

硕士学位论文

**基于社会网络的情感分类算法研究**

学位申请人：员鹏

指导教师： 饶元 副教授

类别（领域）：工程硕士（软件工程）

2016年4月

**Research of Sentiment Classification Algorithm Based On Social Network**

A thesis submitted to

Xi’an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

By

Peng Yun

Supervisor: Accoc.Prof. Yuan Rao

(SoftWare Engineering)

April 2016

**论文题目：基于社会网络的情感分类算法研究**

**类别（领域）：工程硕士（软件工程）**

**学位申请人： 员鹏**

**指导教师： 饶元 副教授**

摘 要

随着 Web2.0 的高速发展，各种社交论坛的出现丰富了用户使用互联网的方式，越来越多的用户参与其中进行互动。这些社交论坛的兴起不仅使网民能更容易表达自己的观点和态度，而且也提升了各种信息的传播速度。在公共管理领域，如果不对网络谣言或舆情进行引导和控制，网民的情感很容易出现群体极化现象。但是，由于情感资源不够完善且情感计算的复杂性等特点，情感传播特征不能得到很好的研究，群体极化现象也不容易发现和预防。

本文以社交网络文本为研究对象，通过文本情感分析技术及社会网络分析方法来对其进行分析，然后情绪传播角度对事件进行阐述，希望能及时发现网络中的负面情绪避免事态严重化。本文主要工作包括以下三个方面：首先，通过现有的情感资源构造了基础情感词典，制定规则运用自然语言处理技术对实体及评价词与实体及其行为进行抽取，并通过共现特性进行极性强度量化构建出实体评价及实体行为情感词典。其次，对不同的情感特征组合设计，利用分类算法进行实验，通过对结果评测来选取最优特征组合，使实验效果达到最佳。最后在情感传播的模型上，运用社会网络分析方法对构建的情感网络进行分析，发现可以通过监控和引导网络中的几个核心节点可以避免事件的严重化，除此之外，发现可以通过事件时间点上的正负向的用户变化情况来对事件及时预警，避免事态的扩大化。

在论坛数据集上，通过实验对比发现，实体词典可以提高情感分析的准确率，同时发现选取合适的特征可以避免噪音数据的影响，在该实验中，情感极性分类准确率可以达到81.4%，召回率可以达到90.3%，情感强度分类准确率可以达到73.3%，召回率可以达到85.62%，通过算法对比发现SVM的分类效果最好。最后对情感网络特征进行分析，发现论坛构建的情感网络满足互惠性理论及同质性理论，不满足平衡理论及状态理论。除此之外，发现用户情感分类趋势图更容易监控舆情爆发的时间点。

**关 键 词**：情感分析；情感词典；社会网络；情感传播；情感分类

**论文类型**：应用研究

**Title: Research of Sentiment Classification Algorithm Based On Social Network**

**Descipline: SoftWare Engineering**

**Applicant: Peng Yun**

**Supervisor: Accoc.Prof. Yuan Rao**

ABSTRACT

With the rapid development of Web2.0, the emergence of various social forums enrich the user use the Internet, more and more users involved interact. The rise of these social media not only enables users to freely express their opinions, attitudes, emotions and other subjective information, but also accelerated the dissemination of information of all kinds. In the field of public administration, if not the network public opinion or rumor control and guide users emotions are prone to the phenomenon of group polarization. However, due to inadequate resources and emotional affective computing complexity and other characteristics, propagation characteristics of emotion not been well studied, the phenomenon of group polarization is not easy to detect and prevent.

In this thesis, a social network text for the study, analysis and social network analysis method to analyze the text emotion and sentiment spread angle of the incident described, hoping to timely detect negative emotions to avoid aggravation of the situation. The main work includes the following three aspects: First, the existing structure of the emotional resources base emotions dictionary rules using natural language processing techniques and evaluation of physical entities and their words and acts of extraction, and polar properties through co-occurrence strength evaluation and quantification constructed entity entity emotional behavior dictionary. Secondly,Different combinations of features of emotional design, the use of classification algorithms conduct the experiments, the results of the evaluation to select the best combination of features, so experiment to achieve the best. Finally, on the model of the spread of emotion, the use of social network analysis to build emotional network analysis, can be found by monitoring and guiding several core network nodes to avoid aggravation of the event, in addition, it can be found through the event positive and negative to the point in time the user changes to the timely warning to the event, to avoid the situation of expansion.

At the forum data sets, by contrast experiments show that an entity dictionary can improve the accuracy of sentiment analysis, also found that select the appropriate characteristics to avoid the effects of noise data, emotional polarity classification accuracy rate can reach 81.4% recall rate can reach 90.3 % emotional intensity classification accuracy rate can reach 73.3% recall rate can reach 85.62%, by comparison found SVM classification algorithm works best. Finally, the emotional network characteristics were analyzed and found to meet the network forum to build emotional reciprocity theory and the theory of homogeneity is not satisfied

**KEY WORDS**: Sentiment analysis; Emotional dictionary; Social networks; Dissemination of emotion; Emotion classification

**TYPE OF THESIS**: Application Research

# 

目 录

[1 绪论 1](#_Toc449535434)

[1.1 研究背景及意义 1](#_Toc449535435)

[1.2 国内外研究现状 1](#_Toc449535436)

[1.3 论文的主要工作 3](#_Toc449535437)

[1.4 论文的组织机构 4](#_Toc449535438)

[2 相关理论与技术 5](#_Toc449535439)

[2.1 文本特征选择 5](#_Toc449535440)

[2.2 情感分类算法及评价指标 6](#_Toc449535441)

[2.2.1 基于词典和规则相结合的方法 6](#_Toc449535442)

[2.2.2 基于统计的方法 6](#_Toc449535443)

[2.2.3 评价指标 9](#_Toc449535444)

[2.3 情感网络特征 11](#_Toc449535445)

[2.3.1 互惠性 11](#_Toc449535446)

[2.3.2 基于结构平衡理论和状态理论的情感关系研究 11](#_Toc449535447)

[2.3.3 情感分布同质性 12](#_Toc449535448)

[2.4 社会网络指标 12](#_Toc449535449)

[2.5 本章小结 13](#_Toc449535450)

[3 情感词典的构建 14](#_Toc449535451)

[3.1 传统情感词典构建与不足 14](#_Toc449535452)

[3.1.1 传统情感词典构建 14](#_Toc449535453)

[3.1.2 传统情感词典的不足 14](#_Toc449535454)

[3.2 基本情感词典的构建 15](#_Toc449535455)

[3.3 基于实体情感词典的构建 17](#_Toc449535456)

[3.3.1 基于实体及属性的情感词典的构建 17](#_Toc449535457)

[3.3.2 基于实体行为情感词典的构建 22](#_Toc449535458)

[3.4 本章小结 24](#_Toc449535459)

[4 情感分类研究 25](#_Toc449535460)

[4.1 传统情感分类及不足 25](#_Toc449535461)

[4.2 特征选择 26](#_Toc449535462)

[4.3 实验及结果分析 26](#_Toc449535463)

[4.3.1 实验数据 26](#_Toc449535464)

[4.3.2 基于词典与规则结合的方法 27](#_Toc449535465)

[4.3.3 基于机器学习的情感分类 29](#_Toc449535466)

[4.4 本章小结 32](#_Toc449535467)

[5 基于社会网络情感分析 33](#_Toc449535468)

[5.1 传统文本情感分析 33](#_Toc449535469)

[5.2 基于社会网络情感分析 33](#_Toc449535470)

[5.3 基本模型与指标 34](#_Toc449535471)

[5.3.1 基本模型 34](#_Toc449535472)

[5.3.2 基本指标 35](#_Toc449535473)

[5.4 实验及结果分析 36](#_Toc449535474)

[5.4.1 数据抓取 36](#_Toc449535475)

[5.4.2 数据预处理 37](#_Toc449535476)

[5.4.3 数据处理 40](#_Toc449535477)

[5.4.4 结果及分析 43](#_Toc449535478)

[5.5 本章小结 51](#_Toc449535479)

[6 结论与展望 52](#_Toc449535480)

[6.1 结论 52](#_Toc449535481)

[6.2 展望 52](#_Toc449535482)

[致 谢 54](#_Toc449535483)

[参考文献 55](#_Toc449535484)

[附 录 57](#_Toc449535485)

[攻读学位期间取得的研究成果 61](#_Toc449535486)

声明

CONTENTS

1 Preface 1

1.1 Background and Significance 1

1.2 Research Status 1

1.3 Main Work of this Thesis 3

1.4 Struture of this Thesis 4

2 Theory and Related Technology of Sentiment Classification 5

2.1 Text Feature Selection 5

2.2 Sentiment Classification and Evaluation 6

2.2.1 Method Based on Dictionary and Rule 6

2.2.2 Based on Statistical Method 6

2.2.3 Evaluation 9

2.3 Emotional Network Features 11

2.3.1 Reciprocity 11

2.3.2 Structural Balance Theory and the Theory of the State 11

2.3.3 Emotional Distribution Homogeneity 12

2.4 Social Networking Index 13

2.5 Summary of Chapter 13

3 Construction of Emotion Dictionary 14

3.1 Construction and Lack of Emotion Traditional Dictionary 14

3.1.1 Construction of the Traditional Emotional Dictionary 14

3.1.2 The Shortcomings of Traditional Emotional Dictionary 14

3.2 Construction of Basic Emotions Dictionary 15

3.3 Entity-based Build Emotional Dictionary 17

3.3.1 Construction of Entities and Attributes Based On Emotion Dictionary 17

3.3.2 Construction Emotional Entity-Behavior Based dictionary 22

3.4 Summary of Chapter 24

4 Emotion Classification 25

4.1 The Shortcomings of Traditional Sentiment Classification 25

4.2 Feature Selection 26

4.3 Experiment and Result Analysis 26

4.3.1 Experimental Data 26

4.3.2 Method Based on Dictionary and Rule 27

4.3.3 Machine learning based approach 29

4.4 Summary of Chapter 32

5 Sentiment Analysis Based on Social Network 33

5.1 Traditional Text Sentiment Analysis 33

5.2 Sentiment Analysis Based on Social Network 33

5.3 The Basic Model and Indicators 34

5.3.1 Basic Model 34

5.3.2 Basic Indicators 35

5.4 Experimental Results and Analysis 36

5.4.1 Data Capture 36

5.4.2 Data Preprocessing 37

5.4.3 Data Processing 40

5.4.4 Results and Analysis 43

5.5 Summary of Chapter 51

6 Summary and Outlook 52

6.1 Summary of this Thesis 52

6.2 Outlook of Research 52

Acknowledgements 54

References 55

Appendix 57

Research Results Obtained During the Degree 61

Declaration

# 绪论

## 研究背景及意义

随着 Web2.0 的高速发展，互联网逐渐形成“以用户为中心，全民参与”的开放式构架理念，注重信息与用户之间的交互。从逻辑上讲，互联网用户由最初的“读”信息向“写”信息互联网发展，并且从简单地接收互联网内容转为主动地创造互联网内容，即用户由浏览者演变为既是浏览者又是创造者。

因此，互联网上出现了大量的用户互动信息，而且形式上也更加丰富，如BBS，电子商务和微博等，用户主动发布博客和参与讨论和的热情与日俱增，带有个人主观情感色彩的文章和评论也随之而来。由于评论信息量的快速增加，仅靠简单的人工方法已无法处理和分析这些海量信息，因此急需机器智能地为用户快速收集、处理和分析这些相关的评论互动信息，在这样的背景下情感分析技术便应运而生。情感分析是指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取原素材中的主观信息。通常来说，情感分析的目的是为了找出说话者/作者在某些话题上或者针对一个文本两极的观点的态度。目前，文本情感分析的各方面技术均有所发展，情感分析方法也逐渐完善，但是由于人们表达情感的随意性及情感本身的复杂性，情感分类的准确度提高并不是很明显，主要是情感词典不完善，情感特征提取的不充分造成的，所以丰富情感词典并充分的挖掘情感特征来提高情感计算的准确度就显得很有必要。

社交媒体的兴起不仅能使网民更自由地表达自己的观点和态度，而且也提升了各种信息的传播速度。在公共管理领域，如果不对网络谣言或舆情进行引导和控制，网民的情感很容易出现群体极化现象，进而导致社交网络中的群体性事件，甚至导致实际生活中的暴力性事件。如2007年的涉日游行事件、2009年的新疆打砸事件和2012年的钓鱼岛事件等均受到了网络谣言的激发和推动。所以，为避免不良情绪的集聚和激化，有必要对网上各种情绪的传播规律进行分析，寻找影响情绪传播的关键因素。随着社会网络分析方法的发展，基于社交网络情绪传播规律的发现及可视化成为可能，它有助于弥补方法论个体主义的局限，将传播主体的个体属性与社会结构因素连接起来，从而为信息传播现象提供了一个富有解释力的全新视角。

## 国内外研究现状

情感分析研究的历史并不是太长，2000年左右开始逐渐成为自然语言处理和文本挖掘领域的热门话题。文本情感分析研究尚处于起步阶段，取得了部分研究成果，但也需要多学科知识的协同研究，面临着许多亟待解决的理论难题。本章节从情感词典构建、情感分类及基于社会网络分析方法的情感研究这三方面来阐述存在的问题和不足。

#### 情感词典

情感词是研究文本情感分析的基础，情感词典研究主要分为情感词获取、情感词极性判定及量化计算与情感词的存储管理等工作。目前基于字、词、段、篇的情感分析，都十分依赖情感词典这一资源。目前的情感词获取主要通过现有情感词典和语料库进行抽取，或者利用这些资源获取具体领域的情感词典，如Baccianella等人[1]以 General Inquirer（GI）词典为基础来构建情感词典，而 Devitt等人[2]和Esuli等人[3]都在SentiWordNet基础上来进行情感词的极性进行判别。Gyamfi等人[4]则在MPQA语料库的基础上构建了情感种子库，并利用 WordNet来对情感词典进行补充。总而言之，这些研究基本上是通过已有的情感词典或其他情感语料库抽取出情感种子词，利用语义相似度和互信息（PMI）等方法来获取情感词并对其极性进行判别。

在国内，情感词典的构建才刚刚开始。比如董振东等人研制的知网[5]（HowNet）目前已包含情感词典，该词典主要包括程度副词、正向评价词语、负向评价词语、正向情感词语以及负向情感词语等。柳位平等人[6]从HowNet的情感词典筛选了基础情感词来对情感词典进行构建，但是其准确度比较依赖于HowNet的情感词集的准确度。徐琳宏等人[7]构建了情感本体词汇，并对每个情感词都进行情感类别、强度和极性的描述。

总的来说，目前的情感词只针对单个字或者词语，由于许多字和词在不同的环境下的情感并不相同甚至相反，所以要对情感词典进行更细化的研究，明确不同字词在不同语义或者领域环境下的情感。而现有的情感词典的研究主要是通过对情感词进行扩展及对这些词进行极性分类，且情感词典的情感值通常需要人工进行标注，情感值量研究比较少。

#### 情感分类

情感分类，即对文本进行情感极性及情感强度进行分类判别。情感极性分类，即将主观本文的进行褒贬两类；情感强度分类，即对文本情感极性的强度更细致的划分。

##### 情感极性分类

目前情感极性的分类主要存在两种研究思路：基于情感词典的方法和基于机器学习的方法。基于情感词典的方法主要是通过现存一些的情感词典、情感领域词典以及带有极性的评价单元来进行计算，来获取主观文本的极性。一些研究者通过分析文本中情感单元的极性，然后对极性进行加权求和，这些方法的研究重点包括对情感词语或组合评价单元的抽取和情感单元的极性判断。基于机器学习的方法主要通过选取大量有意义的特征来对文本情感进行分类。Pang[8]将机器学习的方法首次应用到篇章级文本的情感分类中。他们尝试使用了n-gram词语特征和词性特征，并在NB、ME和SVM这3种分类模型上进行了分类实验，发现unigram作为特征效果最好。Cui[9]通过分类实验证明，当训练语料较少时，unigram作为特征的效果较优；但是，随着训练语料的增加，n-gram(n>3)作为特征的效果越来越好。Kim[10]除了考察传统的n-gram模型之外，还引入了位置特征和评价词特征来对句子级文本进行褒贬分类。

##### 情感强度分类

除了对文本情感的褒贬二分类之外，还有一些研究工作对情感强度进行了更细的分类。Pang[11]将文本情感褒贬等级分为3类，并使用了one-vs-all多元分类算法和回归分类算法对文本完成情感分类。Goidberg[12]使用了基于图的半指导的分类算法把文本的情感强度分为了4个等级。

从上述可以看出，不论是情感极性分类及情感强度分类，情感特征的选择对于情感分类效果有着直接的影响，所以通过选取或抽取有效的情感特征可以提高情感分类的准确度。

#### 基于社会网络的情感分析

国外许多研究人员针情绪传播做了许多研究，他们认为情绪在社交网络中是可以传播的，而且提出了情绪传播模型，Hatfield、Cacioppo和Rapson [13]将情绪感染定义为“人们倾向于自动模仿并自动使自身的行动、表情、姿势和发音与其他人同步，由此达到情绪的一致。”Hatfield等（1993）[14]解释了情绪感染的机制和原理。Hancock 等 [15]研究了在基于网络的交流中人们表述和发现情绪的方式，验证了社会信息处理理论，结果显示在基于网络的交流环境中情绪可以通过消极情感词、标点符号等方式进行表达和接收，在对情绪的察觉方面，参与者对情绪的判断很敏锐，能很快地区分积极和消极情绪。

目前国内情感分析的研究主要集中于对网民发布的信息进行情感的判断和挖掘，并对不同情感的强度进行统计，或者进行情感相关因素的分析研究，如互联网情感与社会重大影响事件的相关程度；通过情感分析结果来预测电影票房、政治选举结果和股市波动等；或将微博中的情感词与网民观点相结合进行分析，而对互联网中网民情感传播特征的研究还比较少。

## 论文的主要工作

不同于以往的只通过文本分类及自然语言处理技术来分析情感问题的方法，本文运用社会网络分析方法来分析参与事件讨论的个体情感及事件的整体情感趋势，主要研究工作如下：

1）在大连理工大学中文情感词汇本体库的基础上，结合知网的“情感分析用词语集（beta版）”、台湾大学NTUSD中文情感极性词典，进行人工筛选、量化并整理得到基本情感词典，该基本情感词典包括褒贬词、否定词及程度副词。

2）利用自然语言处理中的句法分析技术和语义角色标注技术构建实体属性及行为情感词典，并通过与基础褒贬词的共现特性来对实体情感词典进行极性分类，同时制定策略对实体情感词典进行强度量化。

3）在不同的情感词典组合上，运用词典和规则的分类方法对情感进行分类，并对实验结果进行分析；对不同的情感特征组合设计，利用机器学习分类算法进行实验，通过对结果评测来选取最优特征组合，使实验效果达到最佳。

4）通过情感网络数据统计，从互惠性原则、平衡理论及状态理论及同质性理论对论坛情感网络进行特征分析。

5）建立实体用户与文本内容及用户与用户的情感传播模型，通过实体用户与文本内容模型来刻画用户的情感偏向，并通过用户情感之间的感染传播来进一步细化用户的情感传播趋势，最后对事件及用户随时间的情感趋势变化进行分析与阐述。

## 论文的组织机构

第一章 绪论

主要介绍了本文的研究背景和意义，从不同的角度阐述了文本情感分析的研究现状及存在的问题，并提出对部分问题的解决方案。

第二章 相关理论与技术

首先介绍了情感分析相关的一些理论与技术，包括中特征选择方法、情感分类的算法及评价指标等，然后对情感网络及社会网络指标进行了介绍。

第三章 情感词典构建

通过现有的情感资源构造了基础情感词典，基础情感词典包括基本情感词典、否定情感词典及程度修饰词典；通过论坛语料构造了网络情感词典；制定规则运用自然语言处理技术对实体及评价词与实体及其行为抽取，并通过共现特性进行极性强度量化构建出实体评价及实体行为情感词典。

