DOI:10.16644/j.cnki.cn33-1094/tp.2017.03.003

基于情感词典方法的情感倾向性分析*

杨 奎，段琼瑾

(南华大学计算机科学与技术学院，湖南 衡阳 421000)

摘 要：针对网络舆情中观点的获取问题，提出了基于情感词典的情感倾向性分析方法。介绍了情感词的基本概念，给出了基于HowNet概念词典通过计算词汇相似度构建情感字典的方法，探讨了不同类型情感词对文本情感的影响程度并设计了情感得分策略。根据得分挖掘人们对舆情的褒贬态度，从而准确的分析文本的情感走向。

关键词：舆情分析；情感词典；情感倾向性分析；词汇相似度

中图分类号：TP302.7 **文献标志码**：A **文章编号**：1006-8228(2017)03-10-03

Analysis of emotional tendency based on emotional dictionary

Yang Kui, Duan Qiongjin

(*School of Computer Science and Technology, University of South China, Hengyang, Hunan 421000, China*)

Abstract: Aiming at the problem of acquisition of viewpoints in the network public opinion, this paper puts forward the method of emotional tendency analysis based on emotional dictionary. This paper introduces the basic concept of emotional words, gives the method of constructing emotional dictionary by calculating lexical similarity based on HowNet concept dictionary, and discusses the influence degree of different types of emotional words on text emotion and designs emotional score strategy. According to the scores the people's attitude of praise or censure to the public opinion is mined, so as to accurately analyze the emotional direction of the text.

Key words: public opinion analysis; emotional dictionary; emotional tendencies analysis; lexical similarity

0 引言

随着互联网的迅速发展，网络成为了一个巨大的民意聚集地。微博、新闻、论坛等，都成为人们发表言论和观点的场所。因为网络上言论自由度很高，人们对待事物各持己见，想要得到一个正确的观点，便需要对大量的信息进行分析。舆情信息量不断增大，要了解当前社会的舆情走向变得更加困难，网络舆情分析系统便应运而生。

中文语义倾向性分析的研究方法可以分为两类：基于规则和基于统计。基于规则是依据知识库和规则进行文本倾向性分析，比如简单的基于情感词典，统计文本中的正、负面情感词汇的词频；基于统计是将倾向性分析看成是文本对正、负情感倾向性的分类问题，可以使用朴素贝叶斯、SVM等统计学习的方法进行倾向性分析。

本文采用基于情感词典[1]的方法，对舆情信息进行观点挖掘，获取人们对事物的褒贬态度。

1 基于情感词典的文本倾向性分析框架

情感分析是指挖掘文本表达的观点，识别主体对某客体的评价是褒还是贬，根据褒贬态度进行倾向性研究。本文利用HowNet[2]进行语义分析，求出得分，从而来评判文本的褒贬态度。得分结果若为正数，则认为文本表达的是“正面情感”；得分结果若为负数，则认为文本表达的是“负面情感”；得分结果若为0，文本表达的是“中性情感”。具体分析框架如图1所示。

收稿日期：2017-02-14

***基金项目**：2016年度湖南省大学生研究性学习和创新性实验计划项目(湘教通[2016]283号-307)

作者简介：杨奎(1996-)，男，湖南长沙人，本科，主要研究方向：自然语言处理。

通讯作者：段琼瑾(1973-)，男，湖南衡阳人，研究生，实验师，主要研究方向：实验教学与管理。

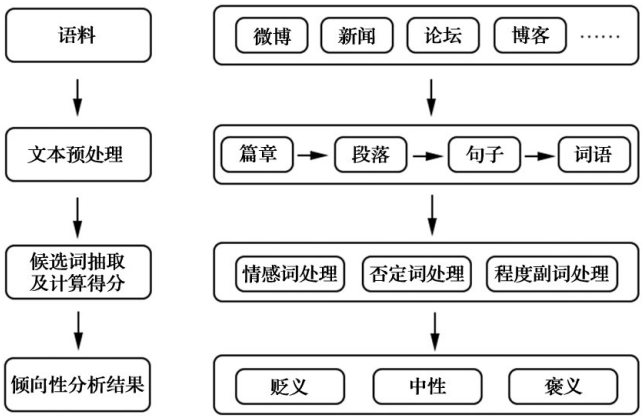

图1 文本倾向性分析框架图

2 情感词典构建

2.1 情感词

情感词,是主体对于某一客体表示内在评价的词语,带有强烈的情感色彩。情感词有两个属性:极性和强度。

根据极性,情感词典可以分解为褒义词典和贬义词典。例如“漂亮”、“善良”为褒义词,表达正面情感;“可恶”、“卑鄙”为贬义词,表达负面情感。褒义词的极性设置为1,贬义词的极性设置为–1。

