基于词典和规则集的中文微博情感分析

王志涛,於志文,郭 斌,路新江
WANG Zhitao, YU Zhiwen, GUO Bin, LU Xinjiang

1.西北工业大学 计算机学院,西安 710129
2.陕西省嵌入式系统技术重点实验室,西安 710129
1.School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China
2.Shaanxi Key Lab of Embedded System Technology, Xi'an 710129, China

WANG Zhitao, YU Zhiwen, GUO Bin, et al. Sentiment analysis of Chinese micro blog based on lexicon and rule set. Computer Engineering and Applications, 2015, 51(8):218-225.

Abstract: This paper identifies the key points of Chinese micro-blog sentiment analysis: How to recognize the new words in micro-blog and comprehend their sentimental implication automatically? How to take advantage of additional information to assist text analysis? And how to construct a structure method based on micro-blog's semantic characteristic? To solve the first problem, this paper utilizes a method using statistical information and point-wise mutual information to implement new words mining and sentiment comprehension, and builds up a sentimental lexicon of new words from 40 million Sina micro-blogs to expand the existing resources. The sentiment analysis method introduced in the paper, based on rule set and lexicon, can cope with the rest questions. With rules defined on different semantic levels and optimized lexicon, micro-blog's sentiment would be calculated on multi-granularity from words to sentences, adding emoticon information as auxiliary element of the total calculation. This method is conducted on original micro blogs of collected data base, and the result demonstrates its effectiveness.

Key words: micro-blog; new word mining; rule set; sentiment analysis

摘 要: 通过对微博文本的特性分析,提取了中文微博情感分析的关键问题:如何识别微博新词并理解其情感含义?如何利用附加信息辅助文本情感分析?如何结合语言特性构造情感计算方法?针对第一个问题,利用统计信息和点间互信息对新词进行挖掘和情感识别,在40万条新浪微博数据中构建了新情感词词典,用于对已有情感词资源的扩充。对于后两个问题,提出了基于词典和规则集的中文微博情感分析方法。根据微博特性,在不同的语言层次上定义了规则,结合情感词典对微博文本进行了从词语到句子的多粒度情感计算,并以表情符号作为情感计算的辅助元素。通过对采集到的原创微博数据集进行实验,验证了该方法的有效性。

关键词: 微博;新词挖掘;规则集;情感分析
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1308-0187

1 引言

自问世至今,微博作为一个新兴的科技信息产物,在全球已成为一个高度互动的信息转播平台。从国外的Twitter(http://twitter.com/),到中国的新浪微博(http://

基金项目: 国家重点基础研究发展计划(973)(No.2012CB316400);国家自然科学基金(No.61222209,No.61103063);教育部"新世纪优秀人才支持计划"(No.NCET-12-0466);教育部高等学校博士学科点专项科研基金(博导类)(No.20126102110043);陕西省自然科学基础研究计划项目(No.2012JQ8028)。

作者简介: 王志涛(1990—),男,硕士研究生,研究领域为普适计算、个性化推荐、移动互联网;於志文(1977—),男,博士,教授,研究领域为普适计算、情境感知系统、个性化服务、移动互联网;郭斌(1980—),男,博士,副教授,研究领域为普适计算、社会智能、移动社会网络;路新江(1984—),男,博士研究生,研究方向为社会化感知计算、社交网络。
E-mail:wztzenk@126.com

收稿日期: 2013-08-15 **修回日期:** 2013-09-23 **文章编号:** 1002-8331(2015)08-0218-08
CNKI **网络优先出版:** 2013-10-12, http://www.cnki.net/kcms/detail/11.2127.TP.20131012.1630.001.html

weibo.com/)等,短时间里以惊人的速度发展并拥有了大量的用户。

微博庞大的用户群体提供了丰富且海量的观点文本数据,越来越多的学者将目光转移到了对微博的研究上来。而在这些研究中,情感分析成为了一个热点话题。微博之所以受到研究人员的高度关注,是因为其相比较于传统的观点文本有很多不同的特性:

数据海量性:微博用户每天所产生的观点信息数量是惊人的,这就提供给了研究者丰富的数据,而同时如何对海量数据进行规模性的情感分析成为了研究的难点。

实时性:用户可以随时随地将自己的观点通过各种渠道发布到微博。所以,微博的实时性相当高。

独有的文本风格:中文微博被限制在140字,具有简短性;但口语化和新潮性是研究难点。

信息元素多样性:微博信息中不仅仅包含文本信息,还有大量的表情符号、图片内容、网页链接、用户的地理位置信息以及时间信息等,如何充分地利用好这些附加信息也是研究的一个重要突破点。

标注数据缺乏:虽然拥有海量的用户观点数据,但是对这些数据的标注和整理工作比较缺乏,而巨大的数据量也为标注工作带来了困难。

微博文本数据独有的特性给情感分析带来了更广阔的研究空间,同时也带来了更大的挑战。目前,国外已有一些对英文微博的相关工作,而针对中文微博研究还比较空缺。本文通过对微博特点的分析,提取了微博情感分析的几个关键问题:问题(1):如何更好地自动识别和理解微博新兴词汇的含义?问题(2):如何充分地利用一些附加信息(例如表情符号)辅助情感分析?问题(3):如何结合微博语言特性构造在句子颗粒度下的情感计算方法,而不是单一地统计情感词个数?针对以上问题,本文提出了基于新词发现和多层规则集的中文微博情感分析方法。在大规模语料库中结合词频、词内耦合度、邻字集信息熵等统计信息构造了新词发现方法,通过点间互信息对发现的新词进行情感自动标注,得到微博特有情感词典。并在不同语素层次上对情感计算进行规则定义,形成多层次规则集。最后结合情感词典对微博进行情感倾向分析。