第四章 情感分类算法研究及评测

首先在不同的情感词典组合上，运用词典和规则的分类方法对情感进行分类，并对实验结果进行分析。然后对不同的情感特征进行组合设计，运用朴素贝叶斯分类算法、SVM分类算法、kNN分类算法进行对比实验，最后通过评测得到最优情感特征组合，选取实验效果好的分类算法来进一步分析。

第五章 基于社会网络方法的情感分析

阐述了目前的几种情绪传播模型、热点事件模型，选取了合适的传播模型，通过抓取数据，对数据进行预处理、指标计算、情感计算及结果数据统计，从客体事件、主体用户及情感特征分析这三个角度对情绪传播进行分析与阐述。

第六章 总结与展望

对本文所做的工作进行总结，基于算法和实验中的问题提出了对未来工作的展望。

# 相关理论与技术

本章主要阐述了情感分类相关技术，主要包括文本特征选择、情感分类算法及评价指标等。

## 文本特征选择

文本的表示方式因特征项不同而存在很大差异，进而影响情感分类的性能。特征

选择方法的研究具有很大实用意义，原因主要有两个方面。一方面，并非所有得到的特征项都会对分类有贡献，部分特征项是噪声反而会降低分类性能。还有一些特征在分类过程中并不存在信息量，这些特征项并不会提高分类效果。另一方面，情感分类所用的特征是由词语特征构成，特征空间的维数通常过高从而影响分类效果。因此，特征选择就是“去其糟粕，取其精华”，去除噪声特征，提取区分度较高的，对情感分类比较有利的特征。

目前，文本分类所使用的特征选择方法基本上是一些过滤方法，即使用特定算法给特征项打分，然后通过设置阈值过滤掉那些分值过大或过小的特征。在特征选择过程中，要使用一个权重函数对每个特征项进行计算，得到标识该特征重要性的权重值。利用权重大小对特征进行排序，筛选出排在前面的特征作为最终的分类特征。对于某个特征的权重函数，计算出的结果越大，该特征则被认为越重要。目前常用的特征选择方法有：文档频率、信息增益、卡方统计等。

#### 文档频率（DF）

文档频率指的是训练集中出现某特征词的文档数量，其以概率统计理论为基础，通过文本特征项的出现次数来作为特征重要性的判定依据。该方法仅选择大于指定阙值的特征项集合，该方法方法计算复杂度小，实现简单，并且在基于主题的文本分类任务中表现出较好的性能。该方法认为某特征在文档中出现次数越多就显得越重要，但实际操作中，某些高频的停用词计算出的值很高但并不具有区别类别的能力，选取这些特征对分类并没有太大作用。

#### 信息增益（IG）

该方法信息论为基础，通过计算信息增益值来完成文本特征的提取。如某个特征相应的信息增益值较高，就说明该特征比较重要。信息增益选择方法与文档频率选择方法类似，都是通过设定信息增益阙值并筛选出大于阙值的特征作为最终特征。信息增益方法的缺限是额外考虑了特征未发生的情况。当文本类别与特征值的概率分布极端失衡时，该方法的特征提取结果会有很大不同。

#### CHI统计选择方法

该方法是在特征t与类别C之间假设具χ2有一阶自由度的分布的基础之上提出的一种特征选取方法，该方法侧重于衡量类别与特征之间的关联程度。其值的大小与特征和类别之间的关联程度成正相关，该值表示对应特征项包含的类别信息量的多少。该方法认为特征的重要性应当由特征的值给出。如果某个特征项的值小于设定的阈值，那么认为该特征项含有较少的类别区分信息。

这些特征选择方法对文本分类有较高的实用价值，可以提高文本情感分类任务的性能与效率。但是并不是所有的特征提取方法都适用于论坛文本特征的提取。这些特征选择方法更多地考虑了文本的概率统计特征，而对自然语言本身语义特性考虑比较少。

## 情感分类算法及评价指标

由于文本的情感分类问题实质上是一个分类问题，因此一般的分类方法均可用于文本情感分类的研究。目前的情感分类有主要存在两种不同的理论方法，这两种方法是基于词典和规则相结合的方法和基于统计的方法。

### 基于词典和规则相结合的方法

基于词典和规则的方法主要是利用情感词典来对文本进行情感分类的一种方法。最简单的基于词典和规则的方法就是通过目前的情感词典来对测试文本中的正负情感词数目进行统计，通过比较正向情感词和负向情感词个数来确定文本的情感倾向。如果测试文本中正向词数目大于负向词数目，则判定文本的情感倾向为正向；如果小于则判定为负向；如果个数相等则判定为无情感倾向。目前广泛使用的情感词典为英文的WordNet词典和中文的HowNet词典。Kamps等[16]人通过计算情感新词同基准词之间的语义距离来判断文本的情感倾向；朱嫣岚等人[17]提出了基于语义相似度和语义相关场的方法，利用HowNet词典来判断情感新词情感极性；李纯等人[18]利用词语在HowNet词典中的定义和描述通过计算新词和种子词之间的语义相似度来判断新词的情感极性。基于词典和规则的方法适用于各种语料库。

### 基于统计的方法

基于统计的方法实质上就是基于机器学习的方法，即利用训练好的分类器测试文本进行情感分析。在基于机器学习的方法中，首先要筛选一出些文本，并对这文本的情感极性进行标注，然后把这些文本作为训练集，通过抽取一些特征并训练处这些特征与类别的关系，即训练出一个情感分类器，最后就可以运用该分类器对文本的情感进行分析。

近年来，许多研究者都致力于设计分类器的算法，比如朴素贝叶斯，KNN和支持向量机等算法，目前这些分类算法都广泛应用到了各个领域当中。

#### K-近邻分类方法

KNN(K-nearest Neighbor)即k最近邻法。该方法的思路是：如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则认为该样本属于这个类别。

为了对测试文本进行分类，就需要从训练集中计算离该测试集最近的k个邻近文档，通过k个邻近文档的类别来判定该测试文档的类别。通过计算找到k个邻近文档后，利用公式（2-1）对k个文档中同类别的文档与测试文档的相似度进行求和，求和结果作为某类别和该测试文档的相似度值。通过对候选分类评分进行总排序后，就可以得到一个阈值。

 （2-1）

公式（2-1）中，y(di，Ci)取值为0或1，当其取值为1时表示文档di属于类别Cj，当其取值为0时表示文档di不属于类别Cj；sim(x，di)表示测试文档x和训练文档di之间的相似度；bj是二元决策阈值，一般采用向量夹角的余弦值表示两个向量之间的相似度。

研究表明，当k>l时，k近邻节点算法的效果很好。由于不平衡的样本对于该算法的影响比较大，一般都会采用权值的方法来对其进行平衡，除此之外，由于k最近邻节点算法的计算效率比较低，为了提高效率需要对文档的样本进行剪辑。

#### 朴素贝叶斯

贝叶斯分类算法是统计学中一种比较简单的[分类方法](http://baike.baidu.com/view/8348989.htm)，该方法是建立在贝叶斯定理基础之上的分类算法。在很多分类过程中，[朴素贝叶斯](http://baike.baidu.com/view/992724.htm)分类算法可以与很多优秀的分类算法相媲美，比如[决策树](http://baike.baidu.com/view/589872.htm)和神经网络分类算法。由于简单、速度快的特点，贝叶斯算法也经常被用到海量数据的处理。由于[贝叶斯定理](http://baike.baidu.com/view/1488573.htm)成立的前提是各个[属性值](http://baike.baidu.com/view/121567.htm)之间是独立的，因此这些属性值对于类别的影响程度也是独立的。但是这种假设在实际分类过程中是不成立的，所以对于某个分类场合中属性之间联系比较少的分类准确率就比较高，属性关联系比较强的分类效果就不会比较差。为了克服这种情况，部分研究者就设计出了许多降低独立性假设的贝叶斯分类算法，比如TAN(tree augmented Bayes network)算法等。

设数据样本中的n个属性值用n维特征向量来进行描述，即：X={x1，x2，…，xn}，假定要对这些样本分到m个类中，m个类用C1，C2，…，Cm表示。对于一个待分类的数据样本X，运用[朴素贝叶斯](http://baike.baidu.com/view/992724.htm)分类法将待分类的样本X分配来进行分类，若该样本分到类Ci中，则该样本满足公式（2-2）。

P(Ci|X)>P(Cj|X) 1≤j≤m，j≠i （2-2）

由贝叶斯定理得知，在各个类别中P(X)为常数，因此[后验概率](http://baike.baidu.com/view/336754.htm)P(Ci|X)可转化为先验概率P(X|Ci)P(Ci)。一般的文本分类的训练[数据集](http://baike.baidu.com/view/262162.htm)都有许多属性或者元组，P(X|Ci)的计算效率会比较低，而贝叶斯定理一般认为各属性的取值都相互独立，因此先验概率P(x1|Ci)，P(x2|Ci)，…，P(xn|Ci)可以通过训练数据集来计算。

依据贝叶斯分类算法，对一个待分类的样本X，首先计算出样本X在类别Ci的[概率](http://baike.baidu.com/view/45320.htm)P(X|Ci)P(Ci)，然后选取先验概率值最大的类别为该样本所属类别。

因为[朴素贝叶斯](http://baike.baidu.com/view/992724.htm)分类算法成立的条件是各个属性之间相互独立，所以数据的各个属性关联性较少时，分类效果会比较好，否则效果就会很差。

#### 支持向量机模型

支持向量机模型(Support Vector Machine)是Cortes和Vapnik在1995年首先提出的，该模型对小样本、非线性及高维模式识别有许多特有的优势，并能够很容易推广应用其他机器学习问题中。支持向量机方法是建立在统计学的VC维理论和结构风险最小原理基础上的，通过有限的样本信息来寻求模型的复杂程度和学习的最佳折衷，来获得最好的泛化能力。支持向量机的优势有如下几个方面：

##### 支持向量机相对于其他分类器所需要需要的样本集较少，该样本集较少不是绝对数量较少，而是对于问题域该方法所需要的样本数较少。

##### 支持向量机模型的训练过程主要是一个二次规划问题，可以采用拉格朗日乘子法来解决。从理论上来说，解决二次规划就可以得到全局最优解，从而可以有效地避免局部极值问题。

##### 支持向量机的最终决策函数只由少数的支持向量所确定,因此计算的复杂程度是由支持向量的数目决定,而不是整个样本空间的维数,从而避免了高维灾难。

在二维空间，仅有两类样本的分类如图2-1所示

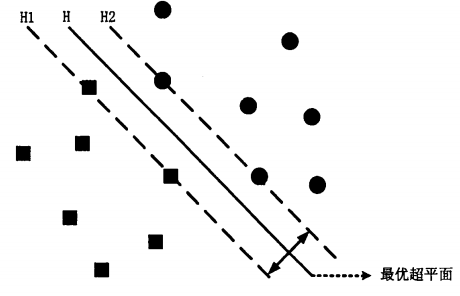


图2-1 两类样本集的分类实例图

图2-1中用圆点和方块分别表示两种不同的样本，实线H表示的是分类的最

优超平面，平行于H平面并在其两边H1和H2表示两类支持向量，离分类超平面H最近的平面H1和H2之间的距离就是分类间隔。支持向量机分类的目标就是使分类间隔达到最大值，因此分类问题也就可以转换成线性可分的样本集(Xi，Yi)，i=1，．．．，n，x∈Rd y∈{+1，-1)，其满足条件如公式（2-3）所示：

 （2-3）

支持向量机方法是基于线性可分情况下的最优分类面而提出的，对于线性不可分的情况，可以把其转化成凸优化问题，即得到的目标函数是二次的，约束条件是线性的。因此这个问题就转化为标准的QP(Quadratic Programming)问题，由于它的特殊结构，可以先通过 Lagrange Duality 变换到对偶变量的优化问题，这样便可以找到一种更加有效的方法来进行求解。把拉格朗日函数将约束条件融合到目标函数中,结果如公式（2-4）所示：

 （2-4）

为了减小最小化分类的错误，就要增大最大化分类的间隔，那么目标函数如

公式（2-5）所示：

 （2-5）

在． 和 约束条件下对αi求解式如公式（2-6）所示：

 （2-6）

其中αi是每个样本相应的Lagrange的乘子，引入核函数K(xi，x)后，得到最优判别分类函数如公式（2-7）所示：

 （2-7）

其中b\*是任意一个支持向量求得的分类阈值。

### 评价指标

实验结果评估是实验中不可分割的组成部分，为了保证实验评价的客观性，就需要考虑各种影响因素之间的平衡。由于实际中往往存在诸多不确定的影响因素，因此实验的结果评估通常都不会精确。目前自然语言处理结果评估往往会参照人工的标注结果，因此也通常认为人工标注的结果为正确的。 情感分析作为自然语言处理的一部分，其结果评估也依赖于人工标注，通过人工标注与实验结果来对比来对实验结果进行评价。

情感分析的结果评价指标与其他自然语言处理的结果评价指标相同，这些评价指标有准确率(Precision)，召回率(Recall)，F一测度值(F-measure)。

在情感分类中，准确率用来衡量其分类结果的准确度。准确率可以用分类结果中应该属于该类且被判定为该类的数目与分类结果中该类总数的比值来表示，如公式（2-8）所示

 （2-8）

在情感分类中，召回率用来衡量该分类模型的完备性。准确率可以用分类结果中应该属于该类且被判定为该类的数目与本应属于该类别总数目的比值来表示，其公式如式（2-9）所示

 （2-9）

若情感分类的实验结果如表2-1所示：

表2-1 文本分类器输出结果

|  |  |  |
| --- | --- | --- |
| 分类器对二者关系的判定 | 测试文本属于实际类别 测试文本不属于实际类别 | |
| 标记为yes | a | b |
| 标记为no | c | d |

表2-1中，a表示属于该类且被判定为该类的个数；b表示不属于该类且被判定为该类个数；c表示属于该类却没被判定为该类的个数；d表示不属于该类且没有被判定为该类的个数。

依据表2-1将准确率p、召回率r的分别用公式表示，结果如公式（2-10）、公式（2-11）所示：

 （2-10）

 （2-11）

准确率和召回率分别从不同的方面反映了分类的效果，许都文献中的实验表明两者存在着潜在的联系。在实际实验过程中，准确率低时，召回率就会高；反之亦然，准确率和召回率通常不能同时较高，这也就说明，准确率和召回率之间是相互制约的，为了从准确率和召回率之间寻求一种平衡以使得准确率和召回率都比较高。针对此定义的评价指标公式如式（2-12）所示：

 （2-12）

公式（2-12）中，Fβ表示准确率和召回率效果的综合度量指标，p表示准确率，r表示召回率，β表示准确率p和召回率r在综合评价中的影响程度。通常认为准确率和召回率的重要程度相同，所以将β的值选定为1，这时的综合评价指标变为（2-13）。

 （2-13）

## 情感网络特征

在社会网络中嵌入情感信息就构成了情感网络，就可以对情感网络进行特征分析。社会网络的情感分布特征可很多方面来进行分析，比如两个或三个节点的情感关系分析、整体网络的情感分布特征分析等。目前两个或三个节点的情感关系分析的理论主要包括互惠性理论、结构平衡理论和状态理论。

### 互惠性

互惠性一般被用在有向网络中，用来衡量网络中两个结点形成相互连接的程度。网络互惠性的研究有很好的指导意义，一方面，互惠性可以体现出网络中个体之间交互的密切程度；另一方面，在实际操作中为了简便起见，我们经常忽略有向边的方向，而互惠性可以揭示出忽略有向边的方向所可能产生的误差。

1983年Kiesler [19]提出了人际互补性的理论框架，该框架认为每一个行动都能激发一个反应，每一个反应都会适应之前的行动，因此就导致了人际交往中关系类型的重复，比如对他人友善就会换来他人对自己的友善。该理论说明了情感关系的双向性，一个节点对另一节点的情感能够激发另一节点对此产生相应的反应和相应的情感态度。在线社交情感网络作为一个有向网络，那么互惠性是否也是情感关系网的一个特征，所以可以对情感网络进行互惠性分析，判断两个相邻节点彼此情感是否是一致的，如图2-2所示，节点A对节点B的态度是积极的，而节点B对节点A的态度却是消极的，那么A和B的情感就是不一致的。本文通过用户之间的交流消息确定有向边的情感，通过两用户之间双向边的情感差值来反应互惠的程度，如果用户之间的情感差值较小，则认为两个用户的情感一致，如果两个用户的差值较大，则认为这两个用户的情感不一致。



图2-2 用户情感不一致情况

### 基于结构平衡理论和状态理论的情感关系研究

互惠性理论主要研究的是两点之间的关系，结构平衡和状态理论主要研究的是情感网络中三个节点之间的关系，这些研究理论都可扩展到整体网络的研究之中。状态理论通常研究的是有向网络，而结构平衡理论可应用到无向网络和有向网络中。

情感平衡研究是建立在结构平衡理论基础上的。结构平衡理论是1956年由Cartwright和 Harary [20]在 Heider [21]提出的平衡理论的基础上形成的通用的结构模型。结构平衡理论主要研究的是网络中三个节点之间的关系，该理论通俗地说就是“朋友的朋友也是朋友，朋友的敌人也是敌人”。如图2-3是一个有向的三方网络关系图，其中A与B之间相互积极认可，A对C持有消极情感，则推测B对C的情感应为消极的。因此情感平衡研究可运用已知的情感关系来推测出未知情感关系。本文通过统计两个用户节点之间的情感关系以及对第三个节点的情感态度数据来对情感网络的平衡性进行分析。

状态理论指出：在社交网络中，如果A对B的情感是积极的，那么说明B在该社交网络中的地位会比A的网络地位高，反之则说明B在社交网络中的地位比A网络地位低。状态理论可用于社交网络中用户地位的评价，也可以来预测和判断网络中的关系。本文首先对用户的情感倾向及用户影响力进行计算，然后对符合状态理论数据进行统计，并对结果进行分析。



图2-3 基于平衡理论的三节点关系

### 情感分布同质性

同质性理论指的是人们通常会与自己相似的人建立联系，Adamic等[22]对斯坦福大学社交网站Club Nexus中的社交网络进行了分析，发现在部门、年级、专业、性别以及爱好等属性方面，随着用户之间连接距离的增加相同属性却变得越来越少，说明用户更倾向于与自己具有相同属性的其他用户取得联系。

情感网络同质性主要研究网络中具有相同情感的用户是否更容易彼此交流，聚集到一起。网络情感同质性的研究主要采用的方法为聚类分析和相关分析，分析相邻节点的情感是否彼此相关或一致。本文在用户节点情感值确定的基础上，通过对相邻用户节点的情感值的一致性统计来判断是否情感网络也满足同质性，即正向的用户是否更愿意与正向用户交流，负向用户是否更趋向于与负向用户进行沟通。

## 社会网络指标

社会网络有自己的一套指标体系，体系中各个指标之间相互联系，根据这些指标可以发现社会节点中的潜在的一些联系，可以挖掘出意见领袖，也可以有助于我们发现社区。

##### 网络规模和直径

网络规模是指网络的大小，网络直径是任意两个节点间距离的最大值。。

##### 网络平均长度

网络平均长度是网络中所有节点对之间的平均最短距离。这里节点间的距离指的是从一个节点要经历的边的最小数目。

##### 密度

密度是一个图各点之间联络的紧密程度。固定规模点之间的连线越多，则说明该图的密度越大。具体来说，是指在网络中实际存在的线与可能存在的所有的线的比值。用来测量社会网络中行动者之间的联系程度，值介于0与1之间。

##### 中心度

中心度描述图中任意一点在网络中占据的核心性。中心度主要可以分为三个方面来衡量：点度中心度、中间中心度和接近中心度。点度中心度又包括绝对中心度和相对中心度，绝对中心度就是指该点的度数，相对中心是绝对中心度与图中点的最大可能的度数只比。中间中心度是来测量社会行动者对资源控制的程度。接近中心度表示一种不受他人控制的程度。