强度,表示情感强弱。例如:①我讨厌你;②我恨你。明显句子②比句子①表现出更多的不喜欢的意思,即句子②的情感强度更强。用数字1~9表示情感强度,数值越大,情感强度越强。

HowNet整理的部分情感词词表如表1。

表1 HowNet部分情感词

类别	基础情感词示例	个数
正向情感词	爱戴 表扬 称赏 崇敬 爱不忍释	836
正向评价词	蔼然 安定 按时 宝贵 安然无事	3730
负向情感词	哀愁 懊恨 抱憾 悲哀 悲痛欲绝	1254
负向评价词	碍眼 暗黑 肮脏 傲慢 卑鄙无耻	3116

2.2 程度副词

程度副词本身没有情感倾向性,但是它能够增强或减弱情感强度。如果不考虑程度副词对于情感词的修饰作用,虽然不一定改变情感倾向性的结果,但一定会改变情感倾向度的结果。HowNet整理的部分程度副词词表如表2。

2.3 否定词

否定词本身也没有情感倾向性,但是它能够改变情感的极性。HowNet没有整理否定词,于是本文在情感词典中添加了19个否定词:

毋 非 莫 弗 勿 否 别 無 休 不 不要 不曾 无 没 没有 难以 未 未曾 未必

表2 HowNet部分程度副词

量级	权值	副词示例	个数
极其\extreme/最\most	2	倍加 备至 极度 绝对 十足	69
很\very	1.75	不过 不少 不胜 分外 特别	42
较\more	1.25	更加 较为 那般 益发 尤甚	37
稍\–ish	0.5	略微 略加 稍许 有点 相当	29
欠\insufficiently	–0.5	半点 不大 不甚 轻度 丝毫	12
超\over	–0.75	超额 过度 过分 过火 开外	30

2.4 基于HowNet构建情感词典

刘群[3]、葛斌[4]等人对词语的相似度计算做了深入的研究,采用“整体相似度等于部分相似度加权平均”的做法。

相似度的计算公式为:

$$Sim(W_1,W_2)=\frac{\alpha}{\mathrm{Dis}(W_1,W_2)+\alpha} \tag{1}$$

$\mathrm{Sim}(W_1,W_2)$为相似度,$\mathrm{Dis}(W_1,W_2)$为词语距离,α是一个正的可调节的参数,其值不大于1。

对于两个汉语词语W_1和W_2,如果W_1有n个义项(概念):$S_{11},S_{12},\cdots,S_{1n}$,$W_2$有m个义项:$S_{21},S_{22},\cdots,S_{2m}$,则$W_1$和$W_2$的相似度为:

$$Sim(W_1,W_2)=\max_{i=1,\cdots,n,\,j=1,\cdots,m} Sim(S_{1i},S_{2j}) \tag{2}$$

由此,就把两个词语之间的相似度问题归结到两个概念之间的相似度问题。概念的相似度计算分为虚词概念的相似度计算和实词概念的相似度计算。

虚词和实词是不能互相替换的,所以,虚词概念和实词概念的相似度为0。虚词概念的相似度计算非常简单,只需要计算其对应的句法义原或关系义原之间的相似度即可。实词概念的相似度计算比较复杂,公式为:

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \tag{3}$$

根据HowNet中总结的义原,与从语料中提取的候选词采用上述公式计算词汇相似度,根据相似度大小筛选出新情感词加入情感词典。参考陈晓东等[5]提出的微博领域情感词获取过程,本文的情感词获取流程图如图2所示。