2　相关工作

2.1　英文微博相关研究

情感资源是情感分析的基础,发挥着举足轻重的作用,情感资源包括:情感词典,已标注语料库以及情感分析工具等。由于对英文文本的情感分析工作开展得比较早,所以英文情感资源也较为丰富。

现在比较热门的英文情感词典包括SentiWordNet(http://sentiwordnet.isti.cnr.it/),Inquirer(http://www.wjh.harvard.edu/~inquirer/)等。这些词典可以提供词语在不同语境下的主客观性和情感倾向,为情感分析研究提供了良好的基础。其次,英文微博已经拥有了一定规模的标注语料。从最早的人工标注,到后来像Pak和Paroubek[1]利用Twitter文本中用户输入的表情符进行自动标注,英文标注语料的数量也在不断地扩充。

对英文微博情感分析的相关研究都主要是针对Twitter数据的情感分析。Kouloumpis[2]等人探讨了语言特征在Twitter情感分析中的实用性,他们评估了现有的英文词汇资源,以及微博数据中口语化和创造性语言的信息特征。Saif[3]等人提出了一种在情感分析训练集中添加语义作为附加特征的方法。实现了微博语义特征从实体聚类转变为抽象概念,并通过实验证明对语义特征提取和分类的方法要好于情感主题提取的方法。Agarwal[4]等人则是通过强化的方式进行特征混合,而并不考虑语义特征。

Barbosa和Feng[5]利用一些网站所提供的对Twitter信息的情感分析结果作为训练数据,然后选用一些特征,采用二步分类法来对微博数据进行分类。Jiang[6]等人运用五折交叉验证的方法对1 939条Twitter信息做训练和测试,实验表明扩充情感词典和主题相关特征对分类结果有很大的提升。

2.2　中文微博相关研究

中文情感分析由于起步较晚,在情感资源方面还比较缺乏。但是随着研究的逐渐兴起,中文情感资源也在逐年增多。

在情感词典方面比较有代表性的就是知网Hownet情感词典(http://www.keenage.com./)。部分高校也提供了一些情感词汇库,但质量参差不齐。在标注语料方面,中文领域也是在近两年来涌现了一些标注文本。诸如中国中文信息学会信息检索专业委员会举办COAE(Chinese Opinion Analysis Evaluation)提供了一些中文情感标注语料(http://www.ir-china.org.cn/Information.html)。就权威性的中文微博情感分析语料而言,目前还十分匮乏,去年计算机学会所举办的中文信息技术专委会微博情感分析评测(NLP&CC2012)提供了腾讯微博的数据。

针对中文微博的情感研究目前还处于起步阶段。谢丽星[7]等人提出了基于层次结构的多策略方法对新浪微博数据展开情感分析研究,并在特征提取时采用了主题相关特征,实验结果显示,使用主题相关的特征后所获得的最高准确率由66.467% 提升到了67.283%。

3　微博情感词典构建

情感资源尤其是情感词典在情感分析中发挥着重要作用,一些研究深入展开了对词典构建的工作[8-9]。就目前中文情感词典的现状,很有必要对其进行整合和优

化。并且,本文针对微博领域提出了新词发现和新词情感词典构建的方法。

3.1 微博情感词典组成

图1是本文构建的微博情感词典资源。其中基础情感词典是对知网情感词典和台湾大学的正、负面词典(http://www.datatang.com/)的整合和优化。微博新词词典则是通过统计信息进行新词挖掘并通过点间互信息对新词进行情感识别。对于新词词典的构建方法将在3.2和3.3节展开介绍。修饰词表包括了程度副词词表和否定词表。否定词表选取了汉语中常用的19个否定词;程度副词词表则是根据汉语言研究把220个程度副词划分6个不同等级中。而表情词表的构建是通过在语料库统计表情符出现频率并排序,筛选出了87个高频表情词,根据其所表达的情感强度,并定义其情感权值分布在[−2,2]的区间内。

图1 情感词典资源组成

3.2 基于统计的新词挖掘

微博语料的情感计算应该针对微博平台上一些新兴的特有词汇和短语进行识别和理解。本文提出了基于统计信息的新词识别。

3.2.1 字串频数

在文本中判断一个字串是否是新词,首先要考虑其在文本中出现的频率,当一个字串重复出现的次数大于某个值时就认为它有可能构成新词。所以给出如下定义:

定义1 在文本域 D 所有可能出现的字串集合 $W=\{w_1, w_2, \cdots, w_i, \cdots, w_n\}$ 中,定义字串频数 $N(w_i)$ 表示 $w_i \in W$ 在文本域 D 中出现的次数。

