兰、任洪广、王鹤、吴诚堃标注 Twitter 观点检索相关数据，感谢王铮标注 Twitter 传播观点检索相关数据。没有你们的无偿帮助，我相信我的课题研究不会如此顺利。感谢 Victor Lavrenko 博士和 Micha Elsner 博士给予我 Twitter 观点检索课题宝贵的意见，和你们讨论是我的荣幸。

感谢我从硕士到博士的室友邹丹、何明、陆化彪，一起研究课题慢慢“变老”的过程值得怀念；感谢我在爱丁堡期间最好的朋友王鹤和黄轩，人生得不多的知己足以；感谢在英国一同留学的杨俊刚、雷鸣、Chee-Ming Ting、陆亮、王铮、刘哲、马瑞、冯翌尧、卢恒、林正帅、曾旋、杨国利、何鑫、胡尽力、贺建森、杨丽莎、魏杰、罗家希，谢谢你们让我拥有那些留学的“回忆”，谢谢你们陪我那时的“孤独”。

感谢我的父母、奶奶、岳父、岳母对我生活上无微不至的照顾；最后感谢我的爱人柳意，谢谢你的支持，爱你！

参考文献

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

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

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

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

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

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

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

[17] Wiebe J M. Tracking point of view in narrative [J]. Computational Linguistics. 1994, 20 (2): 233–287.

[18] 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.

[19] Liu Y, Huang X, An A, et al. ARSA: a sentiment-aware model for predicting sales performance using blogs [C]. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007: 607–614.

[20] Oghina A, Breuss M, Tsagkias M, et al. Predicting imdb movie ratings using social media [M] // Oghina A, Breuss M, Tsagkias M, et al. Advances in information retrieval. Springer, 2012: 2012: 503–507.

[21] Joshi M, Das D, Gimpel K, et al. Movie reviews and revenues: An experiment in text regression [C]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 293–296.

[22] Sadikov E, Parameswaran A G, Venetis P. Blogs as Predictors of Movie Success. [C]. In ICWSM. 2009.

[23] Asur S, Huberman B A. Predicting the future with social media [C]. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. 2010: 492–499.

[24] 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.

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

[26] 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.

[27] Chen L, Wang W, Sheth A P. Are Twitter users equal in predicting elections? A study of user groups in predicting 2012 US Republican Presidential Primaries [M] // Chen L, Wang W, Sheth A P. Social informatics. Springer, 2012: 2012: 379–392.

[28] Metaxas P T, Mustafaraj E, Gayo-Avello D. How (not) to predict elections [C]. In Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom). 2011: 165–171.

[29] Gayo-Avello D, Metaxas P T, Mustafaraj E. Limits of electoral predictions using twitter. [C]. In ICWSM. 2011.

[30] Armstrong J S, Graefe A. Predicting elections from biographical information about candidates: A test of the index method [J]. Journal of Business Research. 2011, 64 (7): 699–706.

[31] Hong L, Dan O, Davison B D. Predicting popular messages in twitter [C]. In Proceedings of the 20th international conference companion on World wide web. 2011: 57–58.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] Liu F, Weng F, Wang B, et al. Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision. [J]. ACL (Short Papers). 2011, 11: 71–76.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

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

[50] Ritter A, Cherry C, Dolan B. Unsupervised modeling of Twitter conversations [C]. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 172–180.

[51] 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.

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

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

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

[59] 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.

[60] Brody S, Diakopoulos N. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs [C]. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 562–570.

[61] Rocchio J J. Relevance feedback in information retrieval [J]. 1971.

[62] Joachims T. Optimizing search engines using clickthrough data [C]. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002: 133–142.

[63] Xue G-R, Zeng H-J, Chen Z, et al. Optimizing web search using web click-through data [C]. In Proceedings of the thirteenth ACM international conference on Information and knowledge management. 2004: 118–126.

[64] Joachims T, Granka L, Pan B, et al. Accurately interpreting clickthrough data as implicit feedback [C]. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005: 154–161.

[65] Radlinski F, Joachims T. Query chains: learning to rank from implicit feedback [C]. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005: 239–248.

[66] Agichtein E, Brill E, Dumais S. Improving web search ranking by incorporating user behavior information [C]. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 19–26.

[67] Agichtein E, Brill E, Dumais S, et al. Learning user interaction models for predicting web search result preferences [C]. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 3–10.

[68] Radlinski F, Joachims T. Active exploration for learning rankings from clickthrough data [C]. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007: 570–579.

[69] Bao S, Xue G, Wu X, et al. Optimizing web search using social annotations [C]. In Proceedings of the 16th international conference on World Wide Web. 2007: 501–510.