##### 中心势

中心势是描述整个图的紧密程度或一致性，也就是一个图的中心度。每一种中心性都有对应的中心度和中心势两个指标。

##### 聚类系数

聚集系数是用来表示网络图形中节点聚集程度，有理论研究表明，在现实中的构建的社会网络中，节点之间更趋向于建立起严密的组织关系。聚类系数表示的是与节点i相连的k个点，k个点之间的实际存在边与总的k个节点的所有可能的边的比值。公式如(2-14)所示

 (2-14)

## 本章小结

本章主要首先介绍了情感分类相关技术，包括文本特征选择方法、常用的情感分类的算法、评估指标等，然后介绍了情感网络的一些特征及社会网络分析的基本指标。

# 情感词典的构建

文本情感计算通常都建立在情感词基础之上，因此情感词的全面程度会很大程度的影响情感分类的效果。由于目前网络上各种评论文本都比较短小，因此细化的情感词典对这些文本情感计算有非常重要的意义。

情感分析过程中使用情感词典来对特征进行选择，即直接根据情感词典对文本中的情感词进行抽取并将其作为特征项。对情感词典中的情感词进行优先级排序，并把该词典加入到分词系统中，这样分词不仅比较简单而且准确。除此之外，分词和文本特征选择可以同时进行，这样处理速度速度有很大提升，进而大大降低对算法的时间复杂度，直接提取的情感词也可以大大提高情感分析的准确性。

## 传统情感词典构建与不足

### 传统情感词典构建

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)，OpinionFinder(OF)，AppraisalLexion(AL)，SentiWordNet以及Q-WordNet。在这些词典中，GI不仅每个词的义项都列了出来，而且每个词情感特性都有标注。但是在国内，情感计算研究才刚刚开始，词典资源也比较少，还没有一部完整且权威性的情感词典。目前常用的情感词典资源包括董振东老师开发的《知网》(HowNet)，清华大学整理的《中文褒贬义词典》[23]，大连理工大学整理的《情感词汇本体》[24]，台湾大学整理并共享的中文情感词典(NTUSD)，张伟等人[25]编撰的《学生褒贬义词典》，杨玲，朱英贵[26]编撰并整理的《贬义词词典》，史继林、朱英贵[27]编撰《褒义词词典》，哈尔滨工业大学信息检索实验室整理的《同义词词林扩展版》[28]。这些词典主要是以手工或半自动方式编辑而成。除此之外，还有许多研究者提出其他的方法来构建情感评价词语，有研究者利用词典将手工采集的种子评价词语进行扩展来获取大量的评价词语；有研究者将词典资源丰富的语种的英语情感词典翻译成中文，供中文情感分析应用。对于情感词典极性的判别，研究者也做了很多工作，部分研究者通过互信息来判断情感词的极性，还有一部分研究者采用基于图的方法来识别评价词语的极性。

### 传统情感词典的不足

从目前情感词典的发展来看，主要集中在字或者词的词典构建与扩充，而且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。由于很多词存在一词多义现象，目前构建的情感词典也通常包含很多存在歧义的词，如“老”在很多场合下表现为“旧，年龄大”的意思，但在部分场合可以扮演修饰成分(如老高、老远)。而且许多不在情感词典内的评价单元，比如“手机耗电高”，明显是对手机的贬义评价，这样的特征在人们的评价中很多，许多也包含明显的情感信息，但是现有的情感词典并不包含。除此之外，人们在自己心中都有自己的行为规则，在描述或者看到一些不文明行为时候，其实已经表达了自己的情感信息，比如“涨工资”，“学生打老师”等，这些包含情感特征的行为要素也不存在现有的情感词典当中，所以说现有的情感词典还有许多不足的地方。

所以为了更多的挖掘文本中的情感特征，针对目前情感词典的不足的问题，本文利用情感词典资源构建了基本情感词典，在此基础上运用自然语言处理技术构建了基于实体属性及行为的情感词典。

## 基本情感词典的构建

本文中的基本情感词典包括三个方面，褒贬情感词典、否定情感词典及程度修饰词典。

#### 褒贬情感词典

##### 基础褒贬情感词典

在中文词语中，很多词存在很强的情感倾向，而且情感倾向也不随语义环境变化，比如美丽、美好、幸福等褒义词语各种文本中都呈现很强的褒义情感倾向；厌恶、残酷、暴力等是在各种文本中呈现很强的负面情感倾向。这类情感词在不同场合的文本情感倾向性都比较稳定，在本文中我们把这些具备跨领域能力的情感词叫做基础情感词。本文的基础情感词典是以《知网》、台湾国立大学的通用情感词典、大连理工大学的《情感词汇本体》等几个比较权威的开源情感词典为基础，通过人工筛选，去掉一些有情感歧义的词语，得到6952个情感词，然后对这些情感词的权值进行人工标注，其中正向情感词3667个，负向情感词为3285个。本文对于词语的情感褒贬程度用区间（-1，+1）来对进行表示，其中标注值大于0为正向情感义词，小于0的为负向情感词，标注值的绝对值越大，情感强度越强。其情感词统计实例如表3-1所示，详细列表见附录。

表3-1 基本褒贬情感词统计表

|  |  |
| --- | --- |
| 情感词类型 | 举例 |
| 正向情感词 | 毕恭毕敬0.6、别具匠心0.8、漂亮0.8、美好0.7、超俗 0.8等 |
| 负向情感词 | 霸道-0.7、傲慢-0.6、卑鄙-0.8、暴戾-0.7、卑猥-0.8、鄙俗-0.7等 |

##### 网络褒贬情感词典

网名在论坛中互动交流时，评论文本都比较偏口语化，因此各种俗话和日常用语都比较常见。除此之外，部分网民素质比较低，经常评论一些脏话或者粗话。通过传统方式构建的情感词并不会包含这些口语化或新型褒贬词语，因此只使用传统情感词典对论坛文本进行情感计算的有效性不会很高，为了提高论坛文本的情感计算的准确性，就需要构建包含网络用语的情感词典。

为了更有效的获取网络情感词，本文人工地从网上搜集网络用语词汇，并对这些词进行筛选，保留存在情感倾向的词汇140个，并对其情感强度进行标注，这些情感词实例如表3-2所示，详细列表见附录。

表3-2 网络褒贬词统计表

|  |  |
| --- | --- |
| 情感词类型 | 举例 |
| 正向情感词 | 大神0.7、NB 0.8、萌萌哒0.6、给力 0.7等 |
| 负向情感词 | 废柴-0.7、怪咖-0.6、煞笔-0.8、碰瓷-0.7等 |

#### 程度副词词典

在句子中，副词主要作用是修饰动词和形容词，其通常会充当补语和状语。副词主要包括程度副词和否定副词。其中程度副词在修饰情感词时，会一定程度上加强或减弱情感词的情感倾向。由于程度副词在带有情感倾向性的论坛文本中出现的频率也是比较高的，因此程度副词也是影响情感倾向性的重要因素。

本文中的程度修饰词典是对《知网》中的程度修饰词进行了权值标注而得到的，标注的区间为（0,2），这些程度副词有178个，这些程度副词实例如表3-3所示，详细列表见附录。

表3-3 程度副词统计表

|  |  |
| --- | --- |
| 程度副词类型 | 举例 |
| 较强修饰词 | 极度1.8、万分 1.8、大为1.5、格外 1.5等 |
| 较弱修饰词 | 轻度0.7、不怎么0.8、微 0.6等 |

#### 否定词词典

否定副词是对文本情感倾向性会有着否定或者反转的作用，如果只待测文本中情感词对于情感倾向的影响，而不考虑否定词副词对于情感词的修饰影响，那么待测情感倾向就很容易判断错误。因此，在情感词典的构建中，需要构建否定情感词典。

本文制定策略对训练语料中的文本情感进行计算，然后从实验结果中筛选了与标注结果不一致的语料，这些语料情感计算异常往往是因为未考虑否定词造成的，由于否定副词作为修饰时往往会在情感词的前面，所以本文抽取文本中的副词，并对提取出来的词踢去程度副词后进行人工筛选得到否定词及组合71个，例如“不”，“难道不”，“岂不”等。

这些否定副词以及组合对于情感倾向有比较大的影响，往往这些词或组合中没有否定含义的词造成单句情感倾向性的置反，而带有否定含义的词的又使得单句情感倾向性保持不变。所以建立否定词以及组合词典很有必要。

## 基于实体情感词典的构建

在以往的研究工作中，研究者们通常会将评价词语的识别作为情感计算的一项基本任务。首先利用各种方法识别并抽取出评价词语，然后运用评价词语来对本文情感进行分析。目前的情感分类方法主要包括基于情感句打分的方法和基于机器学习的分类方法。第一种方法主要通过对评价词的极性的累加结果来确定文本最终的情感倾向性；第二种方法是把评价词作为分类特征，运用机器学习的分类方法来完成文本的情感分类。但是在文本情感分析中只考虑情感词还是远远不够的，大部分的研究工作忽视了以下两个问题：

#### 情感句的情感倾向性并不仅仅由情感句中的评价词语决定，而是由评价词语

及其所修饰的评价对象共同决定的。评价词语中有许多属于上下文相关的评价词，即这个词在不同的上下文中却表现出不同的情感倾向性。比如情感词“高”在以下两个句子中的含义：

例句1：这颗树长真高。

例句2：这个手机能耗比较高。

例句1和例句2是情感句，但是情感词“高”修饰不同的评价对象时有着不同的极性。比如“高”在例句1中是褒义，而在例句2中则是贬义。可以看出，这些词语是没有办法准确划分到情感词典中去。

#### 还有许多没有评价词的实体行为中也存在明显的情感倾向，实体行为情感主要

有实施者、行为及被实施者中的两个或三个决定的。比如“反对独裁”、“涨工资”、“夹带私货”，“学生不尊重老师”等，这些短语都存在很明显的情感倾向。

从上面问题可以看出：文本中的情感倾向，主要是由文本中的修饰的评价对象及对象行为所决定的。所以本文中构建了基于实体的情感词典。实体情感词典构建包括实体属性情感词典构建和实体行为情感词典构建。

### 基于实体及属性的情感词典的构建

实体属性情感词典主要包括评价组合抽取和极性判别。

#### 评价组合的抽取

评价组合的抽取方式有两种，一种是基于手工模板和规则的方法，另一种是基于句法路径的情感评价单元自动识别方法,。第一种方法由于模板或规则是手工构建的，召回率不高；第二种方法为主要是自动构建方法，构建的句法规则结果为更加完整，可以匹配到更多的句法关系，实验效果好,赵妍妍,秦兵,车万翔等[29]就是采用该方法对情感评价单元进行识别。因此本文主要讨论第二种方法，首先对有效的句法路径进行获取并构建出句法路径库，然后采用基于句法路径的匹配算法来对情感句中的情感评价单元进行识别。

##### 构建句法路径库

句法路径是指在句法树上链接任意两个节点之间的句法结构。在本文中，句法路径特指在短语结构句法树。在句法路径的生成步骤中，首先需要确定情感句中的评价词语和评价对象。其中，使用评价词语来自知网HowNet的评价词语，共包含6952个。因为评价对象通常是为名词或代词，所以本文把词性为名词或代词作为候选评价对象。

基于此，如果一个情感句中含有m个候选评价对象和n个评价词语，这些评价对象与评价词语会构成m\*n个候选句法路径。为了筛选出有价值的句法路径，本文对结果进行了两步泛化：如果句法路径含有连续相同的句法标签，则用一个标签取代它；如果某些句法成分标签表达了相似的含义，用规范化标签将它们代替。

在为所有的候选句法路径完成了两步泛化之后，本文统计出各种句法路径的出现频率，并通过阈值th的限定来选取较为频繁的候选句法路径构成句法路径库。这种通过考察频率来确定句法路径正确性的方法基于下面的一个假设：评价词语与其真正具有搭配关系的评价对象之间满足一定的句法关系，并且这些句法关系是有规律可循的、可总结的，而并非杂乱无章的。

##### 基于句法路径匹配的情感评价单元识别

一个情感评价单元包括两个部分：评价词语和评价对象。本文首先通过评价词典获取情感句中的评价词语。为了找到与这些评价词语相搭配的评价对象，本人首先对情感句进行句法分析，然后获取该情感句的句法树，并统计评价词语和候选评价对象所产生的候选句法路径，并对这些候选句法路径进行两步泛化；然后将处理过的的候选句法路径与句法路径库中存在的标准句法路径进行比较，匹配优先级高的候选句法路径，并将链接到的评价对象作为最终的评价对象。当然，如果这种匹配的方法无法成功地匹配句法路径库中的任一条句法路径，则认为该评价词语没有任何价值。

虽然这种句法路径精确匹配的方法可以准确识别出情感句中的情感评价单元，但是仍然存在不足，导致召回率不高：为了改进句法路径精确匹配的不足，本文提出一种基于编辑距离的句法路径匹配算法。如果path与path’之间的编辑距离小于阈值，则本文近似地认为path所链接的评价词语和评价对象组成了最终的情感评价单元。

##### 实体与情感评价单元关联

从抽取出来的评价单元可以发现，有一些是直接修饰实体的，可以直接识别出来，比如“电池不错”；还有一些表面上跟实体没有什么关系的评价组合，也是对于实体的有效评价，比如“待机时间长”。所以本文要对实体对象与相关评价进行关联。既然实体与评价单元为被修饰及修饰关系，那么它们的共现次数比较多，所以本文采用点互信息法来对其进行关联,如果共现率高于一定阈值，则对其进行保留，公式如下（3-1）所示：

 （3-1）

其中，entity表示实体，assessment表示评价单元，P(entity& assessment)表示两个词语entity与assessment共同出现的概率，P(entity)与 P(assessment)分别表示两个词语单独出现的概率。

#### 评价组合极性判别及情感值量化

如何判断一个词语是褒义词或贬义词，目前的研究有两种思路：一种是基于语义计算，例如：朱嫣岚等人采用HowNet语义相似度的方法，计算目标词语跟基准词之间的紧密程度，得以判定情感极性。另一种方法是统计分析法，Turney等人[30]采用统计的方法，首先计算待判别词与基础情感词之间的点互信息值，然后通过点互信息值确定两个词之间的联系程度，把联系程度最高的词的情感倾向作为待判别词的情感倾向。王素格等人[31] 提出了基于同义词的情感倾向判别方法，该方法首先通过词频概率估计找到区别类别能力较强的词语，然后将这些词语与构建的情感词表相结合来获取情感词。

本文在论坛语料集的基础上，运用互信息（PMI）的方法计算评价组合与基准词的共现特性，从而判断候选词的情感倾向，然后利用这些出现共现语料的情感值对该评价词进行情感值量化。

##### 情感词的极性判别

本文将 PMI（点互信息）方法引入来判断词语的情感倾向（Semantic Orientation，简称 SO）中。基于PMI算法的基本思路是：本文首先从知网的HowNet情感词典倾向性非常明显的词语，并把这些词语分为正向情感词和负向情感词，假设分别用 PosWords 与 NegWords 来表示这正向情感词语和负向情感词语，然后计算待判定词语word和PosWords的点间互信息与word和 NegWords的点间互信息的差值，最后就可以根据该值来判断词语 word的情感倾向，公式如式（3-2）所示：

 （3-2）

将点互信息差值与0来进行比较，不同的差值对应不同类型的词，如下所示：

>0; 褒义词

SO-PMI(word) =0; 中性词

<0; 贬义词

运用SO-PMI 公式，就可以判别候选词的情感倾向，对于这些存在情感倾向的候选词语，本文根据其点互信息差值分别将这些词语归入情感词典中的正向情感词典或负向情感词典。

##### 情感评价单元情感值量化

人们在对事物进行评价时，其情感往往会持续一段时间，比如对于美好事物的正向评价具有持续性，即评价过程中会持续使用正向词汇；对较差的事物的负向评价也具有持续性，即评价过程中会持续使用负向词汇。由于情感单元是为了表示文本的情感值，所以情感单元与文档情感值成正相关，所以本文使用文本的情感值来表示情感评价单元的情感值。

对于出现某个情感评价单元的所有文本，本文针对文本设计了一些规则进行情感计算，首先使用句法分析技术，抽取依存关系对组，继而逐词判断是否为情感倾向性词，如果出现倾向词，则将之抽取出并添加入情感词列表；如果有否定词、程度副词出现，则根据情感扩展词典进行相应处理；通过计算包含情感词修饰对的极性强度，求和得出整个句子的情感强度，最后对情感倾向的句子求和得出平均值，得出文本的情感值，大体步骤如下图3-1所示。

图3-1 文本情感计算流程

文本情

感计算

分词

句法分析

评论

文本

抽取情感关系对

关系对组情感计算

文本情感计算具体步骤如下：

###### 利用句法分析技术对文本进行句法分析。

情感极性词多出现在ADV(状中结构)、CMP(动补结构)、DE(“的”字结构)和VOB(动宾结构)这几种结构中，依存句法分析则可以表示这种语义关系，例如“他是个很骄傲的人，简直是不可一世。”句法分析结果如下所示：

“他\_0 是\_1 SBV 是\_1 -1 HED 一个\_2 人\_6 ATT 很\_3 骄傲\_4 ADV 骄傲\_4 人\_6 ATT 的\_5 骄傲\_4 RAD 人\_6 是\_1 VOB， \_7 是\_1 WP 简直\_8 不可一世\_9 ADV 不可一世\_9 是\_1 COO。 \_10 是\_1 WP”

其中数字代表前面对应词在整个句子的位置，比如“他\_0”,说明他为该句第一个字符，字母符号代表代表两个词之间构成的句法关系，比如“很\_3 骄傲\_4 ADV”，说明“很”和“骄傲”构成的是状中结构。

###### 从句法分析结果中抽取情感对组。

本文定义了一个实体关系类RelationPair(firstWord，firstPos，secondWord，secondPos，relation)，其中firstWord为第一个词，firstPos为该词的位置，secondWord为第二个词，seccondPos为第二个词的位置，relation表征两个词之间的修饰关系，把句法结果进行结构化存储，然后遍历存储的结果集，抽取包含极性词、程度副词及否定词的情感对组。比如“骄傲”该词具有情感极性，所以“很\_3 骄傲\_4 ADV”就被抽出为情感对组。

###### 根据情感词典，依据公式对情感对组进行计算。

遍历情感对组，找到其中的情感词，以及修饰该情感词的程度词及否定词，其中否定词会考虑距离的影响程度减弱情况，通过该句中单个情感词的极性强度。单个复合情感词（可由多个情感对组构成）计算公式如（3-3）所示：

 （3-3）

其中sentimentWordVal为情感极性词，intensei为第i个程度修饰副词，θ为距离系数，本文取2，modifiedIntensej为第j个否定修饰词，m为程度副词修饰的个数，n为否定修饰词的个数。

###### 考虑全局程度副词及否定修饰词并算出单个句子的情感值。

首先对单个句子中的复合情感词求和，最后考虑句子未参与计算的全局程度副词及否定修饰词，计算公式如（3-4）所示：

 （3-4）

其中itensei为第i个全局程度修饰副词，modifiedIntensej为第j个全局否定修饰词，n代表全局程度修饰副词的个数，d代表全局否定词的个数，m代表复合情感词的个数，wordSentimentScorek为第k个复合情感词。

###### 对有情感倾向的句子求和，求平均值得出文本的情感值。

对所有文本情感值求平均就得到该情感评价单元的情感值。

#### 实验及分析

本文把天涯论坛2014年共19124帖子及其评论作为实验数据，对该数据进行评价单元抽取，并对这些评价单元进行实体关联、极性判别及情感值量化，实验结果如下：

##### 评价单元抽取

通过构建句法路径库对评价单元进行抽取，筛选出出现次数较高的评价单元共3241个，结果如下表3-4所示：

表3-4 评价单元抽取结果

|  |  |  |
| --- | --- | --- |
| 情感词类型 | 举例 | 数量 |
| 正向 | 绿色：通道，有机：食品，特色：社会主义，新：技术等 | 1146 |
| 负向 | 破：房子，脑残：粉，豆腐渣：工程，邪恶:势力，薄弱：环节等 | 1627 |
| 噪音数据 | 美国：总统，西方：国家，相关：部门，美国：教育，俄罗斯：经济 | 468 |

