基于领域情感词典的中文微博情感分析

肖 江，丁 星，何荣杰
（江苏科技大学 计算机科学与工程学院，江苏 镇江 212000）

摘要： 为了分析中文微博中海量的情感信息，文中提出了一种中文微博情感分析策略，能够有效分析出微博中的情感倾向。为了能准确分析出某领域微博情感倾向，本文构建了领域情感词典，具有自动识别、扩展等功能，减少了人工标注的繁琐。同时考虑到上下文中情感副词等影响，构建了情感副词词典，更加全面的分析情感倾向。最后通过实验表明本文提出的基于领域情感词典的分析策略有一定的可行性和准确率。

关键词： 微博；情感分析；领域情感词典；分析策略
中图分类号： TP393 **文献标识码：** A **文章编号：** 1674-6236(2015)12-0018-04

Analysis of Chinese micro-blog emotion which based on field of emotional dictionary

XIAO Jiang, DING Xing, HE Rong-jie
(*School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang* 212000, *China*)

Abstract: In order to analyze the massive emotional information in Chinese micro-blog, this article proposes a Chinese micro-blog sentiment analysis strategy, and it can analyze the emotional tendencies effectively in the micro-blog. In order to analyze the emotional tendencies of a field more accurately, this paper builds a kind of emotional dictionary through mood words with automatic identification, extended function, reduces cumbersome manual annotation. Considering the influence of emotional adverbs and expressions in the context, we build the emotional adverbs dictionary and micro-blog expressions which will analyze emotion tendency more comprehensively. In order to analyze the emotional tendencies more comprehensive, and solve the problem of more and more internet words, it constructs network vocabulary classifier to find and mark internet words. Finally, experiments show that the proposed analyze strategy which based on emotional dictionary has certain feasibility and accuracy.

Key words: micro-blog; sentiment analysis; field of emotional dictionary; strategy analysis

微博是微型博客的简称，是一个用于信息分享、信息传播以及信息获取的平台。微博用户可以通过各种计算机终端或移动终端，实现信息的即时分享，通常中文微博内容字数限制在 140 字左右，而这仅仅 140 字却能表达出一个人的情感，观点和对某个事件的态度、看法，大大方便了信息的传递、获取与共享，加快了信息的传播速度，微博已经成为一种新兴的网络媒体。同时，微博给予网络用户更加自由、更便捷的方式来沟通信息、分享心情、表达观点，所以深受广大网民朋友的喜爱，一跃成为国内最为热门的互联网应用之一。

随之而来的是海量的情感文本信息，而这些海量的情感信息是非常宝贵的信息资源，通过分析这些情感文本信息可以得到网民对某个观点、某种社会现象的态度和看法，给政府制定新的政策提供了参考的依据，给企业规划新的发展方向提供了帮助，同时能够给广大消费者购买商品提供参考和依据。

1 相关工作

目前文本情感分析大致分为两种方法[1]：基于情感词典和基于机器学习的情感分析方法。

基于情感词典的方法是对文本信息的情感极性进行分析和计算，得到一个情感极性值。Turney[2]等人通过计算基准词与待估词汇的 SO-PMI 值来分析情感倾向性。在国内，朱嫣岚[3]等人则利用 HowNet 提出的基于语义相似度的方法和基于语义相关场的方法分别计算待估词汇与预先选好的基准词对的相似度，最后得到该词的倾向性。文献[4]提出利用表情图片结合情感词语的方法构建中文情感词典，构建贝叶斯分类器，并且利用熵的概念对语料库进行优化，提高了分类的准确性。

基于机器学习的方法，常用的分类方法有：支持向量机分类法、中心向量分类法、K 近邻算法分类法、感知器分类法、贝叶斯分类法和最大熵分类法等，通过此类分类器识别出该文本的倾向性。Wang[5]等人构建一个 Twitter 情感分析系统，能够实时地对有关总统选举的评论信息进行情感倾向性分析；Jiang[6]等人采用主题相关和无关的方式对微博文本进行情感极性分类，将其分为正向情感和负向情感。在国内，谢丽星[7]等提出了基于层次结构的多策略中文微博情感分析方