[70] Duh K, Kirchhoff K. Learning to rank with partially-labeled data [C]. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008: 251–258.

[71] Aslam J A, Kanoulas E, Pavlu V, et al. Document selection methodologies for efficient and effective learning-to-rank [C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009: 468–475.

[72] Qin T, Liu T-Y, Xu J, et al. LETOR: A benchmark collection for research on learning to rank for information retrieval [J]. Information Retrieval. 2010, 13 (4): 346–374.

[73] Xu J, Chen C, Xu G, et al. Improving quality of training data for learning to rank using click-through data [C]. In Proceedings of the third ACM international conference on Web search and data mining. 2010: 171–180.

[74] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent [C]. In Proceedings of the 22nd international conference on Machine learning. 2005: 89–96.

[75] Cao Y, Xu J, Liu T-Y, et al. Adapting ranking SVM to document retrieval [C]. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 186–193.

[76] Xu J, Li H. Adarank: a boosting algorithm for information retrieval [C]. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007: 391–398.

[77] Quoc C, Le V. Learning to rank with nonsmooth cost functions [J]. Proceedings of the Advances in Neural Information Processing Systems. 2007, 19: 193–200.

[78] Xu J, Liu T-Y, Lu M, et al. Directly optimizing evaluation measures in learning to rank [C]. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008: 107–114.

[79] Valizadegan H, Jin R, Zhang R, et al. Learning to rank by optimizing ndcg measure [C]. In Advances in neural information processing systems. 2009: 1883–1891.

[80] Wang L, Lin J, Metzler D. Learning to efficiently rank [C]. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 138–145.

[81] Dai N, Shokouhi M, Davison B D. Learning to rank for freshness and relevance [C]. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011: 95–104.

[82] Chapelle O, Chang Y, Liu T-Y. Future directions in learning to rank. [J]. Journal of Machine Learning Research-Proceedings Track. 2011, 14: 91–100.

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] Stajner T, Thomee B, Popescu A, et al. Automatic selection of social media responses to news [J]. ACM WSDM. 2013.

[93] Kothari A, Magdy W, Kareem Darwish A M, et al. Detecting Comments on News Articles in Microblogs [J]. ICWSM 2013. 2013.

[94] Li C, Sun A, Weng J, et al. Exploiting hybrid contexts for Tweet segmentation [C]. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013: 523–532.

[95] Zhang J, Minami K, Kawai Y, et al. Personalized Web Search Using Emotional Features [M] // Zhang J, Minami K, Kawai Y, et al. Availability, Reliability, and Security in Information Systems and HCI. Springer, 2013: 2013: 69–83.

[96] Luo Z, Osborne M, Petrovic S, et al. Improving Twitter Retrieval by Exploiting Structural Information. [C]. In AAAI. 2012.

[97] Luo Z, Osborne M, Wang T. Opinion Retrieval in Twitter. [C]. In ICWSM. 2012.

[98] Luo Z, Wang T. Propagated Opinion Retrieval in Twitter. [C]. In WISE. 2013.

[99] 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.

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

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

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

[103] Liu B. Sentiment Analysis and Opinion Mining [M]. Morgan & Claypool Publishers, 2012.

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

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

[106] 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.

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

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

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

[110] 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.

[111] 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.

[112] 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.

[113] 知网 HowNet 评价词词典. 2013.

[114] 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.

[115] 情感词汇本体库. 2013.

[116] 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.

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

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

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

[120] 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.

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

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

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

[124] 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.

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

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

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

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

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

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

[131] 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.

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

[133] 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.

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

[135] 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.

[136] 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.

[137] Jiang L, Yu M, Zhou M, et al. Target-dependent Twitter Sentiment Classification. [C]. In ACL. 2011: 151–160.

[138] 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.

[139] 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.

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

[141] 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.

[142] 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.

[143] 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.

[144] 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.

[145] 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.

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

[147] 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.

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

[149] 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.

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

[151] Artzi Y, Pantel P, Gamon M. Predicting responses to microblog posts [C]. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 602–606.

[152] Nagarajan M, Purohit H, Sheth A P. A Qualitative Examination of Topical Tweet and Retweet Practices. [C]. In ICWSM. 2010.

[153] Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in the flickr social network [C]. In Proceedings of the 18th international conference on World wide web. 2009: 721–730.

[154] Salton G, Wong A, Yang C-S. A vector space model for automatic indexing [J]. Communications of the ACM. 1975, 18 (11): 613–620.