$N(w_i)$ 作为新词成词的标准之一。

3.2.2 内部耦合度

一个字串能否成词与这一字串内部的紧密程度也有很大的关系。通过内部耦合度(Inside Coupling)来衡量这种紧密程度,定义如下:

定义2 对字串 w 划分为两个分字串所有的可能组合 $\{(w_{11}, w_{12}),(w_{21}, w_{22}), \cdots,(w_{i1}, w_{i2}), \cdots,(w_{n1}, w_{n2})\}$(例"巨蟹座"所有可能组合:{("巨蟹","座"),("巨","蟹座")})通过如下公式计算:

$$IC(w)=\frac{1}{n}\sum_{i=1}^{n}\frac{P(w)}{P(w_{i1})\times P(w_{i2})} \tag{1}$$

得到的 $IC(w)$ 称为字串 w 的内部耦合度。其中 $P(w)$ 表示字串 w 在文本域 D 出现概率,通过公式:

$$P(w)=\frac{N(w)}{N_D} \tag{2}$$

计算,$N(w)$ 表示 w 字串在文本域 D 中出现的次数,N_D 表示文本域的总字数。

3.2.3 邻字集信息熵

然而,仅对字串内部考察还不够,还需要考察字串面向外部的运用能力。邻字集信息熵能很好地衡量一个字串的外部运用能力。邻字集表示一个词两侧可能出现的单一字的集合;而信息熵可以度量一个事件的不确定性,如果不确定性越大所需要了解的信息量就越大,信息熵也就越高。如果一个字串确实是一个潜在的词,那么其在文本中的运用就更加多样,出现在其两侧的邻字就更加的不确定,这种不确定性就可以通过信息熵计算。

定义3 字串 w 在文本域 D 中所有可能出现在 w 左(右)侧的单字的集合 $C=\{c_1, c_2, \cdots, c_i, \cdots, c_n\}$ 称为 w 的左(右)邻字集。对 C 通过公式:

$$IE(w)=-\sum_{i=1}^{n}\frac{n_i}{n}\mathrm{lb}\frac{n_i}{n} \tag{3}$$

计算得到的 $IE(w)$ 称为 w 的左(右)邻字集的信息熵。其中,n_i 表示 c_i 作为 w 的左(右)邻字出现的次数,n 表示邻字集 C 中的所有字作为 w 的左(右)邻字出现的次数之和。再通过:

$$IE_{\min}(w)=\min\{IE_{\text{left}}(w), IE_{\text{right}}(w)\} \tag{4}$$

求得左、右邻字集信息熵 $IE_{\text{left}}(w)$ 和 $IE_{\text{right}}(w)$ 的较小值作为判断字串 w 能否以一个词灵活运用在文本中的标准。

3.2.4 基于统计信息的新词发现

通过以上的定义,就可以将 $N(w), IC(w), IE_{\min}(w)$ 作为判断字串 w 是否成词的标准。将语料文本整体看为一个长字串,设定词语的成词长度为 $length=7$,再对字串集合进行统计计算,并对三个统计信息设定阈值 $Threshold(N), Threshold(IC), Threshold(IE)$。判断是否在阈值范围内,如果三个参数均满足条件,则认为其构成词语。最后在构成的词语集合中,比对已有词典,若该词未登录(未查找到)则作为新词输出。并构建新词集合 W_{new},可直接应用在分词过程中。

3.3 新词情感识别

新词被挖掘出来后,还需要对这些词进行情感识别,从而构建新情感词典。

对于词语而言,点间互信息(Point-wise Mutual Information,PMI)可以计算词语间的语义相似度[9]。两个词语 w_1 和 w_2 点间互信息计算公式为:

$$PMI(w_1, w_2)=\mathrm{lb}\frac{p(w_1, w_2)}{p(w_1)\times p(w_2)} \tag{5}$$

其中 $p(w_1, w_2)$ 表示 w_1, w_2 共现概率,$p(w_1)$ 和 $p(w_2)$ 分别表示 w_1, w_2 单独出现概率。

可以将PMI公式应用到词语的情感倾向判断上:w_1

是新发现的词，而 w_2 是已知情感倾向的基础词，通过 $PMI(w_1, w_2)$ 来计算两个词的语义相似度，如果相似度高则认为两个词的情感倾向相同，反之则异然。对应微博语料，$p(w_1, w_2)$ 就代表 w_1, w_2 共同出现在同一条微博的概率，$p(w_1)$ 和 $p(w_2)$ 分别表示 w_1, w_2 在微博文本中的出现概率。

但是仅仅判断一对词的相似度是不够的，需要进行多词考察。本文通过统计基础情感词典在微博文本域中的词频，筛选出最高的正、负基础情感词各30个，并结合热门的网络正、负面情感词各5个以及使用频率较高的表情词(正负各5个)构造了正面情感词集合 W_{P} 和负面情感词集合 W_{N}，用于考察新词 w 与多个基础情感词的语义相似度。并对PMI公式加以改进和利用，得出了对新词 w 情感倾向的判别公式：

$$Sen(w)=\sum_{w_p\in W_{\mathrm{P}}}P(w,w_p)-\sum_{w_n\in W_{\mathrm{N}}}P(w,w_n) \tag{6}$$