收稿日期：2014-09-22 **稿件编号：**201409189

作者简介：肖 江(1974—)，男，辽宁营口人，博士，教授。研究方向：互联网信息安全、多媒体通信系统、图像压缩编码。

法，文中对比了表情符号的规则方法、情感词典的规则方法、基于 SVM 的层次结构多策略方法。

纵观以上分析方法，本文通过构建情感词典的方法来构建微博舆情情感分析系统，从而分析情感倾向，和以往的分析不同，本文构建的情感词典具有领域特性，能自动识别标注领域内情感词并且将新词添加到情感词库中，并且考虑情感副词的影响，综合分析情感倾向。

2　情感语料库的构建

2.1　微博数据集

文中针对新浪微博做分析，利用新浪微博 API 抓取特定领域的微博话题，选取 # 巴西世界杯 #,#iphone6#,#NBA 总决赛 # 作为分析语料，根据分析结果将其分为正面话题、一般话题、负面话题。本文分别选取 3 个特定话题各 2 000 条微博数据作为分析语料，然后对其进行人工标注，将该微博数据标注为正向、负向、中性情感微博，标注的统计结果如表 1 所示。

表 1　微博数据信息

Tab. 1　Micro-blog data information

话题	正向情感微博	负向情感微博	中性情感微博	总数
巴西世界杯	662	659	679	2 000
Iphone6	671	663	666	2 000
NBA 总决赛	659	667	674	2 000

2.2　微博数据的预处理

微博文本有其自身的特点，表达方式多样性，且包含网页链接、图片、英文字母等等，本文需要对其进行预处理，步骤如下：

1）对微博数据进行去重、去噪、标签过滤等操作。

2）去除表情符号，为后面的分词处理做准备。

3）对微博进行分词处理，文中使用中科院 ICTCLAS 系统[8]进行分词与词性标注。

4）对分词后的文本进行粗降维，即将停用词，低频词从文档中去除。

2.3　基准情感词典的构建

基准情感词是指具有非常明显褒贬倾向的词汇，是在某个领域内具有明确褒贬意义的基础词汇。由于中文表达方式的多样性，所以词的褒贬倾向在不同领域并不是完全一致，有些词情感倾向非常明确，但是在不同领域相关度就会降低，词汇的敏感度也会降低，所以在相关度低的领域分析该类词汇的意义并不是太大。