[155] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media? [C]. In Proceedings of the 19th international conference on World wide web. 2010: 591–600.

[156] 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.

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

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

[159] 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.

[160] 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.

[161] 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.

[162] 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.

[163] 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.

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

[165] 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.

[166] 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.

[167] 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.

[168] 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.

[169] 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.

[170] 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.

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

[172] 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.

[173] 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.

[174] 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.

[175] Meij E, Weerkamp W, de Rijke M. Adding semantics to microblog posts [C]. In Proceedings of the fifth ACM international conference on Web search and data mining. 2012: 563–572.

[176] Cassidy T, Ji H, Ratinov L-A, et al. Analysis and Enhancement of Wikification for Microblogs with Context Expansion. [C]. In COLING. 2012: 441–456.

[177] Osborne M, Petrovic S, McCreadie R, et al. Bieber no more: First story detection using Twitter and Wikipedia [C]. In Proceedings of the SIGIR Workshop on Time-aware Information Access. 2012.

[178] Dong A, Zhang R, Kolari P, et al. Time is of the essence: improving recency ranking using twitter data [C]. In Proceedings of the 19th international conference on World wide web. 2010: 331–340.

[179] Phelan O, McCarthy K, Bennett M, et al. On using the real-time web for news recommendation & discovery [C]. In Proceedings of the 20th international conference companion on World wide web. 2011: 103–104.

[180] Tsagkias M, de Rijke M, Weerkamp W. Linking online news and social media [C]. In Proceedings of the fourth ACM international conference on Web search and data mining. 2011: 565–574.

[181] Becker H, Iter D, Naaman M, et al. Identifying content for planned events across social media sites [C]. In Proceedings of the fifth ACM international conference on Web search and data mining. 2012: 533–542.

[182] Petrovic S, Osborne M, McCreadie R, et al. Can Twitter replace Newswire for breaking news? [C]. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. 2013.

[183] Chang Y, Dong A, Kolari P, et al. Improving recency ranking using twitter data [J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2013, 4 (1): 4.

作者在学期间取得的学术成果

发表的学术论文

[1] Zhunchen Luo, Miles Osborne, Sasa Petrovic and Ting Wang. Improving Twitter Retrieval by Exploiting Structural Information. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (**AAAI 2012**), Toronto, Canada, July 2012. （**CCF A 类会议，人工智能领域顶级会议**）

[2] Zhunchen Luo, Miles Osborne, Jintao Tang and Ting Wang. Who Will Retweet Me? Finding Retweeters in Twitter. In *Proceedings of the Thirty-Sixth International ACM SIGIR Conference on Research and Development in Information Retrieval* (**SIGIR 2013**), Dublin, Ireland, July 2013.（**CCF A 类会议，信息检索领域顶级会议，获得会议旅行奖金 1300 美元**）

[3] Zhunchen Luo, Miles Osborne and Ting Wang. Opinion Retrieval in Twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (**AAAI-ICWSM 2012**), Dublin, Ireland, June 2012. （**社交媒体领域顶级会议，获得会议旅行奖金 300 美元**）

[4] Zhunchen Luo, Miles Osborne and Ting Wang. An Effective Approach to Tweets Opinion Retrieval. To appear in **World Wide Web Journal**. （**SCI 期刊，影响因子 1.196**）

[5] Zhunchen Luo, Jintao Tang and Ting Wang. Propagated Opinion Retrieval in Twitter. In *Proceedings of the Fourteenth International Conference on Web Information System Engineering* (**WISE 2013**), Nanjing, China, October 2013.（**CCF C 类会议，信息检索与数据挖掘领域重要会议**）

[6] Zhunchen Luo, Jintao Tang and Ting Wang. Improving Keyphrase Extraction from Web News by Exploiting Comments Information. In *Proceedings of the Fifteenth International Asia-Pacific Web Conference* (**APWeb 2013**), Sydney, Australia, April 2013. （**CCF C 类会议，信息检索与数据挖掘领域重要会议**）

[7] Zhunchen Luo, Lan Rao, Chengsen Ru and Ting Wang. Finding High-Quality Posts from Microblogging Conversations. In *the Eighth International Conference on Modeling Decisions for Artificial Intelligence (**MDAI 2011**)*, Changsha, China, July, 2011.

[8] 罗准辰，王挺. 基于分离模型的中文关键词提取算法研究. 中文信息学报, 2009，23 (01): 63-70.

[9] 罗准辰，王挺. 搜索词同现网络研究. 第六届全国信息检索学术会议 (**CCIR 2010**)，镜泊湖，2010 年 8 月.