其中 $Sen(w)$ 表示新词 w 的情感倾向值，而公式右边被减部分表示新词 w 与 W_{P} 中的每个词点间互信息数值的和，减数部分则表示 w 与 W_{N} 中的词语的点间互信息数值的和。一般的，$Sen(w)$ 大于0，w 为正向情感词；等于0，w 为中性情感词；小于0，w 为负向情感词。通过以上方法就实现了对新词的情感自动识别与标注。最终构建了微博新词情感词典。

4　微博情感分析与计算

上一章讲述的新词发现及其词典构造有效地解决了问题(1)。而本章将给出问题(2)和问题(3)的解决方案，即基于规则集和微博情感词典，结合附加信息和语言特点，对微博进行情感分析和计算。

4.1　情感分析流程

图2是本文基于规则集的微博情感分析流程。在文本预处理过程中，需要剔除一些无用字串，例如链接信息(url)等。同时，通过标点符号将每一条微博划分成多个简短的句子，并对表情符号进行了提取。微博数据中，表情符号一般以例如“[哈哈]”这样的字串形式出现。对于一条微博而言，可能含有多个表情符，可以通过正则表达式匹配“[]”内字串提取这些表情符的文字信息，并构建表情集合 $F=\{f_1,f_2,\cdots,f_n\}$，通过查询表情词典得到每个表情 f_i 的情感值并进行加和，求得这条微博对应的表情信息的情感值 $SEN_F=\sum_{i=1}^{n}sen_{f_i}$。在预处理之后，微博被分割成若干个分句，通过规则对这些分句在句型和句式方面进行分析。之后，利用分词和句法分析技术得到每一分句内词语间的依赖关系对，构建词语多元组，基于多元组规则对其进行情感分析。最后，结合预处理环节中提取的表情符集合 F 和表情符规则，以及之前几步得到的句型、句式和词语多元组的分析结果，对整条微博进行综合情感计算。

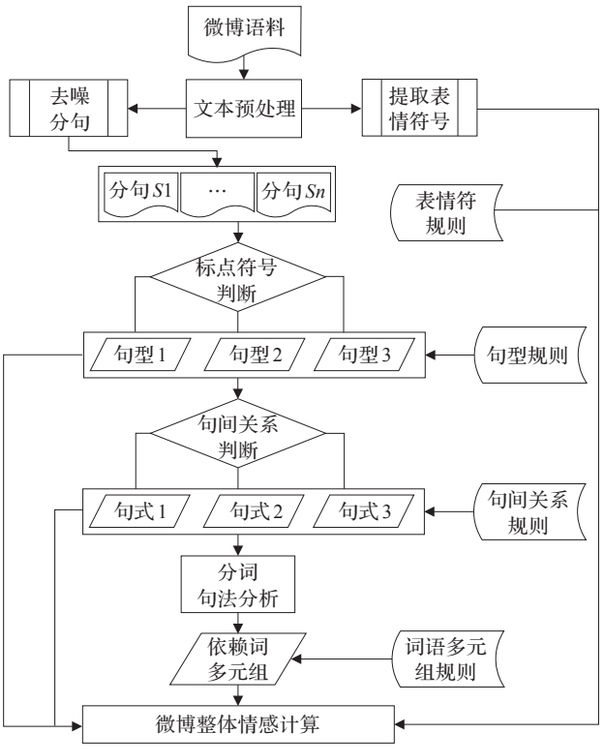

图2　微博情感分析流程

4.2　基于规则集的微博情感分析

如图2所示，本文结合微博文本的语言特性，对不同的语言元素定义了多层次的情感分析规则集：句型规则，句间关系规则、词语多元组规则和表情符规则。下面就结合情感分析的流程对这四个规则进行阐述。

4.2.1　句型分析规则

预处理后微博文本被划分为多个简短分句组成的集合 $S=\{s_1,s_2,\cdots,s_i,\cdots,s_n\}$。在这一语素层次下定义规则文本进行句型分析，这里考虑的是对一个完整句子通过标点符号的句型判断。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)。一个复句用变量 s_{ij} 表示，其中含有 $s_i,s_{i+1},\cdots,s_j$ 共 $j-i+1$ 个分句，各分句多以逗号隔开。定义参数 ST_{ij} 表示复句 s_{ij} 的类型，利用以下规则对 ST_{ij} 赋值：

规则 1.1　如果 s_{ij} 以“!”结尾，则 $ST_{ij}=1.5$，表示句型为感叹句。

规则 1.2　如果 s_{ij} 以“?”结尾且含有反问标志词(如，“难道”)或者句尾不以“?”结尾但含有有反问标志词(有时微博用户可能在操作时忘记加标点，只对反问句的情况考虑)，则 $ST_{ij}=-1$，表示句型为反问句。

规则 1.3　如果 s_{ij} 以“?”结尾且不含有反问标志词，则 $ST_{ij}=0$，表示句型为疑问句。

规则 1.4　如果 s_{ij} 以其他标点结尾，则 $ST_{ij}=1$，表示句型为陈述句。

4.2.2　句间关系分析规则

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，对五类主要关系进行了分析：转折关系、递进关系、假设关系、因果关系和并列关系。这里定义参数 SR_i 表示句式关系，并利用以下规则进行赋值：