##### 实体关联

运用互信息方法对抽取的评价单元与被修饰的实体进行关联，阈值取0.3，对其中存在关联结果的进行评估，结果如表3-5所示：

表3-5 实体关联结果统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | P | R | F |
| 值 | 0.54 | 0.81 | 0.64 |

从上表可以看出，由于统计文本数据量有限，噪音数据较多，准确率有点不理想。

##### 极性判别

通过SO-PMI算法，对评价单元的的极性进行判别，评估结果如表3-6所示：

表3-6 评价单元极性判别结果统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | P | R | F |
| 值 | 0.86 | 0.82 | 0.84 |

从上表3-6可以看出，通过褒贬词共现分析方法对情感词典的极性分类结果还是比较准确。

##### 情感值量化

本文把情感评价单元所在文本的情感值作为该情感评价情感值，计算结果在（-1,1）之间，与标注结果相差±0.2均为正确值，评估结果如表3-7所示：

表3-7 情感值量化结果统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | P | R | F |
| 值 | 0.68 | 0.73 | 0.70 |

从上表3-6可以看出，本文情感单元情感值量化方法的准确率效果比较可观。

### 基于实体行为情感词典的构建

中国有句俗话：没有规矩，不成方圆。其意思就是说，没有规则的约束，人类的行为就会陷入混乱。在我们的生活中有许多规则，道路有交通规则，学校有规章制度，就连游戏也有规则，各种各样的规则约束着我们，让我们在各种环境下能井然有序地生活。每种规则都是针对相应的角色，交通规则针对的在道路上行走的人及驾驶员，学校的规章制度是针对的是在校的学生及老师，游戏规则针对的是参与游戏的玩家。既然是规则，那么它就有约束力，它约束者人们或相应角色的行为，比如“驾驶员要遵守交通规则”，“学生应尊重老师”，“孩子要孝敬父母”等。规则在制约人们的同时，也渐渐地内化到人们的思想中，遵守规则成为了大家普遍认为该有的行为，而那些挑战规则的规则的行为就会引起大家的不满，所以当人们描述不遵守规则的行为时，内心的情绪已经表现出来，即文本中的实体的行为描述是存在情感倾向的。

本文构造了实体行为的情感词典，通过丰富情感词典来提高文本情感计算的准确度。该实体行为情感词典的构建主要包括行为的抽取及极性判别。

#### 实体行为抽取

行为是人类所特有的，由一定原因引起的，为了实现一定[目的](http://www.baike.com/sowiki/%E7%9B%AE%E7%9A%84?prd=content_doc_search)所进行的活动。行为是由满足某种需要的动机引起的，又是达到一定目标使需要得到满足的手段和过程。行为既是某种需要和动机的结果，又是这种需要和动机的反映。构成行为有一下基本要素：①行为的主体是人；②行为有一定的目的性、方向性和预见性；③行为作用于一定的对象(人或物)；④行为总要产生一定的结果。从行为的要素来看，行为主要构成有三个部分，即行为实施者、动作和被实施对象。由于文本中许多的行为描述比较简洁，可能会省略实施者或者被实施对象其中之一，所以本文的实体行为主要抽取文本中的实施者、行为动作及被实施者中的两个或者三个成分。

语义角色标注技术比较利于实体行为成分抽取，所以本文利用哈尔滨工业大学的自然语言处理工具LTP对语料进行语义角色标注，该自然工具对语言角色进行标签说明，其中核心的语义角色为 A0-5六种，A0通常表示动作的施事，A1通常表示动作的影响等，A2-5根据谓语动词不同会有不同的语义含义。比如对“大家可以过来喷一喷越南首都河内的高房价”进行语义角色标注，标注结果如下：“[大家]A0 可以[过来]v 喷 一 喷 越南 首都 河内 的 高 房价”，“ [大家]A0 可以 过来 [喷]v 一 喷 越南 首都 河内 的 高 房价”，“[大家]A0 可以 过来 喷 [一]ADV [喷]v [越南 首都 河内 的 高 房价]A1”。从标注结果可以看出，其中A0代表的实施者，v代表的行为动词，A1代表被实施者，所以通过语义角色标注技术很容易抽取到文本中的行为，只是不够简化，需要进一步处理。

由于被标注的实施者及被实施者会比较长，所以需要对其进行进一步处理，比如“越南 首都 河内 的 高 房价”，这里的被实施者为“房价”，但是标注结果中会有许多附加的修饰部分，所以最后的结果要去除修饰的成分。由于实施者或被实施者一般为名词，所以本文首先利用句法分析方法对该方法对抽取的语句进行分析，通过主谓结构及动宾结构找到实施者及被实施者，如下图3-2所示，“喷”和“房价”构成了动宾结构，所以通过抽取主谓结构及动宾很容易就可以找到实施者、行为及被实施者。

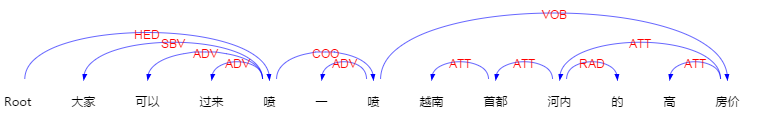


图3-2 语义角色标注分析结构图

实体行为评价单元抽取的具体步骤如下：

1. 首先从数据库中获取帖子及评论暂存到列表中。
2. 调用LTP的语义角色分析及句法分析方法，对列表中的文本进行分析。
3. 通过标签匹配对语义角色分析结果中的实施者、行为动词及被实施者进行抽取。
4. 利用行为动词在句法分析结果中匹配SBV(主谓标签)获取实施者表述。
5. 利用行为动词在句法分析结果中匹配VOB(动宾标签)获取被实施者表述。
6. 对抽取的结果进行保存并统计，保留阈值较高的行为评价单元。

#### 实体行为极性判别及情感值量化

实体行为的极性判别及量化方法与实体及属性修饰情感词的极性判别方法和量化方法均相同，此处不作过多描述。

#### 实验及分析

本文对2014年共19124帖子及其评论数据进行了实体行为抽取实验，对抽取的行为短语进行极性判别，并制定规则对行为短语进行情感值量化。

##### 实体行为抽取

通过语义角色标注技术对实体行为进行抽取，去除附加成分，保留实施者、行为或被实施者，选取出现次数较高前1500个抽取短语，如表3-8所示：

表3-8 实体行为抽取结果表

|  |  |  |
| --- | --- | --- |
| 情感词类型 | 举例 | 数量 |
| 正向 | 顶 楼主，涨 工资，有 道理，成 大器，解决 问题等 | 507 |
| 负向 | 夹带 私货，有 问题，遇到 瓶颈，害 死人，藐视 判决等 | 628 |
| 噪音数据 | 是 的，打 的，说 了，有 人，说 什么等 | 365 |

##### 极性判别

实体行为短语的极性判别与上述相同，也是用到SO-PMI算法进行计算，评价结果如表3-9所示：

表3-9 极性判别结果统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | P | R | F |
| 值 | 0.81 | 0.84 | 0.82 |

从上表3-9可以看出，通过褒贬词共现分析方法对情感词典的极性分类结果还是比较准确。

##### 强度量化

采用的强度量化方法与上面实体及属性评价方法相同，标准值区间在（-1,1），误差在±0.2之间，评价结果如表3-10所示：

表3-10 强度量化结果统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | P | R | F |
| 值 | 0.66 | 0.74 | 0.70 |

从上表3-10可以看出，本文情感值量化方法比传统手工标注的方法好很多。

## 本章小结

本章首先阐述了情感词典在文本情感计算中的重要性，然后通过整理现有的基本情感词典得到本文的基本情感词典，再利用自然语言处理的方法在基本情感词典基础上构建了实体的属性及实体行为的情感词典，通过丰富情感词典，希望可以提高文本情感计算的准确度。

# 情感分类研究

情感分类也被称为意见挖掘。情感分类涉及多个领域，如自然语言处理、人工智能、自动文本分类、文本挖掘、心理学等。与基于主题文本分类方法不同的是，情感分类主要是识别文本中表达的观点和判别网民对于事物的态度。情感分类本身过程比较复杂，首先要对文本进行主客观分类来抽取主观文本，然后再对主观文本进行情感强度和极性的判别。主观文本的抽取只是情感分类的预处理工作，情感的极性判别和情感的强度判别才是情感分类的主要部分。其中极性判别主要用来判别文本中情感是正向积极的还是负面消极的。而强度判别则是判别文本情感倾向性强度，比如强烈正向、一般正向、中立态度、一般负向、强烈负向五个类别。

## 传统情感分类及不足

情感分类的目的是判断文本的情感类别，即该文本对某一主题是持支持还是反对态度。根据其实现的方法可分为基于词典和规则的情感分类和基于机器学习的情感分类。

#### 基于词典和规则的情感分类

基于词典和规则的方法主要是利用情感词典来对文本进行情感分类的一种方法。最简单的基于词典和规则的方法就是通过目前的情感词典来对测试文本中的正负情感词数目进行统计，通过比较正向情感词和负向情感词个数来确定文本的情感倾向。如果测试文本中正向词数目大于负向词数目，则判定文本的情感倾向为正向；如果小于则判定为负向；如果个数相等则判定为无情感倾向，该方法对高质量情感词典的依赖度较高。Turney等人利用词语间的相似度来对电影评论文本进行情感分类，他们首先利用现有的情感词典及电影语料抽取并筛选了一些正向情感和负向情感的基准词，然后利用待测文本中的词与基准词的相似度来对电影评论进行情感分类。Gamon和Aue[32]也是利用词语间的语义相似度来构建合适的情感词典，他们提出了一个新的假设，即极性相反的情感词通常不会出现于同一句子中，然后以Wordnet中的词语关系及词语注释为基准词，通过统计具有正向情感词语和负面情感词语的数目来对文本的情感极性进行判别。

#### 基于机器学习的情感分类

该类方法使用机器学习的模型，首先利用标注好的训练数据来训练分类模型，然后再通过这个模型来对待测文本的情感极性进行预测。目前常用的分类算法包括朴素贝叶斯、K近邻和支持向量机等分类算法。Pang等人首次将机器学习的方法应用到文本的情感分类中，将本文情感倾向分为正向积极和负向消极。他们首先对文本进行预处理并过滤掉无用信息，然后从文本中抽取了词性特征、词的位、unigram、bigram等特征，最后在SVM、NB、ME上分别进行对比实验。该实验结果显示，使用unigram特征时，SVM分类效果最好。Cui等人后来通过实验证明，当训练数据较少时，unigram作为特征的分类效果较好；但当训练数据的较多时，unigram作为特征的分类效果变差， n-gram(n>3)作为分类特征的效果要比unigram作为特征的分类效果好很多。杨震,赖英旭[33]通过上下文重构来进行情感极性判别。张林,钱冠群[34]对轻型评论情感分析做了更深入的研究。

从上可以看出，基于词典和规则的情感分类算法的准确度比较依赖情感词典的完善程度，而且由于情感计算本身的复杂性，只考虑情感词正负词语的个数特征并不能很有效地提高情感分类的准确度；基于机器学习的情感分类算法的情感特征主要为unigram、词性、bigram及n-gram，而这些特征主要是统计特征，由于本文内容灵活多变，所以统计特征就需要通过大量的数据来进行分析其对情感的影响，而且效率也不高。所以发掘更多更有意义的情感特征就势在必行。

## 特征选择

为了训练较好的情感分类模型，本文选取了七类文本特征，包括词性、情感词、否定词、程度副词、实体评价单元、行为评价单元及特殊符号。表4-1列出了分类模型的所有特征类型及其含义：

表4-1 分类模型的特征及含义

|  |  |
| --- | --- |
| 特征类型 | 含义 |
| 词性 | 名词、形容词、动词、介词、副词、叹词的个数 |
| 情感词 | 正负向情感词的个数、情感强度及得分 |
| 否定词 | 是否存在否定词 |
| 程度副词 | 情感词是否存在被修饰的程度副词 |
| 实体评价单元 | 正负向评价单元的数目、强度及得分 |
| 行为评价单元 | 正负向评价单元的数目、强度及得分 |
| 特殊符号 | 问号及感叹号个数 |

## 实验及结果分析

### 实验数据

本文的实验数据是从天涯论坛抓取数据中选取得到，这些数据包含200个训练语料和600个测试语料。表4-2统计了训练语料和测试语料中正向语料和负向语料的分布：

表4-2 语料统计结果

|  |  |  |  |
| --- | --- | --- | --- |
| 语料 | 正向语料 | 负向语料 | 总数 |
| 训练语料 | 100 | 100 | 200 |
| 测试语料 | 300 | 300 | 600 |

### 基于词典与规则结合的方法

#### 规则集

本文结合论坛文本的语言特性，从不同层次定义了情感分析的一些规则，这些规则包括句间关系规则、句型规则及词语关系规则。

##### 句型规则

句型规则主要考虑的不同句型对一个完整句子情感的影响情况，句型通过的标点符号进行判断。一个完整的句子把其称为复句，标点符号上的体现是：只有一个终结标点（句号、问号或是叹号）。那么对于完整句子的情感值为Es=αE，对于不同的句子类型，α值不同。α规则确定如下：

###### 若句子终结标点为“！”，则α值为1.5。

###### 若句子终结标点符号为“？”，且句中存在反问词，则α值为-1。

###### 若句子终结标点符号为“。”，则α值为1。

##### 句间规则

本规则是对各个分句之间关系的分析。本文只考虑在一个复合句内的各分句之间的关系，对三类主要关系进行了分析：转折关系、递进关系和并列关系，整个复句的情感值如公式（4-1）所示，其中E1代表前一个分句的情感，E2为后一个分句的情感，其规则如下所示：

 （4-1）

###### 转折关系

如果前一个分句与后一个分句中出现“可是”、“但是”或“然而”等承接词，则α取值0，β取值1。

###### 递进关系

如果前一个分句与后一个分句中出现“更加”或“更有甚者”等承接词，则α取1，β取值1.5。

###### 并列关系

如果分句间出现因果、并列标志词，或者不存在标志词，则α、β均取值1。

##### 词语多元组规则

可能影响句子情感倾向性的依存关系共有8种，分别为：ATT，ADV，CMP，SBV，VOB，VV，COO和CNJ。所以要对出现这些依存关系的情感关系对进行情感计算，规则如下所示：

###### 规则1

程度副词与情感词或实体评价短语构成的依存关系主要包括动补结构CMP和状中结构ADV，程度副词起到修饰情感词或实体评价短语的作用。记情感词或实体评价短语为word1，程度副词为word2，其组合的情感值的计算规则如公式(4-2)所示：

 （4-2）

###### 规则2

对于否定副词与情感词(或实体评价短语)构成的依存关系，比如状中结构ADV或动补结构CMP。需要分为两种情况考虑，当否定词直接修饰情感词时候，那么当前的词组情感值为直接翻转，计算规则如公式（4-3）；当否定词直接修饰程度副词时，起到了翻转并弱化程度副词作用，其组合情感值的计算规则如公式（4-4），其中word1为情感词，word2为程度副词：

 （4-3）

 （4-4）

###### 规则3

对于由名词与情感词(或情感词短语)构成的依存关系，如主谓关系SBV、动宾关系VOB或定中关系ATT，名词为情感词(或情感词短语)的特征词。记情感词(或情感词短语)记为word1，名词记为word2，其计算规则如公式（4-5）。

 （4-5）

#### 实验设计

为了发现新构建的情感词典对分类结果的影响，本文对实验进行了设计，首先在基本情感词典上进行实验，然后在基本词典和实体属性词典进行实验，最后在基本词典、实体属性词典及实体行为词典上进行实验，为了方便描述和展示，对各个词典进行编号表示，其中a表示基本词典，b表示实体属性词典，c表示实体行为词典，设计结果如表4-3。

表4-3 实验设计组合

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 1 | 2 | 3 |
| 组成 | a | a、b | a、b、c |

#### 实验及结果分析

在不同的情感词典组合上，运用规则集对文本的情感进行了计算，对计算结果进行统计，计算值与标注值相差在（-0.2，+0.2），则认为这两个情感值相同，对实验结果进行指标评价，统计结果如下表4-4。

表4-4 基于词典和规则情感分类实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 准确率 | 召回率 | F值 |
| 1 | 61.3% | 52.4% | 56.5% |
| 2 | 68.4% | 78.1% | 72.9% |
| 3 | 72.2% | 82.5% | 77.0% |

从上表4-4可以看出，加入实体属性词典和实体行为词典对于基于规则集的方法的情感分析的效果有了明显提高。

### 基于机器学习的情感分类

为了更明确一些特征对于情感分类的结果，本文首先对这些特征进行组合设计，然后利用情感分类算法在这些特征组合上进行对比实验。

#### 特征组合设计

对文本进行预处理、分词以及词性标注后，就可以把表中的前四个特征抽取出来，实体评价单元及行为评价单元的抽取方法与情感词典构建中用到的抽取方法相同。在这些文本特征中，不考虑情感词或实体评价单元而把程度副词、否定副词和特殊符号作为特征的实际意义不大。本文对特征组合进行了设计，首先将词性和情感词分别单独作为特征进行实验，然后将词性和情感词作为特征组合进行实验，最后在词性和情感词的特征组合中依次加上否定副词、程度副词等特征。希望通过多组不同的特征组合实验，来评估估各个特征对分类模型的贡献度，并找出对分类最有效的特征组合。

为了方便描述和展示，对各个特征进行编号表示，其中a代表词及词性，b代表情感词，c代表否定词，d代表程度副词，e代表实体评价单元，f代表行为评价单元，g代表特殊符号，设计的特征组合结果如表4-5所示：

表4-5 特征组合设计表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 组成 | a | b | a、b | a、b、c | a、b、c、d | a、b、c、d、e | a、b、c、d、e、f | a、b、c、d、e、f、g |

#### 实验结果及分析

本文使用朴素贝叶斯分类、支持向量机和k近邻分类方法在不同的特征组合上进行了对比实验，实验使用正确率和召回率作为评估指标，对情感极性及情感强度进行了分类。

##### 极性分类

本文在不同的组合特征上进行了极性分类实验，并对分类结果的准确率、召回率及F值进行统计，实验结果如下所示：

图4-1 不同特征上极性分类结果准确率变化曲线

从图4-1可以看出，当使用词及词性、情感词、否定词、程度副词，实体评价单元及行为评价单元的组合特征时效果最好，此时准确率可以达到81.4%，从准确率的增幅可以看出，基本情感词的对分类作用提升最大，程度副词以及特殊符号对于情感极性分类的提升不大，也说明了程度副词及特殊符号不适合作为极性分类的特征。从5、6、7实验结果比较发现，实体评价单元与行为评价单元作为特征对极性准确率均有提升效果。

图4-2 不同特征上极性分类结果召回率变化曲线

从图4-2可以看出，在词及词性、情感词、否定词、程度副词，实体评价单元及行为评价单元的组合特征上，各个分类算法结果的召回率也普遍比较高，而且支持向量机的召回率最高。

图4-3 不同特征上极性分类结果F值变化曲线

从图4-3可以看出，在词及词性、情感词、否定词、程度副词，实体评价单元及行为评价单元的组合特征上，F值也是比较高，而F值是准确率和召回率的综合考虑，因此在该特征组合上的情感分类效果是最好的。从上图可以看出，支持向量机的准确率及召回率都是比较高的，所以本文打算用支持向量机作为情感极性分类器。