基于以上的考虑，构建基准情感词典，流程如下：

Step1：取特定领域一定数量的微博数据集，进行分析。

Step2：将某领域的微博信息进行预处理，步骤如 2.2 节所示，获得预处理后的微博数据。

Step4：设定高频词汇的阈值 P，并且利用 HowNet 情感词汇集对非情感词汇进行过滤。

Step5：利用 HowNet 正负面词汇集判定高频词汇的正负情感倾向。

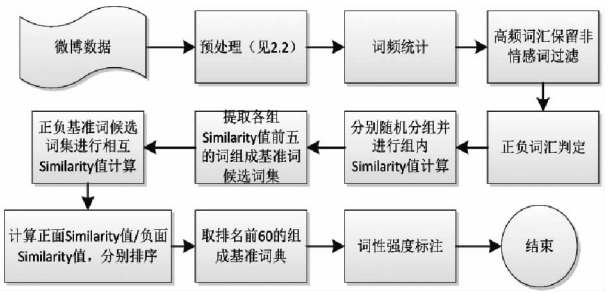

图 1　基准情感词典生成流程

Fig. 1　The generation process of the benchmark emotion dictionary

Step6：在此采用基于知网的语义相似度计算方法来计算词汇之间的 Similarity 值。

Step7–Step10：正负基准词集相互进行相似度计算是为了获得该领域情感倾向更为鲜明的词作为基准情感词集，更准确的识别情感倾向。

Step11：各个基准词代表的情感倾向强度不同，所以需要手动标注基准词的强度值。

2.4　领域相关情感词典的构建

在构建领域情感词基准词集基础上构建领域相关情感词典，再利用基准词集自动识别情感词，实现情感词的自动扩展。本文情感词识别及情感词库构建的整体思路如下图所示。

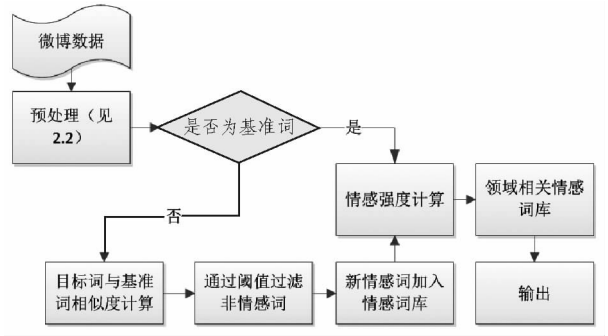

图 2　领域情感词自动识别和标注

Fig. 2　The automatic identification of field of emotional dictionary

如图 2 所示此为领域相关情感词典的构建过程，并且自动将新情感词加入到情感词库，对扩展的词汇进行强度的标注，减少了人工标注过程。目标词与基准词相似度计算，在此采用基于知网的语义相似度计算方法。

2.5　领域情感词情感强度判定

文中计算情感词相似度是基于知网语义相似度计算方法，对领域内情感词进行相似度计算，以均值作为目标词的 Similarity 值。

$$\text{Similarity}(\text{Word},\ \text{WordSet}) = \frac{\sum_{i=1}^{count(WordSet)} Similarity(word, word_i)}{count(WordSet)} \quad (1)$$

目标词的情感倾向判定规则如下，当正向情感的Similarity 值大于负向情感 Similarity 值，则为正向情感词，反之则为负

向情感词。如若正、负向 Similarity 值小于给定的阈值 P 则为非情感词。

$$S(w)=\begin{cases}\text{正向情感词} & \text{Similarity}_{positive}>\text{Similarity}_{negative}\\ \text{负向情感词} & \text{Similarity}_{positive}<\text{Similarity}_{negative}\\ \text{非情感词} & \text{Similarity}_{positive}\ \text{AND}\ \text{Simiarity}_{negative}\end{cases} \quad (2)$$

利用 2.3 和 2.4 构建如表 2 所示的领域情感词典，利用上述计算相似度的方法对词语集 $W=\{w1,w2,w3\cdots\}$分别计算 Similarity 值，剔除不含情感词的词汇，含情感的词汇自动加入情感词库。最后情感词汇集的情感倾向表示为 $W=\{<w1,o1>,<w2,o2>,<w3,o3>,<w4,o4>\cdots\}$，其中 o 表示情感强度值。那么词语的情感倾向值 O(w)可以用如下的公式来计算：

$$O(w)=\sum_{i=1}^{n}Similarity(w,ki)\cdot Ei+\sum_{i=1}^{n}Similarity(w,pi)\cdot Ei \quad (3)$$

其中 $\sum_{i=1}^{n}Similarity(w,ki)\cdot Ei$ 表示了词语 w 和所有褒义词的情感相似度累加，E 表示褒义的强度，k 表示基准褒义词，$\sum_{i=1}^{n}Similarity(w,pi)\cdot Ei$ 表示词语 w 和所有贬义词的情感相似度累加，p 表示基准贬义词。

如下为领域情感词词典，以 #iphone6 为例。

表 2　领域情感词词典

Tab. 2　Field of emotional dictionary

情感词类型	强度值	领域情感词
极强褒义词	+1.0~+1.5	完美、大方、创新、设计美观、人性化、痴迷…
一般褒义词	0~+1.0	不错、合理、还行、灵敏、细腻…
极强贬义词	−1.5~−1.0	讨厌、反感、劣质、易碎、欺骗差劲…
一般贬义词	−1.0~0	不好、费劲、不值当、不舒适、审美疲劳…