转折关系规则　转折关系中往往会发生情感反转，转折前的分句意思会弱化，突出转折后分句情感，所以进行了如下规则定义：

规则 2.1.1　若复句 s_{ij} 中某一分句 s_k 出现单一转折后接词（如，“但是”、“但”、“却”、“可是”），则 $SR_i, SR_{i+1}, \cdots, SR_{k-1}=0; SR_k, SR_{k+1}, \cdots, SR_j=1$。

规则 2.1.2　若复句 s_{ij} 中出现成对转折标志词（如，“虽…但”）且转折后接词出现在分句 s_k 中，则 $SR_i, SR_{i+1}, \cdots, SR_{k-1}=0; SR_k, SR_{k+1}, \cdots, SR_j=1$。

规则 2.1.3　若复句 s_{ij} 中某一分句 s_k 出现单一转折前接词（如，“虽然”），则 $SR_i, SR_{i+1}, \cdots, SR_{k-1}=1; SR_k, SR_{k+1}, \cdots, SR_j=0$。

递进关系规则　递进关系中分句语义的情感将会逐渐增强，所以做出了如下定义：

规则 2.2　若复句 s_{ij} 中出现递进关系标志词（如，“更加”，“更有甚者”），则 $SR_i=1, SR_{i+1}=1.5, \cdots, SR_j=1+0.5\times(j-i)$。

假设关系规则　假设关系往往是对现实状况的设想，而前提条件在语言表达中起到了更重要的作用，有时需要弱化假设句的后半分；而如果出现否定假设，往往是对现实的相反感情假设，所以定义如下：

规则 2.3.1　若复句 s_{ij} 中存在假设关系，分句 s_k 出现关系后接词（如“那么”），则 $SR_i, SR_{i+1}, \cdots, SR_{k-1}=1; SR_k, SR_{k+1}, \cdots, SR_j=0.5$。

规则 2.3.2　若复句 s_{ij} 中存在否定假设关系（如“如果不”），分句 s_k 出现关系后接词，则 $SR_i, SR_{i+1}, \cdots, SR_{k-1}=-1; SR_k, SR_{k+1}, \cdots, SR_j=-0.5$（需要说明的是存在否定假设时，这一规则不会影响后面要提到的词语多元组规则中的否定关系。在之后的分析中否定假设里的否定词还是会作为词语间的否定关系进行词语多元组分析的）。

因果、并列和一般关系规则　对于这一类关系句子情感上的变化不是很大，所以定义如下：

规则 2.4　若复句 s_{ij} 中出现因果、并列标志词，或不存在标志词，则 $SR_i, SR_{i+1}, \cdots, SR_j=1$。

4.2.3　词语多元组分析规则

之后，对每一分句进行了分词和句法分析，进行在词语层面的细粒度情感分析。分词过程中引入了新词词典 W_{new}，进一步提高了分词的准确性，并对得到的分词集合进行句法分析。

本文采用了 Stanford Parser（http://nlp.stanford.edu/software/lex-parser.shtml）句法分析器，提取出了分词集合中主要的依赖关系对。本文所应用到的主要关系对，如表1所示。

表1　词语主要依赖关系

名称	关系	例句	依赖词对
nsubj	宾语和主语	我打扫	（打扫，我）
dobj	宾语和谓语	打扫房间	（打扫，房间）
amod	形容词修饰	房间破旧	（房间，破旧）
advmod	副词修饰	非常辛苦	（辛苦，非常）
comod	并列关系	辛苦而费劲地	（辛苦，费劲）
neg	否定修饰	不高兴	（高兴，不）

利用这些依赖关系，对分句 s_i 构建一个多元组 $Vi=<topic, v, a, adv, neg>$，分量分别表示：主题词（主语和宾语），动词，形容词，修饰副词和否定词。因为情感词典中的词几乎都是名词、动词和形容词，所以，情感词只有可能在 $topic, v, a$ 这三个分量中出现，通过查询情感词典可以找出它们对应的情感极性，分别用 $sen_{topic}, sen_v, sen_a$ 表示。而 adv 可以通过查找程度副词词表确定其数值用 d_{adv} 表示，neg 则是通过否定词表确定个数。本文对多元组情感分析的进行了以下的规则定义：

程度修饰规则　一般副词的修饰主要集中在分量 a 上，定义程度修饰标志 L_a，初始值为1，考虑微博语言简化性，修饰词至多两个，规则如下：

规则 3.1.1　若对于分量 a 存在advmod关系对为 (a, adv)，则 $L_a=d_{adv}$。

规则 3.1.2　若分句中存在comod关系对 $(adv1, adv2)$，且分量 a 与 $adv1, adv2$ 均存在advmod关系，则计算 $L_a=d_{adv1}\times d_{adv2}$。

否定关系规则　一般对情感词的否定会出现在对分量 v 和 a 上，定义否定标志 NO_v 和 NO_a，初始值为1，这里只考虑至多两个neg关系，否定关系规则：规则3.2定义如表2。

表2　否定关系规则

v的neg关系个数	a的neg关系个数	v和a依赖关系	NO_v	NO_a
1个	0个	无	−1	1
0个	1个	无	1	−1
1个	0个	有	−1	−1
0个	1个	有	1	−1
1个	1个	无	−1	−1
1个	1个	有	−1	1
2个	0个	无	1	1
0个	2个	无	1	1
2个	0个	有	1	1
0个	2个	有	1	1