##### 情感强度分类

本文在不同的组合特征上进行了情感强度分类实验，本文打算对强度分为强、中、弱三类，并对分类结果的准确率、召回率及F值进行统计，实验结果如下图4-4所示：

图4-4 不同特征上强度分类结果准确率变化曲线

从上图4-4可以看出，当使用词及词性、情感词、否定词、程度副词，实体评价单元、行为评价单元以及特殊符号的组合特征时强度分类效果最好，此时准确率可以达到70.3%，从准确率的增幅可以看出，情感词及程度副词对分类效果提升较明显，否定词对于情感强度分类效果提升不大，也说明了否定词不适合做为极性分类的特征。

图4-5 不同特征上强度分类结果召回率变化曲线

从上图4-5可以看出，当使用词及词性、情感词、否定词、程度副词，实体评价单元、行为评价单元以及特殊符号的组合特征时强度分类效果最好，各个分类算法结果的召回率也普遍比较高，而且支持向量机的召回率最高。

图4-6 不同特征上强度分类结果F值变化曲线

从上图4-6可以看出，当使用词及词性、情感词、否定词、程度副词，实体评价单元、行为评价单元以及特殊符号的组合特征时效果最好，各个分类算法结果的F值也普遍比较高，而F值是准确率和召回率的综合考虑，因此在该特征组合上的情感分类效果是最好的。最优组合特征时，支持向量机的准确率及召回率都是比较高的，所以本文打算用支持向量机作为情感强度分类器。

## 本章小结

本章针对情感分析的不足，运用现有的情感分类算法在构建的情感词典上进行实验。首先基于词典和规则方法进行实验，制定情感计算规则，并在规则上进行实验，并对结果进行分析；然后设计特征组合，在不同的特征组合在数据集上运用机器学习的方法进行情感极性及强度分类实验，并对实验结果进行分析。

# 基于社会网络情感分析

## 传统文本情感分析

目前国内情感分析的研究主要集中于对网民发布的信息进行情感的判断和挖掘，并对不同情感的强度进行统计，或者进行情感相关因素的分析研究，如互联网情感与社会重大影响事件的相关程度；通过情感分析结果来预测电影票房、政治选举结果和股市波动等；或将微博中的情感词与网民观点相结合进行分析，而对互联网中网民情感传播特征的研究还比较少。

## 基于社会网络情感分析

个人情绪的表达会影响到社会环境中的其他人。情绪不仅能展示表达者的心理状态，而且还能传递出社会环境信息，如对事物的观点、对其他人的态度等，观察者还能通过个体情绪来推测情绪表达者对某种状况的认知评价，这就是情绪-社会信息模型（EASI）。EASI模型说明了情绪是一种社会信息，可以在人们交流过程中传播，不同的情绪传递的信息不同。情绪通过表达可以给观察者提供一些信息，从而影响到观察者的情绪或行为。EASI 模型为人际交往中情绪相互影响的过程提供了一个框架，VanKleef[35]给出了该框架的详细解释，论证了情绪可以通过激发观察者情绪反应和推理过程进而影响观察者的行为，不同的观察者的推理过程和情绪反应的强度不尽相同，这主要是由于不同的观察者的社会关系或信息处理能力不同造成的。

社会网络分析方法简称sNA，是近年来兴起的一种新的研究范式。sNA是主要分析某一特定空间范围内行动者的关系状况，寻找关系的特征以及发现关系对整个网络的影响。在构建的网络图中，通常用节点来表示行动者，用边表示行动者之间的关系，边的数值用来表示两个节点连接关系的强弱；在论坛情感网络构建中，我们可以把参与情感传播的用户作为节点，利用他们之间的评论互动关系来构建情感传播网络。在sNA视角下，行动者通过评论互动的关系而形成了情感传播网络，情感信息正是通过这种网络结构进行传播。情感传播网络整体上会影响群体的互动行为和情绪状态，而单个行动者在整个网络中的位置也决定了个体的信息行为及效果；除此之外，行动者之间的互动会产生新的连接关系，这些关系又会反过来来影响情感传播网络的结构。社会网络分析正是通过对基于互动关系形成的情感传播网络的结构分析来理解情绪的传播机制。该方法有利于解决个体主义的局限性问题，将社会结构因素与传播主体的个体属性结合起来，可以从全新的视角来对情感传播进行分析。

情绪感染理论和 EASI 理论为互联网情感传播研究提供了理论支持。情绪感染理论与EASI 理论均说明了情绪在人与人之间传播的机制，情绪感染理论将模仿视为形成情绪传播的主要因素，而EASI 理论将情绪视为一种信息，将情绪信息接收者的推理和情感反应视为形成情绪信息传递的主要因素，通过社会网络分析方法可以将情感传播更直观的展现出来。

## 基本模型与指标

### 基本模型

#### 实体用户与文本内容情感模型

实体用户在发帖及评论过程不仅展示了他对一些问题的见识及看法，而且这些文本也显露出该用户的基本情感状态。所以本文制定公式计算实体用户的情感，Et为用户所发帖子的情感，Ec为参与互动评论的情感，α、β分别为帖子情感与评论情感的权重比例，公式如（5-1）所示：

E=αEt+βEc （5-1）

本文中的α值及β值是根据帖子数量及评论数量比例进行确定的，α与β的比值等于评论数量与帖子数量的比值。

#### 情感传播模型

实体用户在互动过程中会自动地、有意识或无意识地模仿他人的表情、声音、动作，在模仿过程中感受他人情绪且影响自我情绪体验，社会心理学把该过程命名为情绪感染。群体中的个体情绪在情绪感染的作用下会逐渐融合为同一情绪，即群体情绪。

在计算机领域对群体情绪模拟研究主要是以情绪感染理论作为基础，其中构建群体情绪模型的方法最常见的理论包括传染病机制和热传递机制。部分研究表明基于传染病机制的传染模型与情绪传染最接近，基于传染病机制的情绪模型主要有以下特点：a)存在感染者和传染者；b)传染者的情绪会自然的复制给感染者。而热传递理论下的情绪模型说明情绪在个体间可以相互感染，且情绪在每一次感染过程中仅作局部改变。

许多研究者在研究情人际交往感传播过程时，考虑了个体之间不对等的情感关系，把人际关系划分为三类：强情感关系(家庭成员或亲属之间的关系)、弱情感关系(同学或朋友之间的关系)和陌生关系(陌生人)，来表示不同的人际关系。本文中的数据主要来自论坛，论坛中个体用户基本是平等的、自由的，虽然个体之间没有人际交往中那么明显情感依赖关系，但是也存在情感依赖不对等的关系。本文将情感关系分为两类：强情感关系（意见领袖对其他用户）与弱情感关系（普通用户之间，普通用户对意见领袖），来模拟情感传播过程。本文制定公式（5-2）及（5-3）来描述不同情感关系Ei表示个体i的情感状态，Ej表示个体j的情感状态，β表示个体i对个体j的影响系数，α表示个体j对个体i的影响系数，α与β值确定如公式（5-4）及（5-5）所示，其中Ejn表示第n个与个体i互动的用户，Ein表示第n个与个体j互动的用户。

 （5-2）

 （5-3）

 （5-4）

 （5-5）

#### 热点帖子模型

热点事件主要用来度量舆情受网络用户关注的程度，并可以通过网络参数：点击量、回复量和参与人数来进行组合计算，其中公式如（5-6）及（5-7）所示：

 （5-6）

，当t>0时：

 （5-7）

若某一天的N大于所设定阈值，则该天为有效天数，tef为有效天数总和，进行归一化处理后，得到最终结果。其中，CH为舆情热度，N1为回复数量，N2为参与人数，相应权值为g1、g2、g3；N3为点击量，N为中间计算结果，AN 为有效天数的中间计算结果总和g3，为点击量所占权重，g4为中间计算结果所占权重。

### 基本指标

#### 用户节点指标

用户节点指标主要表示节点的一些特征或属性，本文用户节点指标包括用户影响力及价值观，这些指标有助于情感网络特征的分析。

##### 用户影响力

用户影响力是用来度量用户在社区的作用与影响力的大小。该指标通过浏览次数、发帖数、帖子回复数和粉丝数四个网络参量来进行度量。公式如（5-8）及（5-9）所示：

 （5-8）

 （5-9）

其中，AI用户影响力，N1为总文章被浏览次数，N2为发布文章数，N3为文章被回复数，N4为粉丝数，权重分别为g1、g2、g3、g4；N为中间计算结果，g5为其所占权重。在计算作者影响力中用到了simgoid函数，如（5-10）所示：

 （5-10）

由于当x=6的时候该函数已经非常接近1，而且随着x继续增大该值变化特别小，所以通过归一把x控制在0-6之间，因为sigmoid函数的值域为[0.5,1),2\*simoid(x)-1的值域为[0,1),故最后的函数定义如上所示。

##### 价值观

价值观就是指的是在网络的世界中，节点用户体现出来的对事物的评价和态度取向。

#### 舆情传播类指标

舆情传播类指标主要目的在于确定舆情在网络中的传播渠道和传播影响的环节，并着重可以分析网络舆情所处的传播和扩散阶段。它包括以下三个子指标：

##### 舆情影响力

舆情影响力主要是指舆情传播过程中对网民的影响程度，通过某一专题下所有帖子的日流量和总流量随时间变化的趋势图来衡量。日流量即所有帖子的日回复量之和，总流量即为所有帖子的总回复量之和

##### 发展峰值

舆情讨论最激烈的点，即单位时间内回复量最高的点即为发展峰值，监控该点可以得知导致舆情变化的原因与趋势。

##### 发展拐点

即情感倾向发生变化的点，通过正面回帖数与负面回帖数的差值来进行计算，监控该指标可以有助于分析引发网络大众情感发生变化的拐点与原因。

#### 舆情预警指标

舆情预警指标主要目的在于发现舆情的爆发的时间点。本文的舆情预警主要是考虑事件的负面指数来进行预警，该指标计算公式（5-11）如下：

 （5-11）

其中negNum表示负面回复数或负面参与人数，posNum表示正面回复数及正面参与人数，dReplyNum代表日回复数或日参与人数。

## 实验及结果分析

### 数据抓取

通过对网络论坛文本的属性建模与特征分析，发现这些论坛网站的文本信息几乎都是按照固定的主题板块进行划分的，而且在不同主题板块下均按照发表的时间先后顺序排列。因此针对这些比较规则的信息，在设计网络爬虫时可以根据网站的顶层URL进入到网站的网站地图中，从网站地图获取每个板块对应的URL，将这些板块的URL存起来作为爬虫的入口。

网络爬虫包括三个主要的模块，第一模块为URL收集部分，这部分是用来收集所有要抓取的网页的URL，将其按照顺序存入到系统的队列中；第二模块是对网页信息的解析，即针对相应页面上的结构特征与元素特征进行解析。由于网页上的特征信息较多，先过滤掉帖子页面中的存在的广告，并将需要的部分进行内容解析并存储，例如对于论坛中的帖子，就需要收集帖子的标题、内容、发表时间、所属板块、作者、回复、回复人等信息；第三个模块是存储模块，将第二部分获取到的数据信息与索引关系存储到相应的数据库中。

为了将网络爬虫获取的文本数据有效存储，本报告设计了4个关键的数据库表，即：数据源表、评论表、用户表。数据源表主要是用于存储抓取的文本数据的主体，包括数据URL、标题、内容、发布时间、来源网站、访问数、回复数、作者、所属板块等关键字段。评论表用于存储每个数据源的评论情况，包括评论内容、评论时间、所属数据源以及评论作者等字段，评论表与数据源表相关联。用户表主要是用来存储发布文章或是发表评论的作者的基本属性信息，如：作者名称、URL、所属网站、发表文章数、积分、最后登录时间、注册时间、在线时间、精华数、评论数、个人简介等主要字段。形成了如图5-1所示的数据存储的底层结构。

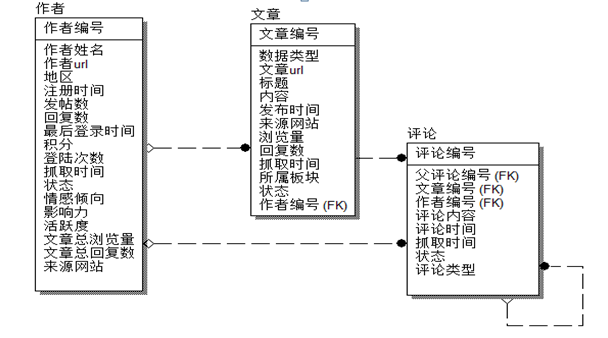


图5-1 爬虫数据存储的底层结构示意图

另外，为了保证爬虫抓取数据的效率和质量，在设计过程中增加了对重复URL以及无法访问URL的过滤，并对不同的属性特征进行融合，例如在有的网站上用户在线时长是按照登录的天数计算的，但是有的网站上用户在线时长是按照在线的小时计算的，为了统一可以按照两小时算作一天的规则将其转换成按照在线天数计算；另外，对一些名称不同但是实际意义相同的属性进行合并或是转换，例如，有的网站上对于用户观点表示支持的属性为支持数，但是有的网站同样意义的属性称为“赞”，因此，可以通过对所有类似的同义但不同名的属性进行转换并实现数据的统一，从而保证了通过不同页面规则的定义来实现了网页内容的自动抓取。

本文在上述模型的基础上，采用基于Jsoup开源工具来开发和实现了一个网络爬虫，并对国内天涯论坛2014年的全部数据进行抓取和存储。该网站具有大量的用户群体，并且发布的文本内容丰富并具有较大的影响力，具有较高的样本示范作用。因此，从该网站上收集到的网络文本数据对本报告的深入研究提供了基础数据来源，具有较好的研究价值与实际应用价值。

### 数据预处理

在数据挖掘整体过程中，海量的原始数据中存在这大量复杂的，重复的，不完整的数据，不仅会严重降低算法执行的效率，而且甚至影响数据挖掘结果的准确性，因此，在对数据进行处理之前，必须要对采集到的数据进行预处理，希望通过提高数据的质量，来使数据的处理效果达到最好。数据预处理阶段主要包括数据清理，数据集成，数据变换与数据规约等技术。

#### 数据清理

数据清理主要是去除数据集中的无关数据和噪声数据，主要包括处理缺失数据、脏数据清洗以及删除重复信息等。

在抓取论坛数据过程中，由于网络或者其他原因，抓取的数据中存在乱码，这些乱码会严重影响后续结果的处理；除此之外，本文的研究的重点是情感分析中的舆情分析，舆情分析主要是研究人们对于热点事件或者焦点问题的看法或者观点，而抓取的数据中存在许多非热点或者普通数据，这些数据对于结果的分析并没有实质性的作用，所以要合理的删除这些数据。

##### 乱码监测

为了清除这些脏数据，必须监测到这些乱码数据，本文针对制定一些策略对乱码进行监测。对于获取的文本数据，首先去除文本中的空白字符，对于不是数字或者字母的字符进行判断，如果该字符是不是汉字，则对其进行统计，如果统计结果大于阈值，则判定该文本为乱码。流程图如图5-2所示：



图5-2 乱码处理流程图

##### 无关数据清理

抓取数据中存在许多不重要的帖子，许多帖子的评论都比较少，这些数据不仅对结果没有帮助，而且还会造成处理资源的浪费，影响处理速度，所以这些数据必须得到清理，本文设计了公式对这些不重要的帖子及评论进行过滤并删除。公式如（5-12）所示：

 (5-12)

其中N1 为帖子的浏览量，N2为帖子的评论量，g1与g2 为浏览量与评论量的权重。

#### 数据变换

数据变换是通过平滑聚集，数据概化，规范化等方式将数据转换成适用于数据挖掘的形式。为了把数据转换成适合挖掘的数据，本文进行了数据合并及数据抽取。

##### 数据合并

抓取的数据中有许多帖子是描述同一个事件，所以为了数据的完整性，必须对这些数据进行合并，本文制定了一些策略，首先对帖子标题进行分词，通过停止词库过滤掉停用词，然后计算tf-idf值，进行相似度比较，对于相似度高于阈值的，且发表时间间隔小于十天的直接合并为同一事件，流程图如图5-3所示：



图5-3 数据合并流程图

##### 数据抽取

论坛中的评论互动主要是通过“@”符号来展示的，要对评论内容进行分析，必须先对评论内容进行抽取，所以本文设计了一些策略对用户评论进行抽取。首先遍历评论文本，找到存在“@”符号的文本，如果对所评论的用户有评论内容，则再开始查找“@”的用户是否在数据库中存在，如果存在则直接加入把该条评论加入评论列表中，如果该用户不存在，则去除“@”内容，保留文本内容。流程图如图5-4所示：



图5-4 评论数据抽取流程图

### 数据处理

数据处理是从大量的原始数据抽取出有价值的信息，即数据转换成信息的过程。本文的数据处理主要包括指标计算、情感计算与统计及情感网络数据统计。

#### 指标计算

由于指标计算中需要许多中间数据，所以需要先对中间数据进行统计，然后再通过指标公式进行计算。本文指标计算包括数据统计及指标值计算。

##### 数据统计

预处理过程中会删除一些无用的帖子，还会抽取一些评论加入到评论列表中，所以要对一些数据进行重新统计，例如事件回复数，发帖数等，除此之外还要统计一些指标计算过程中的影响因子，例如事件参与人数，参与回复数等。最后把这些统计的因素数据暂存到数据库中。

##### 指标值计算

统计完一些计算需要的影响因子，然后就要对指标值进行计算。本文首先把指标公式转换为sql，然后要对sql进行优化，提高sql的执行效率，通过计算完的指标结果进行对比及分析，合理的调节影响因子的权重，使指标值更加合理。

#### 情感计算与统计

通过情感分类实验得知，基于机器学习的情感分类算法要比基于词典和规则的情感分类算法效果好。在设计不同的特征组合上，通过朴素贝叶斯、kNN分类算法和SVM分类算法的比较得知，SVM分类器在的分类效果要比其他分类器相对好一些，所以本文通过SVM分类器来对帖子及评论进行情感值计算，首先运用对极性分类特征效果较好的特征组合（词及词性、情感词、否定词、实体评价单元及行为评价单元的组合特征）对文本进行极性分类，分为褒义、贬义及中性三类；然后运用情感强度效果较好的组合特征（词及词性、情感词、程度副词、特殊符号、实体评价单元及行为评价单元的组合特征）进行强度分类，分为强、中、弱三类（对应数值对0.8，0.5，0.2），最后根据极性分类结果与情感强度分类结果对帖子及评论文本情感值进行打分，例如对于强的正向情感值对应打分为0.8，弱的负向情感值对应分值为-0.2。

在文本情感分值确定的情况下，就可以通过用户与帖子情感模型可以给用户进行情感打分，把这些情感值嵌入到情感网络中，就可以对情感网络进行特征分析，也可以通过社会网络分析方法对情感网络进行分析，研究情感传播规律。除此之外，对于具体的热点事件，本文可以分析一个用户情感随时间的走势，这样可以及时的发现存在负面的用户。通过统计事件中的不同日期中的正负向评论数量，可以展示该事件随时间的情感走势。也可以分析一个事件中某个时间点正负向用户的数量来对事件进行监控，及时发现事件的舆情爆发点，防止事态严重化。

#### 情感网络特征数据统计

由于需要对于情感网络特征需要进行分析，所以首先要对情感网络特征所需数据进行统计，针对互惠性分析、结构平衡理论及同质性理论，本文分别设计策略对其所需要的数据进行统计。

##### 互惠性分析数据统计

情感网络的互惠性分析，即分析两个用户在相互交流过程中情感的互动过程，所以需要双方均有评论互动，可以通过统计双方在互动过程中的情感差值来表示互惠程度，情感差值越小，说明该互动越满足情感互惠性。本文制定了一些策略来统计这些数据，首先定义实体Entity来存储互动关系及情感值，然后获取到一个帖子，对其评论按照时间进行排序，得到排序好的评论列表，开始从评论列表中取数据，存储区队列是否包含这样的这样的实体 Entity，如果包含则加入该队列中，如果不包含，则创建队列，加入队列后存储到列表中，最后求出所有队列里的互惠程度，帖子评论数据遍历完毕，如果队列中不存在互动（只有单一用户给另一用户发消息，没有反馈信息），舍弃该队列，存在互动则计算互惠值并存储。