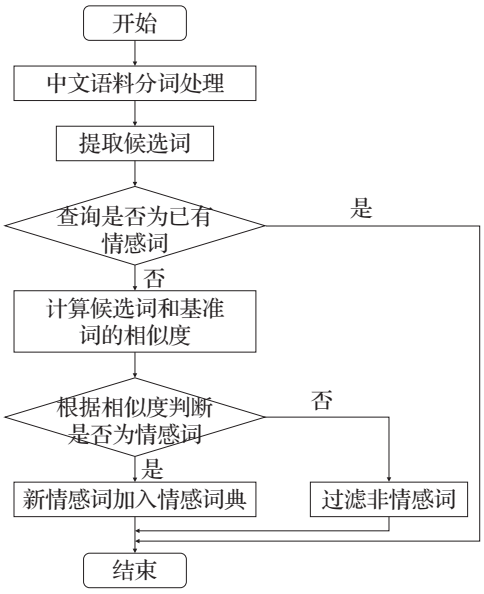

图2 情感词获取流程图

3 设计计分策略

娄德成等人[6]认为计算语义极性时,如果忽略上下文极性,可能会使得极性倾向判断错误或者极性倾向虽然判断正确,但是强度不够准确。因此,加入了程度副词和否定词以提高语义极性分析的准确率。本文在其计算方法基础上稍作改进,情感计算策略如下:

⑴ 假设某语句中有情感词word,Polarity和Strength分别为它的极性和强度,Polarity的值为1或-1。该语句的情感得分为:

ContextScore (word)=Polarity (word)*Strength (word)

⑵ 如果语句中有程度副词intensifier修饰情感词,Weight为它的权值。该语句的情感得分将受程度副词的影响,得分为:

ContextScore (word)=ContextScore (word)*Weight (intensifier)

⑶ 如果该语句中有否定词privative修饰情感词,否定词能改变语句的极性,此时的情感得分为:

ContextScore (word)=-ContextScore (word)/2

这里要除以2,是因为否定词能够减弱情感强度。比如:对于“喜欢”这个词,得分为+2,其反向语义应该是“讨厌”,得分为-2。而不喜欢并不代表讨厌,其情感强度减弱了,因此获得的得分为-1。

⑷ 当一个句子中同时出现否定词和程度副词时,由于否定词和程度副词相对位置的不同,会引起情感的不同。例如:①我很不高兴;②我不很高兴。前者表达的是一种很强烈的负面情感,后者则表达的是一种较弱的正面情感。因此,如果否定词在程度副词之前,起到的是减弱情感强度的作用,得分为:

ContextScore=ContextScore (word)*Weight (intensifier)/2

如果否定词在程度副词之后,则起到的是逆向情感的作用,得分为:

ContextScore=-ContextScore (word)*Weight (intensifier)

4 倾向性分析过程

4.1 中文分词

根据文本的粒度不同,情感分析的任务可以分为“篇章级”、“句子级”和“词语级”。首先,要进行中文分词。本文采用开源的HanLP汉语言处理包中的CRF分词方法对语料进行分词处理。

算法设计的最大分析对象为篇章Document。中文分词过程的伪代码描述如下。

⑴ 将文档以换行符“\r\n”进行分割得到段落顺序表Paragraphs。

⑵ 从左到右扫描Paragraphs的每一个段落执行下述操作,直至遍历完顺序表:

① 将段落中的句号、分号、问号、感叹号等划分句意的符号作为分隔符,对段落进行切割得到句子顺序表Sentences。

② 从左到右扫描Sentences中的每一个句子执行下述操作,直至遍历完顺序表:

a. 通过CRF分词法对句子进行分词得到词语顺序表Words。

b. 基于情感词典识别情感词,计算得分。

4.2 计算得分

分词结束后,对每个句子逐个计算得分。将处理后得到的单词,依次与预先构建好的情感词表逐个查找,若能找到,则是情感词,记录该情感词的位置,表示为(词语位置, 情感词, 得分)。然后以每个情感词为基准,向前依次寻找程度副词、否定词,并作相应分值计算。随后对分句中每个情感词的得分作求和运算。每个句子的得分再求和即得到文章的情感得分,根据分值可分析情感倾向性。

例:这顿饭太好吃了,太美味了!