当分量 v 和分量 a 存在依赖关系,且 v 出现neg关系时,会对 a 的否定标志造成极性反转的影响。这是因为动词对形容词的主导作用,而形容词对动词不具有主导作用,所以在 a 存在neg关系,即使两者存在依赖关系也不会对 v 有影响。

主题词相关性规则 微博的情感计算要考虑主题词相关性,与主题词无关的情感词会给情感计算带来干扰。主题主要通过 *topic* 分量体现,所以考察分量 v 和分量 a 与 *topic*(主语和宾语)依赖关系,定义主题相关性参数为 T_v 和 T_a,通过如下规则进行主题相关性分析:

规则 3.3.1 若分量 v(或分量 a)与 *topic*(主语和宾语)存在依赖关系 $T_v=1$(或 $T_a=1$)。

规则 3.3.2 若分量 v(或分量 a)与 *topic*(主语和宾语)不存在依赖关系 $T_v=0$(或 $T_a=0$)。

4.3 微博综合情感计算

通过4.2节分别在复句,分句以及词语多元组三个层次上得到了情感分词的相关参数和标志,本节给出利用这些参数的微博情感综合计算方法。情感计算的颗粒度由小到大:

(1)词语多元组:对 V_i 情感值 $SEN(V_i)$:

$$SEN(V_i)=sen_{topic}+SEN(v)+SEN(a) \tag{7}$$

其中 $sen_{topic}, SEN(v), SEN(a)$ 分别表示词语多元组 V_i 的三个分量 *topic*, v, a 的情感倾向值。而 sen_{topic} 通过在查询情感词典即可得到,$SEN(v)$ 和 $SEN(a)$ 则需通过下边的公式计算得到:

$$SEN(v)=NO_v\times T_v\times sen_v \tag{8}$$

$$SEN(a)=NO_a\times T_a\times L_a\times sen_a \tag{9}$$

其中,sen_v 和 sen_a 查询词典可得,NO_v 和 NO_a 则是分量 v 和 a 的否定标志,T_v 和 T_a 则是它们的主题相关性标志,而 L_a 是分量 a 独有的程度修饰标志,以上参数都可以通过前述规则得到。

(2)分句:对 s_i 情感值 $SEN(s_i)$:

$$SEN(s_i)=SEN(V_i)\times SR_i \tag{10}$$

其中 $SEN(V_i)$ 表示分句 s_i 所对应的分解的得到的词语多元组 V_i 的情感倾向值。SR_i 表示分句 s_i 在句间关系上的考量指数,通过4.2节所述规则可得。

(3)复句:对 s_{ij} 情感值 $SEN(s_{ij})$:

$$SEN(s_{ij})=ST_{ij}\cdot\sum_{k=i}^{j}SEN(s_k) \tag{11}$$

是对复句中每个可能的分句的情感倾向值进行加和,并考虑复句的整体句式,句式参数 ST_{ij} 通过规则得到。

(4)微博文本:情感值 SEN_{text}:

$$SEN_{\text{text}}=\sum SEN(s_{ij}) \tag{12}$$

对于微博的文本信息的情感倾向值通过对所有复句的情感值进行加和得到。

(5)整条微博(文本和表情符):情感值 SEN:

$$SEN=\begin{cases}SEN_{\text{text}}+SEN_F(|SEN_{\text{text}}|<1)\\SEN_{\text{text}}(|SEN_{\text{text}}|\geqslant 1)\end{cases} \tag{13}$$

这步计算就是图2所示的表情符规则的应用,其表述如下:

规则 4 当通过纯文本分析所得到的情感计算的绝对值小于1(情感倾向不明显)时,引入表情集合情感值与纯文本情感值进行加和。

一般的,当 SEN 取值为0时,表示微博情感为中性;大于0时,情感倾向为正面;小于0时为负面。

5 实验结果

5.1 新词挖掘实验

本文随机选取了40万条通过利用新浪API采集到的原创微博数据,这些数据时间跨度从2012年9月23日到2013年1月30日。对这些数据进行了分析,按照内容对数据进行了分类,并对每一分类数据进行了条数和字数信息统计,如表3。其中,媒体是指图片、视频以及一些网页链接信息等。可以看出,若是文字与媒体信息一起出现时,字数会有一定的减少,而对纯媒体信息而言,微博采集数据的文本内容会出现一些"分享图片"或是描述媒体信息的一些字符串,所以表中会有字数的统计信息。

表3　新词挖掘数据集

类型	微博条数	字数	平均字数
纯文字	206 446	13 150 610	64
文字和媒体	150 839	7 447 961	49
纯媒体	42 715	233 187	5
总计	400 000	20 831 758	52