互惠性数据统计的流程图如图5-5所示：



图5-5 互惠性数据统计流程图

##### 基于结构平衡理论和状态理论数据统计

###### 结构平衡理论数据统计

结构平衡理论主要是研究三个用户节点之间的关系，所以需要获取共同出现的三个用户节点，很明显三个用户节点不会仅局限于一个帖子，所以要对全局数据进行统计。本文制定了一些策略来进行数据的统计，使用map来存储互动关系，由于评论较多，一定数据量后计算机内存会被耗尽，为了防止这种问题发生，本文首先把帖子分成块，通过map对关系进行暂存，如果map中已经包含某互动关系，则对计数器进行加1，块中的帖子已统计完时，将统计结果写入txt保存，直到所有的帖子数据统计完，然后对txt中的数据进行两两合并，最后对结果进行筛选，保留共现率超过10的用户，流程图如图5-6所示，在共现用户的基础上，就需要通过这些评论用户的交互情况来判断这些用户之间的关系，如果两个用户交互频繁且交互的情感倾向值为正向，则认为这两个用户是“友好”关系，如果两个用户交互频繁且交互的情感倾向为负向，则认为这两个用户为“敌对”关系，如果两个“友好”关系的用户对其他某个用户均为“友好”关系，则统计数据加1，如果两个友好关系用户对某个用户均为“敌对”关系，统计数据加1，最后统计结果进行保存。



图5-6 平衡理论分析数据统计

###### 状态理论数据统计

状态理论认为，用户A对用户B的情感态度与用户A及用户B在该网络中的地位有关，所以本文对所有用户间的评论进行统计及用户影响力进行统计，如果评论中用户A与用户B的情感态度与其网络地位（通过用户影响力来表示）相符，则计数器加1，不符则不记录。

##### 同质性分析数据统计

在对情感平衡性数据统计时已对评论共现用户次数较高的进行保存，所以本文只需要统计这些共现用户的情感情况，如果两个用户的价值观相同，即两个用户的情感值相差在（-0.3,0.3）区间内，则计数器加1,最后保存结果。

### 结果及分析

#### 互惠性分析

数据处理过程中已对用户评论互动数据进行了的统计，若两用户的情感差值±0.3，则认为双方情感相同，统计结果如下表5-1所示

表5-1 互惠性分析数据统计结果

|  |  |  |
| --- | --- | --- |
| 互动情感相同 | 互动情感不同 | 总数据 |
| 432737 | 165793 | 598530 |

通过上表得出，互动过程中情感相同的比例可以达到72.3%，表明用户之间倾向于彼此表达相同的情感，情感网络满足互惠性原则。

#### 基于结构平衡理论与状态理论的情感分析

##### 结构平衡理论数据分析

对互动次数超过10的用户进行平衡理论统计，结果如表5-2所示：

表5-2 平衡理论数据统计结果

|  |  |  |
| --- | --- | --- |
| 满足平衡理论 | 不满足平衡理论 | 总数据 |
| 5122 | 5776 | 10898 |

通过上表看出，满足平衡理论的比例占47%，可以看出该情感网络不满足平衡理论，可能与论坛的弱社交因素有关。

##### 状态理论数据分析

对所有的评论数据进行状态理论分析，统计结果如表5-3所示：

表5-3 状态理论数据统计结果

|  |  |  |
| --- | --- | --- |
| 满足状态理论 | 不满足状态理论 | 总数据 |
| 371748 | 2262782 | 598530 |

从上表看出满足状态理论的数据占62%，可以看出论坛用户在其情感网络基本地位平等，除了意见领袖与普通用户之间网络地位有所差别，其他用户之间地位基本平等，无明显区分，这可能与论坛的弱社交因素相关。

#### 情感网络同质性分析

通过对评论数据进行同质性分析，统计结果如表5-4所示：

表5-4 同质性分析数据统计结果

|  |  |  |
| --- | --- | --- |
| 满足同质性 | 不满足同质性 | 总数据 |
| 8298 | 2667 | 11595 |

通过上表可以看出，满足同质性分析的为77%，可以看出用户更倾向于跟自己价值观相同的进行交流，即正向用户比较愿意与正向用户交流，负向用户更倾向于负向用户进行互动，情感网络满足同质性理论。

#### 情感网络传播分析

随着社交网络媒介的迅速发展，热点事件的传播也呈现出新的特点。与传统媒介相比，新媒介中的热点事件不仅传播速度快而且主观性比较强，会造成较强的影响力和爆发力。在群体传播过程中，事件很容易传播失实且被恶化，事态更容易往极端的方向发展，所以对热点事件的传播及情感传播的研究就变得很有必要。

##### 热点事件

热点是比较受广大群众关注或者欢迎的新闻或者信息,或指某时期引人注目的地方或问题。热点事件就是广大群众比较关注的事件或问题，对于这样的事件，网民的参与度会比较高，讨论会比较激烈，流露出的情感比较丰富，情感传播效果比较明显。通过热度模型计算了帖子的热度，选取了热度较高前10的帖子，帖子列表如5-5所示：

表5-5 热点事件Top 10列表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 帖子标题 | 发帖人 | 发帖时间 | 点击量 | 正面回复 | 中性回复 | 负面回复 |
| 1 | 山西煤老板吕中楼豪赌欠债遭追索 欠赌资35亿港元 | 薄荷  茶hoho | 2014/5/10 | 2680144 | 3397 | 398 | 4101 |
| 2 | 出大事了，深圳市长市委书记管不管？ | 杨真M | 2014/7/28 | 603285 | 3692 | 605 | 6530 |
| 3 | 盘锦会计“被有罪”肇州法院是否“葫芦僧乱判葫芦案” | 沙滩上的鱼2014 | 2014/3/2 | 27950 | 4854 | 276 | 3447 |
| 4 | 文成县常务副县长李某某为妻弟谋利非法占用农保田 | 习小兵 | 2014/2/5 | 134957 | 2746 | 741 | 4738 |
| 5 | 北大生命科学院院长：决不能因为一小撮极端偏执的人，错过转基因技术革命 | 神完意适 | 2014/8/12 | 166855 | 3600 | 477 | 2769 |
| 6 | 马来西亚失联航班原定飞往北京，航班号是MH370，机上239人 | 刘文宪 | 2014/3/8 | 1834980 | 1984 | 239 | 2033 |
| 7 | 跟随背夹巡线人挑战悬崖 | 宇航摄行者 | 2014/7/8 | 1003375 | 2204 | 180 | 1781 |
| 8 | 大庆市肇州法院，判决书“张冠李戴”天下奇闻 | 龙哥123 | 2014/3/6 | 20634 | 919 | 68 | 1237 |
| 9 | 致黑龙江省委、公安厅及全体干警的一封信 | 梁革亭治 | 2014/8/4 | 19549 | 1460 | 220 | 1343 |
| 10 | 党的群众路线教育实践活动，能保群众不受政府职能部门伤害不？ | 新手L | 2014/2/9 | 626817 | 1649 | 197 | 1691 |

从上表5-5可以看出，较热的帖子的回复数量和点击量都比较高，而且许多帖子多涉及到官员或政府部门，这与许多中国人民的“仇富”或“仇官”心态有很大的关系，网民们很乐意把对于官员普遍的不信任感和“有罪推定”心理定势通过讨论发泄出来。从上面表还可以看出，大多数存在负面倾向帖子（1,2,4,8）的评论数量中，负面评论的数量也相对较多，可以激发人们正向情绪的帖子（5,7）的正面评论较多，这也说明情绪具有传染性，网民们对于负面情绪更容易引起负面的回应。

##### 情感传播网络分析

为了更清晰的展现情感传播过程，本文对帖子情感传播网络进行构建。首先通过遍历帖子评论，然后把互动用户及其情感值进行保存，最后运用gephi对其进行可视化。通过对top10的帖子传播特征进行比较，最终选择了传播效果比较好情感网络传播图进行展示，如下图5-7所示：

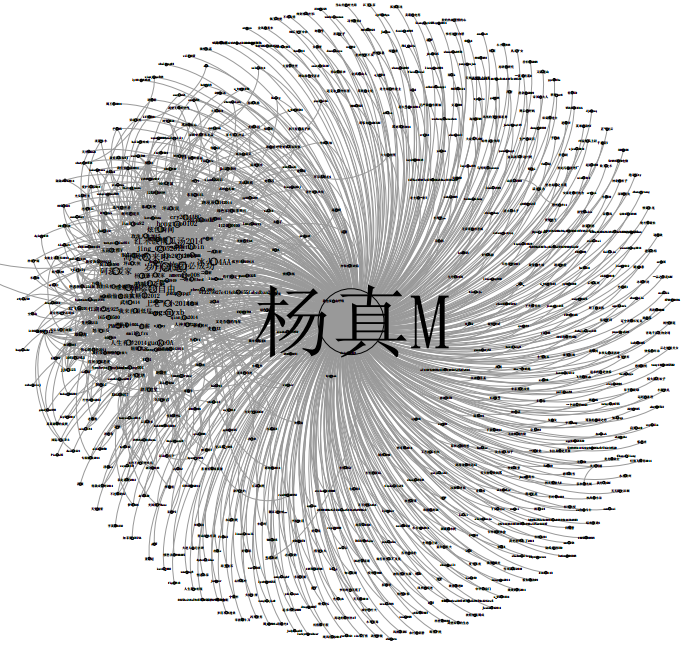


图5-7 帖子情感网络传播图

从上图可以看出，度数比较大“杨真M”为该传播网络的中心，旁边还有一些度数相对较大的节点，这些节点主要互动频繁的一些用户组成。

对该情感传播网络社会网络指标进行分析，结果如下所示：

###### 基本指标

该网络共由 735 个节点和 1929 条边构成，基本指标结果如表5-6所示：

表5-6 基本指标表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 指标 | 平均度 | 图密度 | 网络直径 | 模块化指标 | 平均路径长度 | 平均聚类系数 |
| 数值 | 5.249 | 0.007 | 3 | 0.001 | 1.998 | 0.681 |

平均度：较小；图密度：该传播网络图比较稀疏；网络直径：较小，该网络的传播特性不明显；平均路径长度：居中，网络传播较广；平均聚类系数：比较高。

###### 中心性

使用 UCINET 对该网络的中心度指标进行计算，并列出这些节点的点度中心度、中间中心度和接近中心度三项指标的绝对值、相对值，结果如下表5-7所示。

表5-7 点度中心度Top10列表

|  |  |  |  |
| --- | --- | --- | --- |
| 用户名 | Degree | NrmDegree | Share |
| 杨真M | 9891 .000 | 1.404 | 0.428 |
| smile\_frx | 987.000 | 0.140 | 0.043 |
| 弹奏我自由 | 382.000 | 0.054 | 0.017 |
| 岁月流逝J | 350.000 | 0.050 | 0.015 |
| 红米饭南瓜汤2014 | 324.000 | 0.046 | 0.014 |
| 齐心抱团必成功 | 323.000 | 0.046 | 0.014 |
| 冰火14AA | 308.000 | 0.044 | 0.013 |
| denhbin | 287.000 | 0.041 | 0.012 |
| 上老下小2014 | 269.000 | 0.038 | 0.012 |
| amengye008 | 267.000 | 0.038 | 0.012 |

在表5-7中，列Degree是绝对点度中心度值，列NrmDegree是相对点度中心度值，点度中心度表示某个节点与其周围节点直接相连数目。从上表可以看出，“杨真M”节点度数最大，说明其在网络中与其他用户节点直接交互比较多，其居于该网络的核心。根据情绪感染理论，情绪可以通过文本交互进行传播，某个用户的节点度数越大，那么该用户的情绪就会越多地感染给其他用户，对于整个情感网络中的情绪值有较大的贡献，所以积极监控或引导这些用户的情绪可以有效引导网络中与其直接相连其他用户的情绪，对于事件的积极传播有很大的重要意义。

中间中心度统计Top10如表5-8所示：

表5-8 中间中心度Top10列表

|  |  |  |
| --- | --- | --- |
| 用户名 | Betweenness | nBetweenness |
| 杨真M | 148625.734 | 27.624 |
| denhbin | 32428.033 | 6.027 |
| djh20142014 | 25344.193 | 4.711 |
| 一点薪 | 22304.961 | 4.146 |
| 乖乖不坏 | 21438.770 | 3.985 |
| 上老下小2014 | 17693.338 | 3.289 |
| 岁月流逝J | 15225.149 | 2.830 |
| qianjin2014 | 11003.548 | 2.045 |
| 齐心抱团必成功 | 9230.020 | 1.716 |
| 红米饭南瓜汤2014 | 8944.784 | 1.663 |

在表5-8中，列Betweenness表示中间中心度，列nBetweenness表示相对中间中心度，中间中心度测量的是行动者对资源信息的控制程度。如果一个点处在其他点的交通路径上，则该点的中间中心度就越高。在情感网络中，一个节点的中间中心度越高，说明该节点情感在情感传播中充当的媒介的作用越大，那么该节点用户的情绪值对于整个网络情绪贡献值也越大。所以适当的引导这些用户的情绪可以控制整个网络的情感值。从上表可以看出，“杨真M”的中间中心度值最大，同时其节点中心度值可以最大，说明该用户不仅对其直接相连的用户的情绪造成影响，而且会对全网的情感情绪造成很大影响，所以这样的用户必然是舆情监控的重点用户。从上表可以看出许多用户的点度中心度与中间中心度都比较高，那么通过控制与引导这些节点用户完全可以控制整个情感网络情绪值。

接近间中心度统计Top10如表5-9所示：

表5-9 接近中心度Top10列表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 用户名 | inFarness | outFarness | inCloseness | outCloseness |
| 杨真M | 2202.000 | 363709.000 | 33.333 | 0.202 |
| zhyf112 | 2839.000 | 539490.000 | 25.854 | 0.136 |
| 上老下小2014 | 2902.000 | 363659.000 | 25.293 | 0.202 |
| BuleBlood | 2912.000 | 363713.000 | 25.206 | 0.202 |
| 一点薪 | 2913.000 | 363645.000 | 25.197 | 0.202 |
| 武松1414 | 2913.000 | 363948.000 | 25.197 | 0.202 |
| 20140809ABD | 2915.000 | 363944.000 | 25.180 | 0.202 |
| djh20142014 | 2919.000 | 363625.000 | 25.146 | 0.202 |
| jhsjsky | 2920.000 | 363948.000 | 25.137 | 0.202 |
| denhbin | 2921.000 | 363613.000 | 25.128 | 0.202 |

在5-9表中，inCloseness为内接近中心度，outCloseness为外接近中心度，这里本文只用到内接近中心度，接近中心度衡量的是该节点到达其他节点的距离，距离越短，接近中心度值越高。情绪传染过程中，某节点到达其他节点的距离越短，那么该节点的情绪值影响到其他用户的情绪值的权重越大，所以这些用户也是舆情监控的重点。从上表可以看出，节点“杨真M”的点度中心度、接近中心度及中间中心度均是最大的，同时还有许多用户如“上老下小2014”，“denhbin”等的中心性指标值都比较高，说明这些用户在该情感网络中趋于核心地位，他们的情感值基本上决定了整个情感网络的情感值，所以对这些用户进行监控就可以捕获事件的情感传播趋势。

##### 事件详细分析

本文对上述帖子进行了分析，针对该帖子绘制了“日流量-时间”趋势图、“情感总值-时间”趋势图。并在情感分值基础之上将评论文本情感倾向划分为3个等级，褒义(Z≥α)；中立态度(β<Z<a)；贬义(z<=β)，a和β分别表示褒义情感和贬义情感的阈值，根据这三个等级划分，绘制了“情感分类-时间”趋势图，通过这些图可以明显看到事件随时间发展的舆论倾向变化。

该帖子的日流量图如图5-8所示：

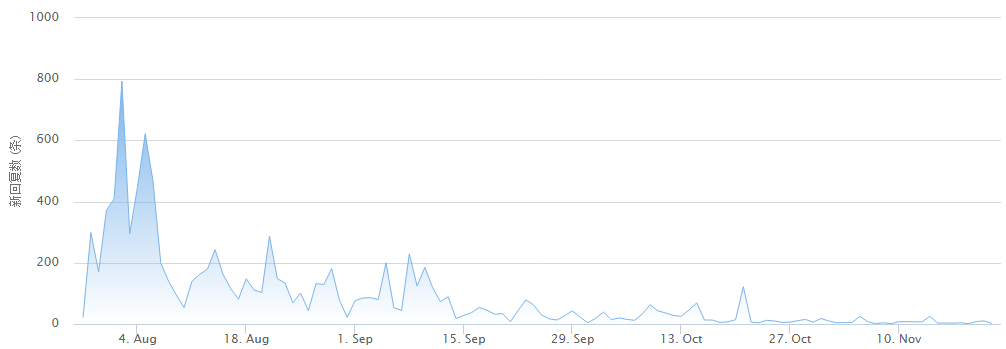


图5-8 日流量趋势图

从上图5-8可以看出，该帖子在2014-07-28发表以后，立刻引发了激烈的讨论，2014-08-02日流量达到峰值，随着时间的推移，日回复量也渐渐减少，说明大家对于此事的关注度也逐渐下降，中间的关注度有回涨趋势，可能是因为其他杂志媒体的滞后性或者新的曝光材料出现及相关事件的发生。

情感倾向值随时间的趋势图5-9如下所示：

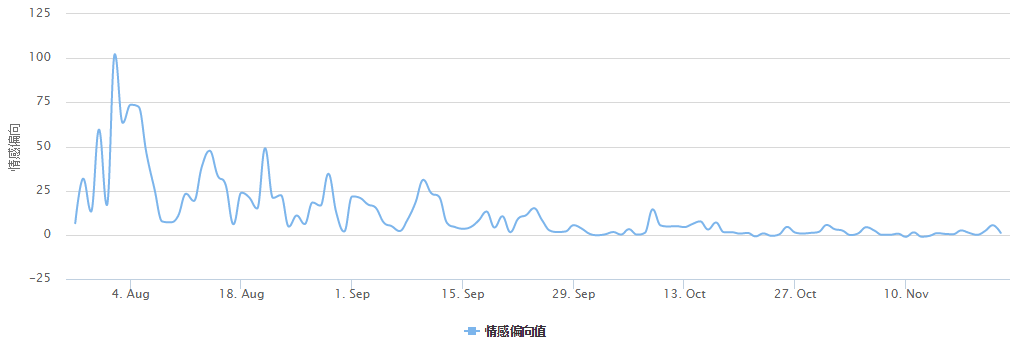


图5-9 评论情感趋势图

从该情感趋势图5-9可以看出，该情感趋势走势与日流量的大体走势相同，说明网民的情绪还比较稳定，正向评论数量与负向评论数量相当。

评论情感分类随时间趋势图如下所示：

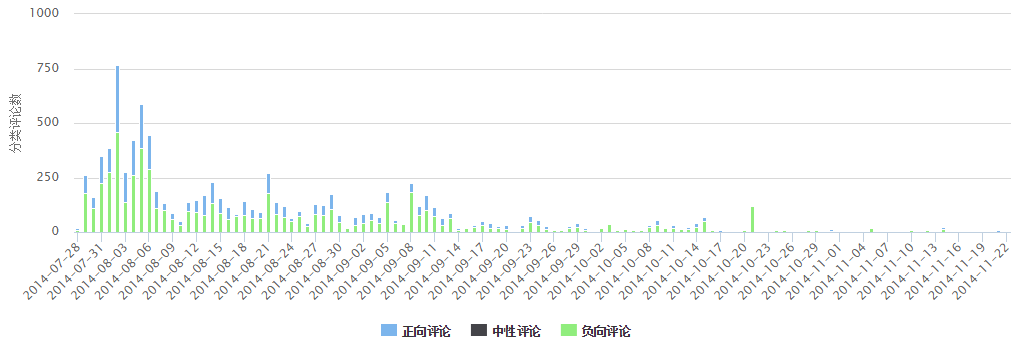


图5-10 评论情感分类图

从图5-10可以看出，在2014-08-02日回复量达到峰值，但正负向的评论数量比例相当，说明大家虽然比较关注该事件，但是情绪比较稳定。2014-08-05的日回复量再一次达到峰值，且该日的负向评论比例比较大，可以对这样的时间点进行监控来防止事件进一步加重。