对上述句子进行中文分词,结果如下。

[这/rzv, 顿饭/nz, 太/d, 好吃/a, 了/ule, ,/w, 太/d, 美味/n, 了/ule, ! /w]

得分详细计算过程如下:

⑴ 从左到右扫描词语集合,得到情感词“好吃”,得分+2,记录当前词语的位置,表示为(3,“好吃”,2)。

⑵ 向前寻找程度副词或否定词，直至遇到分隔符结束。找到程度副词“太”，该程度副词的权值为1.75，计算得分2*1.75=2.5。更新情感词得分为：(3,“好吃”,2.5)。

⑶ 在位置3向后继续扫描词语集合，找到情感词“美味”，记录为：(7,“美味”,2)。

⑷ 在位置7向前寻找程度副词和否定词，当到达位置3时停止寻找，找到程度副词“太”，更新情感词得分为：(7,“美味”,2.5)。

⑸ 再从位置7向后扫描，直至句子结束。

⑹ 最终求得该句子的情感得分为：2.5+2.5=5，句子表达“正面情感”。

5 结束语

本文研究了基于情感词典对文本情感进行倾向性分析的方法，重点阐述了情感词典的构建过程和情感得分的设计策略，主要解决从文本中获取人们对事物褒贬态度的问题。基于词典的方法主要是使用词典中词语之间的词义联系挖掘情感词，所以获取的情感词语的规模非常可观，从而提高了准确率。但影响文本情感的因素很多，不能仅凭借本文方法就能准确分析所有文本的情感倾向。在未来的工作中，还可以做出以下改进：提高情感词典的覆盖率；结合上下文语境或结尾标点符号等改进计分策略。

参考文献(References):

[1] 李婷婷，姬东鸿.基于SVM和CRF多特征组合的微博情感分析[J].计算机应用研究，2015,04:978-981.

[2] HowNet[R].HowNet's Home Page.http://www.keenage.com.

[3] 刘群，李素建.基于知网的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会，2002:59-76

[4] 葛斌，李芳芳，郭丝路，汤大权.基于知网的词汇语义相似度计算方法研究[J].计算机应用研究，2010.9:3329-3333

[5] 陈晓东.基于情感词典的中文微博情感倾向分析研究[D].华中科技大学，2012.

[6] 娄德成，姚天昉.汉语句子语义极性分析和观点抽取方法的研究[J].计算机应用，2006.11:2622-2625

(上接第9页)

分别使用DSL(数字用户线路)服务和卫星遥感服务。日本人均耕地仅有0.7亩，但通过农业信息网络、农业数据库系统、精准农业、生物信息、电子商务等现代信息技术，实现了播种、控制与质量安全及农产品物流等方面的智慧化，农业安全生产和农产品流通效率位居世界前列。目前我国智慧农业呈现良好发展势头，但整体上还属于现代农业发展的新理念、新模式和新业态，处于概念导入期和产业链逐步形成阶段，在关键技术环节方面和制度机制建设层面还面临支撑不足的问题，且缺乏统一、明确的顶层规划，资源共享困难和重复建设现象突出，一定程度上滞后于信息化整体发展水平[6]。

4 结束语

农业是国家发展进步的基础，通过不断的改革创新，我国农业迎来了前所未有的发展机遇。设施农业能有效弥补自然环境的缺陷，提升农业生产效率，是我国农业现代化转型的必由之路。目前，各类设施温室大棚不断推广普及，但在大规模应用方面仍有很多问题需要解决，如数据安全、系统维护、偏远恶劣环境下的电源问题等，只有将这些问题都合理解决后，我国的农业物联网技术才能真正走向成熟。

参考文献(References):

[1] 杨利春.大棚温室环境监控系统的设计[J].湖南科技学院，2010.31(4).

[2] 赵忠彪.nRF401在温室大棚监控系统中的应用研究[J].工业控制计算机，2008(3):21.

[3] 托普农业物联网整体解决方案[J].中国农业物联网，2012.

[4] 奚强.温室大棚监测控制系统研究[D].中国农业大学，2005.

[5] 梁建华，肖申平.基于NRF04I的无线监控探头的设计[J].机电产品开发与创，2006.1:19

[6] 陈世清.对称经济学[M].中国时代经济出版社，2010.