由于对新词挖掘不会有贡献,对纯媒体信息进行了过滤去除。最终对357 285条原创微博数据利用第3章提到的方法进行了新词挖掘实验。在实验中,首先需要确定三个统计量的阈值作为新词成词的衡量标准。在大量的文本数据中,新词的成词阈值可以参考已知词汇的统计量取值。通过计算采集数据中的已知常用词(1 000个)的三个统计量参数,并求得三个参数的最小值的近似值作为新词挖掘实验中的参数阈值,词频阈值 $Threshold(N)$ 为10,内部耦合度阈值 $Threshold(IC)$ 为2.5,邻字集信息熵阈值 $Threshold(IE)$ 为0.75。在此文本数据中,计算满足三个阈值并且未在已知词典中登陆的字串的统计信息。通过计算,检测到了2 847个未登录字串,这些字串按照词频从高到低排列下来。表4是按词频排名前10的未登录字串及其各项统计信息的得分,其中 N 表示词频,IC 表示内部耦合度,IE 表示邻字集信息熵。

在新词挖掘实验的过程中会出现一些固定词组的搭配,这样的字串也因为无法在已有词典中匹配而被认

为是新词,这些固定词组搭配的字串称为“噪音”。通过利用常用的搭配词表来进行降噪,最终得到了如表4的结果。从表中可以看出,降噪后仍然会出现一些词组搭配字串,但这些字串表现了一定的微博特性,例如,“请关注”、“分享自”等都是微博所独有的固定字串,这也体现了对微博语言特性的挖掘。之后人工校对进行了再去噪,最终得到新词1 087个。

表4　未登录字串Top10

排名	字串	*N*	*IC*	*IE*
1	微博	9 885	147.94	2.68
2	请关注	3 934	46.25	1.82
3	分享自	3 207	6.94	1.82
4	给力	2 729	3.18	3.05
5	在评论中	2 715	7.05	1.03
6	有木有	2 092	4.92	3.53
7	包邮	1 383	14.37	2.38
8	卖萌	1 266	68.85	3.57
9	尼玛	1 207	92.43	2.98
10	坑爹	622	79.35	3.75

得到新词后对其进行了自动情感识别之后,又进行了人工情感校对,以这1 087个词构造了新情感词典,新情感词典的统计信息如表5。

表5　新情感词词典

情感倾向	词数	例子
正面	97	“正能量”
中性	749	“穿越剧”
负面	241	“吐槽”
总计	1 087	

这里中性词占得比例很大是因为网络大量涌现的新鲜名词,这些名词是对一些新事物或是事件的概括,而大部分都是情感倾向很中立的一些事物和事件。负面情感词多于正面情感词原因可能是因为现在的网民更习惯于运用一些负面新词或是讽刺一些负面事件。将本实验得到的情感词典应用在了情感分析实验中。

5.2　情感分析实验

5.2.1　实验数据

在情感分析实验中,首先在采集到的微博数据集(共1 000万条微博,包括原创和非原创)中提取了其中的西安用户,共53 661个(注册信息的地点显示是西安)。之后,提取了这些用户所发的全部原创微博共2 763条。经过分析,发现这些用户大部分是对微博进行转发,本文提取的微博数据都是原创数据,所以会出现比例很小的情况。选择西安用户的原因是:在对数据进行人工标注时,这些数据信息更贴近标注人员(西安大学生)的生活,他们更容易做出判断。选取了三名志愿者对这2 763条微博,进行了讨论式的人工标注,并按照新词挖掘实验的方式进行了分类,数据如表6。

纯媒体信息无法通过字符文本判断其情感倾向,所以均标注为中性。可以看出,实验数据中正面倾向所占比例较大,负面倾向的数据较少,倾向性分布并不均匀。这样的不均匀与地域、时间以及事件都有着紧密的关系。例如,在这一时间段中该地区有一些负面事件,该地区用户在这段时间所发微博可能倾向性更加偏于负面。这是真实情况的反映,所以并没有对这种不平衡再做处理。

表6　微博情感分析实验数据

	纯文字	有媒体	纯媒体	总计
正面	887	413	0	1 300
中性	588	374	73	1 035
负面	282	146	0	428

5.2.2　实验性能评估指标

本实验基于第4章的方法对每条微博进行情感倾向性分析,将自动分析的结果与手工标注的结果对比。这里只考虑极性方向分析是否正确。

评价指标采用目前广泛接受的正确率(*Precision*)和召回率(*Recall*),选用综合度量指标 *F* 值(*F*)作为 *Precision* 和 *Recall* 两者的调和平均数来衡量[10]评估分析的准确率。它们的计算公式如下所示:

$$Precision=\frac{\text{判断正确的该类别微博数}}{\text{判断为该类别的微博数}} \tag{14}$$

$$Recall=\frac{\text{判断正确的该类别微博数}}{\text{应判断为该类别的数目}} \tag{15}$$

$$F=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{16}$$

5.2.3　实验结果与分析

为了验证之前构建出的新情感词典的作用和本文提出的基于规则集的情感分析方法的有效性,分别通过表7中的四种方法对数据进行了实验,并对结果进行了指标评价。

表7　微博情感分析实验分析

实验方法	新情感词典		*Precision*	*Recall*	*F*
基于词典的传统统计法	未加入	正面	0.621	0.426	0.505
		负面	0.477	0.231	0.311
		中性	0.633	0.647	0.619
		平均	0.577	0.435	0.478
基于词典的传统统计法	加入	正面	0.613	0.478	0.537
		负面	0.498	0.223	0.308
		中性	0.653	0.603	0.627
		平均	0.588	0.435	0.491
基于规则集计算	未加入	正面	0.711	0.834	0.768
		负面	0.546	0.558	0.552
		中性	0.765	0.596	0.670
		平均	0.674	0.663	0.663
基于规则集计算	加入	正面	0.717	0.840	0.773
		负面	0.556	0.568	0.563
		中性	0.777	0.604	0.680
		平均	0.683	0.671	0.672