2.6　修饰副词的构建

和英文微博相比中文的语法非常丰富，词与词之间不是独立存在的，他们之间还存在着一定的依存的关系，即受前面的修饰词的影响，最常见的就是否定词和副词对情感倾向的影响。目前对于副词和否定词等修饰情感词大多采用的是建立修饰词典，在判断句子的情感倾向时，首先检查情感词前的词语是否属于修饰词，如果是则根据词典中修饰词的强度值计算出情感倾向值。某个词的修饰副词集 $w=\{<q1,<at1,at2..>>,<q2,<at1,at2..>>,<q3,<at1,at2..>>\cdots\}$，其中 q 表示情感词，at 表示修饰副词。那么加入情感副词后的情感值 O(at)用如下公式计算：

$$O(at)=\sum_{i=1}^{n}(qi\cdot\prod_{j=1}^{m}atj) \quad (4)$$

3　微博情感分析策略

前一节完成了情感词典的构建工作，本节在情感词典基础上提出一种分析方法，分析并判定微博情感倾向值。

输入：中文微博句子 S

Step1：对 S 进行预处理，得到预处理后的句子 ST(见 2.2 节步骤 1、2)；

表 3　修饰副词词典

Tab. 3　Adverb dictionary

修饰副词类型	副词	强度值
极量	最、最为	+2.0
高量	更、更加、还、愈、越加…	+1.75
中量	较、比较、较为…	+1.5
低量	略、略微、多少…	+0.5
否定	不是、不好…	−1.0
	……	

Step2：对 ST 进行分句处理，将长句分成多个短句，得到短句集 $ST=\{S1,S2\cdots Si\}$，并进行词性标注（见 2.2 节步骤 3、4)，得到处理后的词语集 $Si=\{<w1,ad1><w2,ad2>\cdots\}$，其中 w 表示情感词语，ad 表示修饰副词；

Step3：计算 Si 的情感倾向值 Oi；

Step3.1：根据情感词典中公式(3)计算情感词的倾向值 $Oi(w)$；

Step3.2：在 $Oi(w)$的基础上根据情感副词词典计算 $Oi(Si)$；

Step4：得到每个短句的情感倾向值 Oi，然后利用加权平均法得到整个句子的情感倾向值 O；

Step5：计算总的情感倾向值，输出总的情感强度 O。

4　实验数据

文中利用爬虫从新浪微博中抓取 3 个话题各 2 000 条微博作为数据集进行分析，经人工筛选后得到正向、负向、中性微博各 650 条。利用文中提出的基于领域词典的情感分析方法进行分析，最终获得各个微博数据的情感强度值，然后通过设定的阈值判定情感倾向。

本实验对比基于基础情感词典、领域情感词典的微博情感分析，对比加入情感副词对整个微博情感分析的影响。实验采用的评价指标有召回率 R、准确率 P 和 F 值，如下公式：

召回率：
$$R=\frac{\sum C_1}{\sum D_1} \quad (5)$$

准确率：
$$P=\frac{\sum C_1}{\sum E_1} \quad (6)$$

F 值：
$$F=\frac{2P\times R}{P+R}\times 100\% \quad (7)$$

其中 Ci 为实验分类为 c 的微博数，Di 为实验总的微博数，Ei 为微博分类为 c 的微博条数，c 表示微博类别(正、负、中性情感类别)。

表 4　不同词典对测试集的分类数值对比

Tab. 4　Comparison of different of classification dictionary

测试集	基础词典			领域			领域+其他词典		
	正面	负面	中性	正面	负面	中性	正面	负面	中性
巴西世界杯	578	590	782	595	623	732	623	643	684
iphone6	626	583	741	636	593	721	671	598	681
NBA 总决赛	548	612	790	597	610	743	620	617	713

由以上数据对比可知，基于领域情感词典的分析系统具有一定的可行性，对微博情感的判定具有一定的准确率。由

表 5　不同词典对测试集的评价指标对比

Tab. 5　Comparison of different evaluating indicator of classification dictionary

测试集	基础词典			领域			领域+其他词典		
	召回率	准确率	F 值	召回率	准确率	F 值	召回率	准确率	F 值
巴西世界杯--正面	0.632	0.681	0.656	0.652	0.730	0.689	0.701	0.802	0.748
巴西世界杯--负面	0.564	0.685	0.619	0.603	0.698	0.647	0.756	0.864	0.806
巴西世界杯--中性	0.586	0.654	0.618	0.610	0.689	0.647	0.720	0.801	0.758
iphone6--正面	0.684	0.731	0.707	0.636	0.701	0.667	0.762	0.692	0.725
iphone6--负面	0.645	0.703	0.673	0.662	0.654	0.658	0.809	0.703	0.752
iphone6--中性	0.655	0.651	0.653	0.675	0.792	0.729	0.792	0.806	0.799