由于许多负面评论可能是部分网络推手或者只有很少部分人造成的，所以单从负面评论的数量不能确定该事件的是否达到预警程度，针对此问题，本文绘制了“参与人数趋势图”，如下图5-11所示：

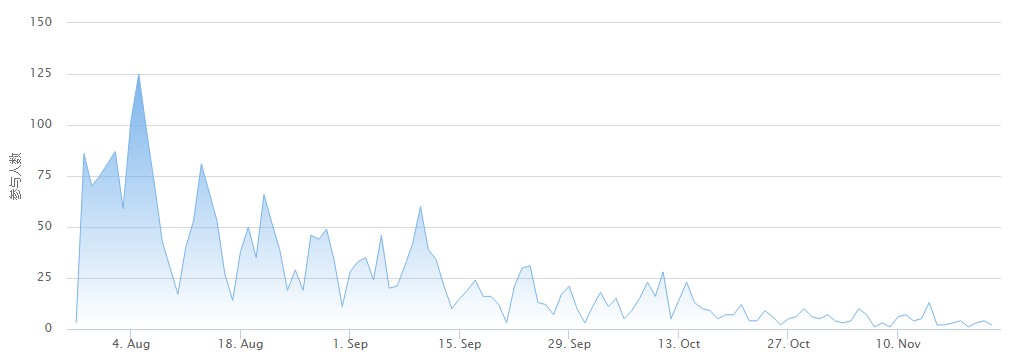


图 5-11 参与人数趋势图

通过图5-11与日流量图对比可以看出，在2014-08-02日流量虽然达到了峰值，但此时的参与讨论的人数并未达到峰值，可能该事件还未引起广大网友的关注，在2014-08-05日，参与讨论的人数达到了峰值，但是回复量并未达到峰值，说明许多新用户加入了该事件的讨论，但是由于用户对该事件的讨论热度已经有所下降，因此回复量并未突破最大值。

既然通过参与人数趋势图可以分析出网民们对事件的关注程度及事件的传播程度，那么网民在参与评论时的情绪状态是怎样的，情绪状态分布又是怎样的，针对该问题，本文绘制了参与“网民情感分布趋势图”，如下图5-12所示：

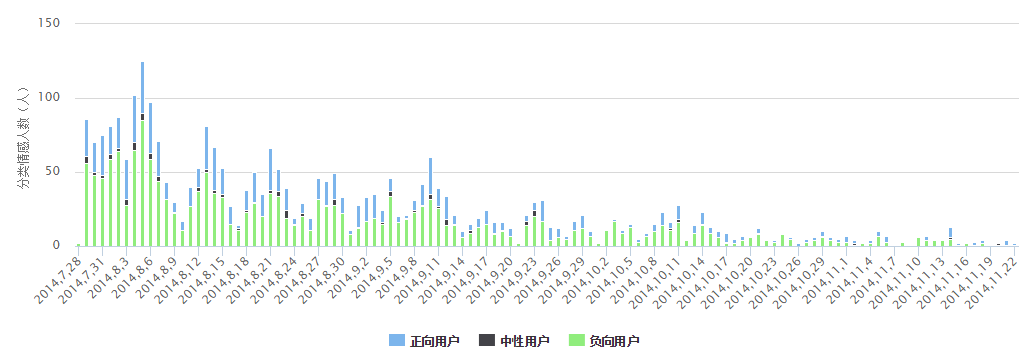


图5-12 网民情感分布趋势图

从图5-12可以看出，在2014-08-02日流量达到了峰值时，虽然参与人数较少，参与讨论的负面用户占大多数，而且相邻的几天里负面用户数均相对比较多。在2014-08-05日，用户参与人数最多且负面也占很大比例，所以通过网民情感分布趋势图可以很明显的发现舆情监控的时间段。

以日流量为基础，对评论文本情感分类图与网民情感分布趋势图的分析与比较，发现网民情感分布趋势图更能准确反映事件的发展态势，通过分析参与讨论的用户的情感分布可以更好发现舆情监控的时间点，可以对舆情进行有效的预警。

## 本章小结

本章在构建的模型基础上，通过抓取数据，对数据进行预处理、指标计算、情感计算及结果数据统计，对情感网络特征进行分析，选取热点帖子并运用社会网络分析方法对情感网络进行分析。

# 结论与展望

## 结论

针对传统情感分类算法不足的问题，本文通过丰富情感词典，选取最优情感特征组合来提高情感分类的准确性，并运用社会网络分析方法，在情感传播模型基础上，从新的视角上对文本情感进行分析，本文主要做了以下几个方面的工作：

#### 构建情感词典

通过现有的情感资源构造了基础情感词典，基础情感词典包括基本情感词典、否定情感词典及程度修饰词典；并利用网络词库构建了网络情感褒贬词典。运用句法分析技术对实体及其评价词进行抽取，运用语义角色标注技术对实体行为进行抽取，在基础情感词典上通过共现特性对抽取的情感单元进行极性判别，制定策略对其情感值进行量化，构建出实体评价及实体行为情感词典。

#### 情感分类研究

针对现有情感分类研究的问题，本文首先在不同的情感词典组合上，运用词典和规则的分类方法对情感进行分类，并对实验结果进行分析，验证了实体情感词典对于情感分类准确率提高的有效性。然后罗列了对情感分类有帮助的特征，并对不同的情感特征进行组合设计，在不同特征组合上运用朴素贝叶斯分类算法、SVM分类算法、kNN分类算法进行对比实验，通过评测得到最优情感特征组合，再次验证了实体情感词典对准确率提高的有效性。

#### 基于社会网络情感分析

在用户与文本内容情感模型基础上，通过抓取数据，对数据进行预处理、指标计算、情感计算及结果数据统计，对情感网络特征进行分析。选取热点帖子，在情感传播模型基础上，构建帖子的情感传播网络，通过gephi对情感网络进行可视化，运用社会网络分析方法对情感网络进行分析，发现网络中的核心节点情绪值对于情感网络情绪值有很大的贡献。通过日流量图、评论文本情感分类图及用户情感分类图对热点帖子进行分析，发现通过用户情感分类图更容易发现事件监控的时间点。

## 展望

本文虽然对基于社会网络的情感分类算法进行了一定研究，但是还存在以下不足之处：

1）本文对于情感词典构建方法进行了改进，但由于自然语言处理技术不是很成熟，实体情感词典的抽取需要大量文本语料，处理速度比较慢，且抽取的实体情感词典中噪声数据比较多，强度量化准确率也有些低，下一步希望可以通过方法改进来降低噪声数据，并提高词典量化的准确率。

2）本文设计特征组合对情感分类算法进行实验，通过结果评估找到了最优特征组合，由于影响情感有很多因素，本文罗列的情感特征还不是很完善，比如人们在评价事物的时候还喜欢运用对比，对于这类特征本文没有进行考虑，下一步希望引入更多的情感相关特征对来提高情感分类的准确率。

3）由于情感传播是动态的，本文通过社会网络分析方法对其进行分析，更多分析的是静态特征，而且分析粒度比较浅，虽然通过“用户情感分类趋势图”可以看出动态特征，下一步希望可以通过其他方法来深化对情感传播的研究。

# 致 谢

时光如梭，岁月如流，转眼间已经在西安交通大学度过了三个春夏秋冬。三年学中发生的点点滴滴，都深深的印在我的脑海中，这样的充满财富的经历使我感到荣幸和自豪。

首先感谢我尊敬的恩师饶元老师，在我三年的研究生生活学习中给予了我最无私的帮助和关心。饶老师是一个在生活上和蔼可亲，在学习上认真严谨的老师。饶老师的务实的工作作风、乐观向上的生活态度深深的感染着我，让我在面对生活、学习中的困难挫折能更加坚定的往前走。每次在我遇到问题困难的时候，饶老师都会耐心认真的帮我解答。在论文工作过程中，饶老师尽心尽力的指导并帮助我。我很感恩能在我人生的道路上遇到一位这么好的导师，衷心感谢饶老师对我的帮助。

同时还要感谢软件学院的各位授课老师，朱利老师、宋永红老师、吴晓军老师、陈建明老师以及学院教育的刘楠老师和秦老师，他们在学习和生活中给了我无尽的帮助。

感谢朝夕共处的SICDP实验室的成员们，有照顾我指导我的师兄师姐们，和共同努力的同一届同学们，以及帮助过我的师弟师妹们，跟大家一起度过的这段美好的时光必定是我人生中最幸福的时光之一。

在漫漫的求学道路中，我非常感谢我的家人，感谢他们多年来对我的关心和激励，为我提供了良好的生活和学习的环境，使我懂得了一个个做人的道理。

最后，感谢所有参加论文评审和答辩的专家和老师，各位专家和老师在百忙之中仍旧不辞辛苦地为我们这些毕业生进行指导和帮助，向各位专家和老师表示由衷的敬意和深深的谢意。

同时本论文得到了国家科技部“火炬计划”（2012GH571817）、2013年国家社科基金重大项目“基于多学科理解的社会网络分析模型研究－－虚拟Web网络空间中的社会网络模型与个体行为机制研究”、2015年度陕西省科技厅陕西省协同创新计划（2015XT-21），2015年西安市科技局产学研协同创新计划 (CXY1514(5))等项目的联合资助，在此一并表示衷心的感谢。

# 参考文献

[1] Baccianella S.,Esuli A.,Sebastiani F.Multi-facet Rating of ProductReviews[C]. In:Proceedings of ECIR’2009，2009: 461-472.

[2] Esuli A.，Sebastiani F.PageRanking WordNet Synsets: An Application to Opinion Mining[C]. In:Proceedings of ACL’ 2007,2007: 424-431.

[3] Gyamfi Y.,Wiebe J.，Mihalcea R，Akkaya C. Integrating Knowledge for Subjectivity SenseLabeling[C]. In: Proceedings of NAACL-HLT’2009，2009: 10-18.

[4] Devitt A，Ahmad K.Sentiment Polarity Identification in Financial News: A Cohesion-based Approach[C]. In: Proceedings of ACL’2007,2007: 984-991.

[5] HowNet[R/OL]．HowNet’s Home Page．http://www.keenage.com. 201 l,12,10.

[6] 柳位平,朱艳辉，栗春亮,向华政,文志强.中文基础情感词词典构建方法研究[J].计算机应用. 2009，29(10): 2875-2877.

[7] 徐琳宏，林鸿飞，潘宇，任惠，陈建美.情感词汇本体的构造[J].情报学报，2008，27(2):180-185.

[8] Pang B，Lee L，Vaithyanathan S. Thumbs up?: Sentiment Classification using Machine Learning Techniques[C]. In: Proceedings of EMNLP’2002，2002. 79-86.

[9] Cui H，Mittal VO，Datar M．Comparative experiments on sentiment classification for online product reviews．In：Gil Y，Mooney ILl，eds．Proc．of the AAAI 2006．Menlo Park：AAAI Press．2006．1265-1270

[10] Kim SM; Hovy E Automatic identification of pro and con reasons in online reviews 2006

[11] Pang B;Lee L Seeing stars:Exploiting class relationships for sentiment categorization with respect to rating scales 2005

[12] Goldberg AB;Zhu X Seeing stars when there aren't many stars:Graph-Based semi-supervised learning for sentiment categorization 2006

[13] Hatfield E, Cacioppo J T, Rapson R L. Primitive emotional contagion[J]. Review of personality and social psychology,1992, 14: 151-177

[14] Hatfield E, Cacioppo J T, Rapson R L. Emotional contagion[J]. Current Directions in Psychological Science, 1993, 2(3):96-99.

[15] Hancock JT, Landrigan C, SilverC.Expressing emotion in text-based communication[C].Proceedings of the SIGCHIconference on Human factors in computing systems,New York,NY,USA,2007: 929-932.

[16]Kamps J，Marx M J，Mokken R J'et a1．Using worduet to measure semantic orientations of adjectives[C]．Proceeding of the 4th International Conference on Language Resources and Evaluation．2004：11 15—1l 18．

[17] 朱嫣岚，闵锦，周雅倩，等．基于HowNet的词汇语义倾向计算[J]．中文信息学报，2006

[18] 李钝，乔保军，曹元大，等．基于语义分析的词汇倾向识别研究[J]．模式识别与人工智能，2008，21(4)：482—487

[19] Kiesler D J. The 1982 Interpersonal Circle: A taxonomy for complementarity in human transactions[J]. Psychological Review,1983, 90(3): 185-214.

[20] Cartwright D, Harary F. Structural balance: a generalization of Heider's theory[J]. Psychological review, 1956, 63(5): 277-293.

[21] Heider F. Attitudes and cognitive organization[J]. The Journal of psychology, 1946, 21(1): 107-112

[22] Adamic L, Buyukkokten O, Adar E. A social network caught in the Web[J]. First Monday, 2003, 8(6).

[23] 李军．中文评论的褒贬义分类试验研究[D]．[硕士学位论文]．北京：清华大学，2008

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

[25] 张伟，刘缙，郭先珍．学生褒贬义词典．北京：中国大百科全书出版社，2004

[26] 杨玲，朱英贵．贬义词词典．成都：四川辞书出版社，2006．

[27] 史继林，朱英贵．褒义词词典．成都：四川辞书出版社，2005．

[28] 同义词词林扩展版．http：//www.ir-lab.org.2011,12,15

[29] 赵妍妍,秦兵,车万翔等.基于句法路径的情感评价单元识别[J].软件学报,2011

[30] Turney P.D. Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews[C]. In: Proceedings of ACL’2002，2002: 417-424.

[31] 王素格,李德玉,魏英杰等.基于同义词的词汇情感倾向判别方法.中文信息学报.2009

[32] M Gamon,A Aue,S Corston-Oliver, E Ringger. Mining Customer Opinions from Free Text[A] .In:

Proceedings of IDA 2005[c]

[33] 杨震,赖英旭,段立娟,李玉鑑.基于上下文重构的短文本情感极性判别研究[期刊论文]-自动化学报 2012(1)

[34] 张林,钱冠群,樊卫国等.轻型评论的情感分析研究?[J].软件学报,2014,(12):2790-2807.

[35] Van Kleef G A. How emotions regulate social life : the emotions as social information (EASI) model[J]. Current Directions in Psychological Science, 2009 18(3): 184-188.

[36] 赵妍妍,秦兵,刘挺等.文本情感分析[J].软件学报,2010,21(8):1834-1848. DOI:10.3724/ SP.J.1001.2010.03832.

[37] 纪雪梅,王芳.在线社交网络用户情感传播研究[C].//2013中国信息经济学会学术年会暨博士生论坛论文集.2013:1-13.