前两次实验中用到的分析方法就是简单地比较微

博句子中的情感词正负个数,在一定程度上考虑否定词和副词的修饰,这也是较为传统的微博情感分析方法。

这里,考察四种实验方法对不同情感倾向微博的分析下的 F 值变化。并进行了以下分析:

(1)正负情感微博分析上 F 值在同一方法中的差异较大,是由于数据分类上的不平衡造成的。数据分析评估指标上的不均衡是由于数据集的不均匀造成的。

(2)在未使用规则集的方法中,三类微博中对中性的情感分析 F 值更高一些,且加入新词典和引入规则集后提高幅度为6.1%在三类微博中最小。这是因为,中性微博中很可能不会出现情感词,而情感词典就足以满足对这类无明显情感倾向微博的判断需求。

(3)在前两次实验中,负面微博的情感判断都低于50%,处于不可接受范围,在引入了基于规则集的后两次实验 F 值分别达到了55.2%和56.3%,有20%左右的提高,并达到了可接受范围。而加入情感词典后,F 值的增加并不明显。正面微博的提高效果最为明显,由最初的50.5%提高到了77.3%,可见文本提出的基于规则集的方法在正面微博的情感分析中表现更好。

(4)本文提出的基于规则集的方法较传统的统计法而言,多层次的考虑微博在不同语素中所要表达的情感倾向。通过对句子结构的语法分析提取句子中的关键因素,并考虑微博句子中汉语的语言习惯。而传统的统计方法,在句子较为复杂的环境下效果一般,而微博文本的表达既有简单易懂的句子,同时也有很多含有了转折、讽刺以及类比等句子手法,在这样较复杂的句子环境中,基于规则的方法能够更加还原微博本意。

(5)虽然总体的 F 值67.2%,达到了较好的水平。但是对于负面微博的 F 值56.3%而言,还是处于偏低的水平,其效果还有待于提高。且实验中依赖于分词和句法分析过程,其性能的优劣也直接影响到了实验结果,所以应进一步减少文本处理过程对情感分析的影响。

实验表明,加入新词词典和使用基于规则集的方法的都使情感分析的效果有了提高。基于规则集的方法的使用对效果的提高更加明显,而加入词典后提高效果相对较小。在综合考虑准确率和召回率的评价标准下本文基于新词词典和规则集的情感分析方法具有一定的有效性。当然,本文提出的方法在一些方面的表现还略显不足,在以后的工作中会有进一步的改进。

6　结束语

面向微博的情感分析,新词的挖掘和情感识别是其中的基础和关键问题。本文提出了基于统计信息的新词挖掘和基于点间互信息的情感识别的方法,构建了新情感词词典。利用这一词典,提出了基于规则集和词典的微博情感分析方法,并通过实验验证了此方法的有效性。

中文微博的情感分析研究工作才刚刚起步,还有许多研究需要完成,未来的工作首先是改进现有方法,并实现更加细粒度的情感分析。汉语文化博大精深,表达方式多种多样,各领域都有一些特定的用法,所以针对微博领域接下来会考虑利用已有的工作为其构建一种语义情感模型。微博的独有实时性和用户的情境信息也是可以利用的。进一步的工作还可以集中在情感分析与情境信息的结合应用上。

参考文献:

[1] Pak A, Paroubek P.Twitter as a corpus for sentiment analysis and opinion mining[C]//Proceedings of the International Conference on Language Resources and Evaluation, Valletta, Malta, 2010.

[2] Kouloumpis E, Wilson T, Moore J.Twitter sentiment analysis: the good the bad and the OMG![C]//Proceedings of the International AAAI Conference on Weblogs and Social Media, North America, 2011.

[3] Saif H, He Yulan, Alani H.Semantic sentiment analysis of twitter[C]//Proceedings of the 11th International Conference on the Semantic Web-Volume Part I (ISWC'12), Boston, USA, 2012: 508-524.

[4] Agarwal A, Xie B, Vovsha I, et al.Sentiment analysis of twitter data[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2011: 30-38.

[5] Barbosa L, Feng Junlan.Robust sentiment detection on twitter from biased and noisy data[C]//Proceedings of Coling, 2010: 36-44.

[6] Jiang Long, Yu Mo, Zhou Ming, et al.Target-dependent twitter sentiment classification[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, USA, 2011: 151-160.

[7] 谢丽星,周明,孙茂松,等.基于层次结构的多策略中文微博情感分析和特征抽取[J].中文信息学报,2012,26(1):73-83.

[8] Ku L W, Lo Y S, Chen H H.Using polarity scores of words forsentence-level opinion extraction[C]//Proceedings of the 6th NT-CIR-6 Workshop Meeting, 2007.

[9] Kaji N, Kitsuregawa M.Building lexicon for sentiment analysis from massive collection of HTML documents[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.

[10] Li Guangxia, Hoi S C H, Chang Kuiyu, et al.Micro-blogging sentiment detection by collaborative online learning[C]//Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 2010: 893-898.