表 1 可以看出，正面情感识别的微博条数基本维持在 600 条左右，没有较明显的倾倒现象。领域情感词典效果较基础词汇好，在基于领域情感词典基础上添加其他词典，识别效果更佳。表 2 是测试集召回率、准确率和 F 值的对比，更加直观的显示出了基于领域词典和其他词典的优势所在，比以往的基于一般词典效果更加显著，无论是召回率、准确率和 *F* 值基本都维持在了 0.7 以上。

实验过程中只使用基础词典识别微博情感倾向，明显识别条数偏低，这是因为基础词典无法准确判定带有鲜明领域特色的情感词汇，例如“大方”，“人性化”等词，如果仅仅依靠基础词汇是无法辨别出这两个词所含情感倾向。领域词典却能解决这一问题，并且在识别的同时将新词加入情感词库，大大节省了人工标注的时间，同时又提高了识别的准确度。

通过以上实验表明本文构建的领域情感词典、情感副词词典结合情感分析策略在分析情感过程中具有一定的可行性，并且取得了一定的效果。

5　结束语

文中为了更好更准确的分析微博情感构建了领域情感词典、情感副词词典，和以往不同的是文中构建的领域情感词典能够准确、快速识别和标注情感词汇的强度，并且在标注的同时将新词和其情感强度值自动加入情感词库中，为后续的情感词汇标注做准备，大大节约了人工再标注的时间，同时又提高了标注的准确度。通过实验又证明了该方法的可行性，并且取得了一定的效果。

在实验过程中发现，由于微博数据的短小，表达方式多种多样，上下文关系衔接不紧密，造成的识别错误占据了一部分，还有现如今越来越多的网络词、变形词的出现同样也导致了实验过程中识别的错误。所以在接下来的工作中会把研究的重点放在上下文的关系、语境和网络词汇发现上，提出一套更加合理更加全面的新的算法。

参考文献：

[1] 周胜臣，瞿文婷. 中文微博情感分析研究综述[J]. 计算机应用与软件，2013，30(3):162–164.
ZHOU Sheng-cheng，QU Wen-ting，et al. The analysis on the Chinese micro-blog emotion[J]. Computer Applications and Software，2013，30(3):162–164.

[2] Turney，Peter D，Littman L. Measuring praise and criticism: Inference of semantic orientation from association. In: ACM Transactions on Information Systems [M]. New York: ACM Press，2003.

[3] 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(1):140–146.
ZHU Yan-lan，MIN Jin，ZHOU Ya-qian，et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Process，2006(1):140–146.

[4] 张珊，于留宝，胡长军，等. 基于表情图片与情感词的中文微博情感分析[J]. 计算机科学，2012，39(11A):146–148.
ZHANG Shan，YU Liu-bao，HU Chang-jun，et al. Sentiment analysis of chinese micro-blog based on emotions and emotional words[J]. Computer Science，2012，39(11A):146–148.

[5] WANG Xiao-long，WEI Fu-ru，LIU Xiao-hua，et al. Topic Sentiment Analysis in Twitter:A Graph-based Hashtag Sentiment Classification Approach [C]//Proceeding of 20th ACM Conference on Information and Knowledge Management (CIKM)，Glasgow，2011.

[6] JIANG Long，YU Mo，ZHOU Ming，et al. 2011. Target-dependent Twitter Sentiment Classifica-tion[C]//ACL 2011.

[7] 谢丽星. 基于SVM的中文微博情感分析的研究 [D]. 北京：清华大学，2011.

[8] ICTCLAC汉语分词系统 [EB/OL]. [2012–07–02] http://ictclas.org/ictclas_download.aspx.