# 附 录

基本褒贬词统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 基本情感词及情感值 | | | | | |
| 僄 -0.5 | 啰唆 -0.5 | 啰嗦 -0.5 | 哙 -0.2 | 婞 -0.6 | 婞直 -0.6 |
| 崒 -0.2 | 弇陋 -0.5 | 惛 -0.6 | 惼 -0.6 | 梼昧 -0.7 | 狯 -0.5 |
| 瘆 -0.8 | 矮 -0.5 | 碍难 -0.5 | 碍眼 -0.4 | 暗 -0.4 | 暗暗 -0.4 |
| 暗沉沉 -0.4 | 暗淡 -0.4 | 暗地 -0.4 | 暗黑 -0.4 | 暗里 -0.4 | 暗昧 -0.4 |
| 暗弱 -0.4 | 暗无天日 -0.8 | 暗中 -0.5 | 暗自 -0.5 | 暗朦 -0.5 | 岸然 -0.6 |
| 肮脏 -0.7 | 昂贵 -0.6 | 凹凸 -0.5 | 凹凸不平 -0.6 | 傲 -0.6 | 傲岸 -0.6 |
| 傲慢 -0.6 | 八面玲珑 -0.6 | 跋扈 -0.7 | 霸道 -0.7 | 霸气 -0.7 | 白 -0.3 |
| 白白 -0.6 | 白搭 -0.6 | 白忙 -0.6 | 白忙活儿 -0.6 | 百孔千疮 -0.7 | 败坏 -0.7 |
| 稗 -0.5 | 板 -0.2 | 板滞 -0.5 | 半半拉拉 -0.6 | 薄 -0.5 | 薄情 -0.5 |
| 薄弱 -0.3 | 薄幸 -0.5 | 保残守缺 -0.6 | 保守 -0.6 | 抱残守缺 -0.6 | 暴 -0.4 |
| 暴烈 -0.6 | 暴虐 -0.7 | 暴躁 -0.6 | 暴戾 -0.7 | 爆炸性 -0.6 | 悲惨 -0.7 |
| 悲观 -0.7 | 悲剧 -0.7 | 悲凉 -0.7 | 卑 -0.6 | 卑鄙 -0.8 | 卑贱 -0.8 |
| 卑劣 -0.8 | 卑陋 -0.8 | 卑怯 -0.8 | 卑俗 -0.8 | 卑琐 -0.8 | 卑微 -0.8 |
| 卑污 -0.8 | 卑下 -0.8 | 卑猥 -0.8 | 背地 -0.6 | 背光 -0.5 | 背后 -0.4 |
| 背悔 -0.4 | 背静 -0.3 | 背理 -0.4 | 背人 -0.4 | 背时 -0.5 | 背阴 -0.4 |
| 被动 -0.3 | 被动式 -0.3 | 被动性 -0.3 | 本本主义 -0.4 | 笨 -0.5 | 笨手笨脚 -0.6 |
| 笨头笨脑 -0.6 | 笨重 -0.5 | 笨拙 -0.5 | 鄙 -0.6 | 鄙贱 -0.6 | 鄙吝 -0.7 |
| 鄙陋 -0.7 | 鄙俗 -0.7 | 鄙俚 -0.7 | 蔽塞 -0.7 | 闭塞 -0.7 | 必修 -0.2 |
| 变化多端 -0.5 | 变化万千 -0.5 | 变化无常 -0.5 | 变化无穷 -0.5 | 变幻不定 -0.5 | 变幻莫测 -0.5 |
| 变幻无常 -0.5 | 变态 -0.7 | 变相 -0.3 | 憋拗 -0.6 | 别扭 -0.5 | 别有用心 -0.6 |
| 冰冷 -0.6 | 病弱 -0.4 | 病态 -0.6 | 病歪歪 -0.6 | 病殃殃 -0.6 | 病恹恹 -0.6 |
| 波谲云诡 -0.6 | 驳杂 -0.6 | 捕风捉影 -0.5 | 不便 -0.3 | 不才 -0.2 | 不成器 -0.7 |
| 不当 -0.4 | 不道德 -0.6 | 不得当 -0.5 | 不得劲 -0.4 | 不得了 -0.4 | 不得体 -0.6 |
| 不等 -0.2 | 不端 -0.6 | 不对 -0.2 | 不法 -0.6 | 不公 -0.5 | 不轨 -0.6 |
| 不和 -0.4 | 不合时令 -0.5 | 不合时宜 -0.5 | 不济 -0.6 | 不济事 -0.6 | 不俭省 -0.5 |
| 不尽 -0.2 | 不绝如缕-0.4 | 不堪入耳-0.5 | 不堪设想 -0.5 | 不堪一击 -0.6 | 不可告人 -0.4 |
| 不可理喻 -0.6 | 不可逆转 -0.6 | 不可收拾 -0.6 | 不可一世 -0.7 | 不可逾越 -0.5 | 不冷不热 -0.5 |
| 不利 -0.5 | 不力 -0.5 | 不良 -0.6 | 不伦不类 -0.6 | 不妙 -0.6 | 不明不白 -0.6 |
| 不平 -0.6 | 不起眼 -0.3 | 不起眼儿 -0.3 | 不巧 -0.4 | 不切实际 -0.6 | 不仁 -0.5 |
| 不仁不义 -0.6 | 不三不四 -0.7 | 不善 -0.4 | 不慎 -0.5 | 不胜 -0.4 | 不痛不痒 -0.4 |
| 不妥 -0.3 | 不为人知 -0.5 | 不祥 -0.6 | 不详 -0.4 | 不像话 -0.5 | 不孝 -0.7 |
| 不肖 -0.6 | 不行 -0.2 | 不幸 -0.7 | 不修边幅 -0.6 | 不要紧 -0.2 | 不宜 -0.3 |
| 不易 -0.1 | 不择手段 -0.7 | 不周 -0.7 | 不着边际 -0.6 | 不足道 -0.3 | 不足为外人道-0.2 |
| 不足为训 -0.2 | 不羁 -0.4 | 才疏学浅 -0.7 | 财迷心窍 -0.8 | 残 -0.8 | 残败 -0.8 |
| 残暴 -0.8 | 残毒 -0.8 | 残酷 -0.8 | 残酷无情 -0.8 | 残虐 -0.8 | 残破 -0.8 |
| 残缺 -0.8 | 残缺不全 -0.8 | 残忍 -0.8 | 残损 -0.8 | 惨 -0.8 | 惨不忍睹 -0.8 |
| 惨淡 -0.8 | 惨毒 -0.8 | 惨绝人寰 -0.8 | 惨苦 -0.8 | 惨厉 -0.8 | 惨烈 -0.8 |
| 惨痛 -0.8 | 惨无人道 -0.8 | 苍白 -0.6 | 苍凉 -0.7 | 苍茫 -0.5 | 操切 -0.7 |
| 糙 -0.6 | 草 -0.9 | 草草 -0.6 | 草荒 -0.6 | 草率 -0.6 | 策略 -0.1 |
| 策略性 -0.1 | 差 -0.4 | 差点儿 -0.4 | 差劲 -0.6 | 豺狼成性 -0.7 | 缠手 -0.6 |
| 颤颤巍巍 -0.4 | 颤颤悠悠 -0.4 | 颤巍巍 -0.4 | 猖 -0.6 | 猖狂 -0.7 | 长篇大论 -0.4 |
| 长线 -0.1 | 超标 -0.6 | 超常 -0.3 | 超然 0.5 | 超重 -0.4 | 朝不保夕 -0.6 |
| 朝不谋夕 -0.6 | 朝三暮四 -0.6 | 潮 0.5 | 吵吵闹闹 -0.4 | 沉闷 -0.4 | 沉痛 -0.6 |
| 沉滞 -0.6 | 陈 -0.2 | 陈腐 -0.7 | 陈旧 -0.3 | 逞性 -0.6 | 逞性子 -0.6 |
| 吃劲 -0.1 | 吃力 -0.4 | 痴 -0.6 | 痴痴 -0.6 | 痴呆 -0.6 | 痴呆呆 -0.6 |
| 痴傻 -0.6 | 痴愚 -0.6 | 迟钝 -0.6 | 侈 -0.7 | 侈靡 -0.8 | 侈糜 -0.8 |
| 赤贫 -0.6 | 冲昏头脑 -0.7 | 丑 -0.8 | 丑恶 -0.8 | 丑陋 -0.8 | 臭 -0.7 |
| 臭哄哄 -0.7 | 臭烘烘 -0.7 | 臭名昭彰 -0.7 | 臭名昭著 -0.7 | 臭味 -0.7 | 初出茅庐 -0.4 |
| 触目惊心 -0.7 | 串秧儿 -0.6 | 蠢 -0.7 | 蠢笨 -0.7 | 蠢头蠢脑 -0.7 | 刺鼻 -0.6 |
| 刺耳 -0.6 | 刺眼 -0.6 | 次 -0.4 | 从严 0.5 | 粗 -0.4 | 粗暴 -0.7 |
| 粗笨 -0.7 | 粗鄙 -0.7 | 粗糙 -0.7 | 粗放 -0.7 | 粗拉 -0.7 | 粗劣 -0.7 |
| 粗陋 -0.7 | 粗鲁 -0.7 | 粗率 -0.7 | 粗蛮 -0.7 | 粗莽 -0.7 | 粗浅 -0.7 |
| 粗涩 -0.7 | 粗手笨脚 -0.7 | 粗疏 -0.7 | 粗俗 -0.7 | 粗线条 0.1 | 粗心 -0.6 |
| 粗野 -0.7 | 粗枝大叶 -0.7 | 粗制滥造 -0.7 | 粗重 -0.7 | 粗拙 -0.7 | 粗犷 -0.7 |
| 促狭 -0.7 | 脆弱 -0.3 | 村气 -0.3 | 村野 -0.2 | 寸草不生 -0.6 | 错 -0.3 |
| 错乱 -0.6 | 错误 -0.6 | 错杂 -0.6 | 错综 -0.6 | 错综复杂 -0.5 | 大 -0.1 |
| 大大咧咧 -0.4 | 大海捞针 -0.6 | 大谬不然 -0.8 | 大手大脚 -0.7 | 大肆 -0.6 | 大摇大摆 -0.6 |
| 大意 -0.4 | 大咧咧 -0.5 | 呆 -0.6 | 呆板 -0.7 | 呆笨 -0.7 | 呆痴 -0.7 |
| 呆呆 -0.7 | 呆钝 -0.7 | 呆气 -0.7 | 呆傻 -0.7 | 呆头呆脑 -0.7 | 歹 -0.8 |
| 歹毒 -0.8 | 殆 -0.6 | 怠惰 -0.7 | 单 -0.4 | 单薄 -0.6 | 单调 -0.6 |
| 单弱 -0.6 | 胆怯 -0.7 | 胆小 -0.7 | 淡 -0.1 | 淡薄 -0.4 | 淡淡 -0.4 |
| 淡漠 -0.5 | 淡然 -0.5 | 诞 -0.4 | 荡 -0.7 | 刀光剑影 -0.6 | 倒胃口 -0.7 |
| 德行 -0.6 | 德性 -0.6 | 得寸进尺 -0.7 | 灯红酒绿 -0.6 | 等因奉此 -0.3 | 低 -0.5 |
| 低卑 -0.8 | 低层 -0.6 | 低档 -0.7 | 低等 -0.8 | 低端 -0.6 | 低级 -0.7 |
| 低贱 -0.8 | 低劣 -0.7 | 低迷 -0.6 | 低能 -0.8 | 低人一等 -0.7 | 低俗 -0.8 |
| 低下 -0.7 | 低效 -0.7 | 低值 -0.6 | 低智 -0.7 | 低质 -0.8 | 滴里嘟噜 -0.5 |
| 敌对 -0.6 | 地下 -0.1 | 颠倒 -0.7 | 颠连 -0.8 | 凋敝 -0.7 | 刁 -0.7 |
| 刁恶 -0.8 | 刁悍 -0.8 | 刁滑 -0.8 | 刁赖 -0.8 | 刁蛮 -0.8 | 刁钻 -0.8 |
| 调皮 -0.3 | 鼎沸 -0.3 | 丢魂 -0.6 | 丢脸 -0.6 | 东倒西歪 -0.7 | 冬烘 -0.7 |
| 动荡 -0.7 | 毒 -0.8 | 毒辣 -0.8 | 独裁 -0.8 | 独断 -0.8 | 短浅 -0.7 |
| 短视 -0.7 | 钝 -0.6 | 多变 -0.5 | 多余 -0.5 | 惰 -0.6 | 惰性 -0.7 |
| 讹 -0.8 | 恶 -0.8 | 恶毒 -0.9 | 恶贯满盈 -0.9 | 恶狠狠 -0.9 | 恶劣 -0.9 |
| 恶心 -0.7 | 恶浊 -0.7 | 饿殍遍野 -0.8 | 耳生 -0.4 | 二手 -0.2 | 二五眼 -0.3 |
| 发狂 -0.5 | 发腻 -0.4 | 乏 -0.5 | 乏味 -0.5 | 翻云覆雨 -0.7 | 繁复 -0.6 |
| 繁乱 -0.7 | 繁难 -0.7 | 繁冗 -0.7 | 繁琐 -0.7 | 繁芜 -0.7 | 繁杂 -0.7 |
| 繁重 -0.7 | 繁缛 -0.7 | 烦 -0.6 | 烦难 -0.8 | 烦冗 -0.8 | 烦琐 -0.8 |
| 烦嚣 -0.8 | 反 -0.4 | 反常 -0.5 | 反反复复 -0.6 | 反面 -0.5 | 反叛 -0.7 |
| 饭桶 -0.4 | 泛 -0.3 | 泛泛 -0.2 | 放诞 -0.7 | 放荡 -0.8 | 放荡不羁 -0.8 |
| 放浪 -0.8 | 放肆 -0.7 | 放纵 -0.6 | 菲 -0.2 | 菲薄 -0.6 | 非 -1 |
| 非法 -0.8 | 非分 -0.7 | 非婚生 -0.8 | 非礼 -0.7 | 非人 -0.8 | 非生产性 -0.4 |
| 非正常 -0.5 | 非正统 -0.6 | 非正义 -0.5 | 废 -0.6 | 废弛 -0.7 | 废旧 -0.5 |
| 废物 -0.8 | 沸沸扬扬 -0.7 | 费 -0.5 | 费工夫 -0.3 | 费劲 -0.5 | 费力 -0.5 |
| 费时 -0.5 | 费事 -0.5 | 纷 -0.4 | 纷繁 -0.7 | 纷乱 -0.7 | 纷扰 -0.7 |
| 纷杂 -0.7 | 封闭 -0.4 | 封闭式 -0.4 | 封闭型 -0.4 | 封建 -0.8 | 锋芒毕露 -0.7 |
| 风吹日晒 -0.6 | 风刀霜剑 -0.7 | 风风火火 -0.5 | 风流 -0.2 | 风骚 -0.2 | 风声鹤唳 -0.6 |
| 风雨飘摇 -0.6 | 疯疯癫癫 -0.7 | 疯狂 -0.6 | 疯癫癫 -0.7 | 否 -1 | 否定 -1 |
| 肤泛 -0.7 | 肤浅 -0.7 | 浮 -0.2 | 浮泛 -0.7 | 浮光掠影 -0.6 | 浮滑 -0.7 |
| 浮漂 -0.7 | 浮浅 -0.7 | 浮躁 -0.7 | 浮噪 -0.7 | 腐败 -0.8 | 腐臭 -0.8 |
| 腐恶 -0.8 | 腐化 -0.8 | 腐旧 -0.8 | 腐烂 -0.8 | 腐朽 -0.8 | 复 -0.1 |

网络褒贬词统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 网络褒贬词及其情感值 | | | | | |
| 小鲜肉 0.8 | 萌萌哒0.8 | 心塞 -0.9 | 逼格 -0.7 | 逗比 -0.6 | 你行你上-0.7 |
| 也是醉了 -0.8 | 不忍直视 -0.8 | NoZuoNoDie -0.8 | 上天台 -0.3 | 伐开心-0.7 | 脑洞大开 0.8 |
| 吓尿了 -0.8 | 大神0.7 | NB 0.8 | 给力 0.7 | 强 0.8 | 弓虽 0.8 |
| 顶 0.8 | 废柴-0.7 | 怪咖 -0.6 | 煞笔-0.8 | 碰瓷-0.7 | 晕 -0.6 |
| 狂顶 0.9 | 拍砖 -0.7 | 养眼 0.8 | sb -0.8 | 汗 -0.7 | 吐 -0.8 |
| 变态bt -0.9 | High 0.8 | 陈水 -0.8 | 大虾 0.8 | 菜鸟 -0.8 | 恐龙 -0.9 |
| 脑残 -0.9 | 衰 -0.8 | 8错 0.8 | 去死 -0.8 | 74 -0.8 | 晕倒 -0.8 |
| 色狼sl -0.8 | 靠kao -0.8 | 妈的MD -0.9 | TMD -0.9 | TNND -0.9 | JJWW -0.8 |
| 叽叽歪歪 -0.8 | 神经病SJB -0.9 | 拍马屁PMP -0.8 | KICK -0.9 | 果酱 0.7 | 抓狂 -0.9 |
| 酷 0.8 | 蛋白质 -0.9 | 结结巴巴 -0.7 | 打酱油 -0.4 | 绿茶婊 -0.8 | 奶茶婊 -0.8 |
| 腐女 -0.8 | 干物女 -0.7 | 布波女 0.8 | 乐活族 0.7 | NONO族0.8 | 剩女 -0.7 |
| 走召弓虽 0.9 | 吼吼 0.8 | 咔咔 0.8 | 西西 0.7 | 酷毙 0.9 | 垃圾 -0.8 |
| 牛逼 0.8 | 没品 -0.8 | yy -0.7 | 型男 0.8 | 渣男 -0.8 | 渣女 -0.8 |
| 油墨 0.8 | 灌水 -0.7 | 纯净水 -0.7 | 轻舞肥羊-0.7 | 青蛙 -0.8 | 马桶文章 -0.7 |
| 飞仔 -0.7 | 太妹 -0.7 | 飞女 -0.7 | 坑 -0.8 | 萝莉 0.8 | 正太 0.8 |
| 欧巴 0.8 | 兄贵 0.7 | 未够班 -0.7 | 逆天 0.7 | 素敌 0.8 | 残念 -0.7 |
| 苦手 -0.8 | 恶趣味 -0.7 | 收声 -0.8 | 白痴BC -0.8 | 奸商JS -0.8 | 流氓LM -0.8 |
| 红茶婊 -0.8 | 龙井婊 -0.8 | 污 -0.7 | 大牛 0.8 | BB -0.8 | 装逼ZB -0.8 |
| 丢雷楼谋 -0.9 | 泪奔 -0.8 | 妈宝男 -0.7 | 月经贴 -0.8 | 西八 -0.8 | 席巴 -0.8 |
| 土锤 -0.8 | 猴赛雷 0.8 | 亮了 0.8 | 哥屋恩 -0.8 | 学屌 -0.8 | 喷子 -0.8 |
| 干货 0.8 | 学婊 -0.7 | 学霸 0.8 | 高富帅 0.8 | 白富美 0.8 | 直男癌 -0.8 |
| 直女癌 -0.8 | 深井冰 -0.9 | 文豪 0.8 | 脏蜜 -0.8 | 火星贴 -0.8 | 爆肝 -0.8 |
| 呆逼 -0.9 | 绿茶屌 -0.8 | 五毛党 -0.7 | 自干五 -0.7 | 速冻矮子-0.7 | 正能量 0.8 |
| 负能量 -0.8 | 喜当爹 -0.7 |  |  |  |  |

程度副词统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 程度副词及其修饰强度 | | | | | |
| 倍加 1.6 | 不得了 1.7 | 不堪 1.6 | 不可开交 1.5 | 不亦乐乎 1.5 | 不折不扣 1.6 |
| 彻头彻尾 1.6 | 充分 1.5 | 到头 1.5 | 地地道道 1.6 | 非常 1.7 | 极 1.8 |
| 极度 1.8 | 极端 1.8 | 极其 1.8 | 极为 1.8 | 截然 1.5 | 尽 1.8 |
| 绝 1.8 | 绝顶 1.8 | 绝对 1.8 | 绝对化 1.8 | 刻骨 1.6 | 酷 1.6 |
| 满 1.5 | 满贯 1.5 | 满心 1.5 | 莫大 1.7 | 奇 1.6 | 入骨 1.6 |
| 十分 1.5 | 十足 1.5 | 死 1.6 | 滔天 1.7 | 痛 1.6 | 透 1.4 |
| 完全 1.5 | 万 1.6 | 万般 1.8 | 万分 1.8 | 万万 1.8 | 无比 1.8 |
| 无度 1.8 | 无可估量 1.8 | 无以伦比 1.8 | 要命 1.7 | 异常 1.7 | 贼 1.6 |
| 至极 1.8 | 卓绝 1.8 | 最为 1.7 | 佼佼 1.6 | 郅 1.7 | 綦 1.8 |
| 齁 1.7 | 最 1.7 | 不过 1.2 | 不少 1.4 | 不胜 1.4 | 惨 1.5 |
| 沉 1.3 | 沉沉 1.3 | 出奇 1.5 | 大为 1.5 | 多 1.4 | 多多 1.4 |
| 多么 1.4 | 分外 1.5 | 格外 1.5 | 好 1.3 | 好不 1.3 | 何等 1.3 |
| 很 1.5 | 坏 1.3 | 可 1.3 | 老 1.3 | 老大 1.3 | 良 1.2 |
| 颇 1.3 | 颇为 1.3 | 甚 1.3 | 实在 1.2 | 太 1.4 | 特 1.4 |
| 特别 1.4 | 尤 1.5 | 尤其 1.5 | 尤为 1.5 | 远 1.4 | 着实 1.3 |
| 曷 1.2 | 碜 1.3 | 大不了 1.2 | 多 1.3 | 更 1.3 | 更加 1.3 |
| 更进一步 1.3 | 更为 1.3 | 还 1.2 | 还要 1.2 | 较 1.2 | 较为 1.2 |
| 进一步 1.2 | 那么 1.1 | 那样 1.1 | 强 1.2 | 益 1.2 | 益发 1.3 |
| 逾 1.3 | 愈 1.3 | 愈发 1.3 | 愈加 1.3 | 愈来愈 1.3 | 愈益 1.3 |
| 远远 1.3 | 越发 1.3 | 越加 1.3 | 越来越 1.3 | 越是 1.3 | 这般 1.1 |
| 这样 1.1 | 足 1.3 | 足足 1.3 | 点点滴滴 1.2 | 怪 1.3 | 好生 1.2 |
| 还 1.2 | 或多或少 1.1 | 略 1.1 | 略微 1.1 | 蛮 1.1 | 稍 1.1 |
| 稍稍 1.1 | 稍微 1.1 | 稍为 1.1 | 稍许 1.1 | 挺 1.2 | 未免 1.2 |
| 相当 1.2 | 些 1.1 | 一点 1.1 | 一点儿 1.1 | 一些 1.1 | 有点 1.1 |
| 有点儿 1.1 | 有些 1.1 | 半点 0.8 | 不大 0.8 | 不怎么 0.8 | 聊 0.7 |
| 轻度 0.7 | 弱 0.7 | 丝毫 0.5 | 微 0.6 | 相对 0.9 | 超 1.8 |
| 超额 1.8 | 超外差 1.8 | 出头 1.5 | 多 1.4 | 浮 1.3 | 过 1.5 |
| 过度 1.8 | 过分 1.8 | 过火 1.8 | 过劲 1.8 | 过头 1.8 | 过于 1.8 |
| 何止 1.7 | 何啻 1.7 | 开外 1.7 | 苦 1.5 | 老 1.5 | 偏 1.6 |
| 强 1.6 | 溢 1.6 | 忒 1.7 | 不 -1 |  |  |

# 攻读学位期间取得的研究成果

1. 冯妮，眭欣阳，员鹏，宋明爽.“武大吉奥杯”第二届全国研究生智慧城市技术与创意设计大赛.创意设计赛优胜奖.教育部学位与研究生教育发展中心.2015.

[2] 饶元，冯妮，宋明爽，员鹏等．数据分析-基于内容与结构的网络舆情分析报告（2015）[M]．北京：电子工业出版社，2016

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．交回学校授予的学位证书；

2．学校可在相关媒体上对作者本人的行为进行通报；

3．本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。

4．本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）： 日期： 年 月 日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．学校可在相关媒体上对本人的失察行为进行通报；

2．本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。

3．本人接受学校按照有关规定做出的任何处理。

指导教师（签名）： 日期： 年 月 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）： 日期： 年 月 日

指导教师（签名）： 日期： 年 月 日

(本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用)