文章编号:1671-9352(2012)03-0047-04 DOI:10.6040/j.issn.1671-9352.2012.03.010

一种基于极性词典的情感分析方法

张成功[1,2],刘培玉[1,2*],朱振方[1,2],方明[1,2]
(1.山东师范大学信信息科学与工程学院,山东 济南 250014;
2.山东省分布式计算机软件新技术重点实验室,山东 济南 250014)

摘要: 极性词典是文本情感分析和倾向性分析的基础。本文构建了一个全面、高效的极性词典,包括基础词典、领域词典、网络词词典以及修饰词词典,深入研究了修饰词对极性词的影响,将极性词与修饰词组合成极性短语作为极性计算的基本单元,提出了一种基于极性词典的情感分析方法。实验结果表明,利用本文构建的词典进行倾向性分析效果不错。

关键词: 极性词典;修饰词;极性短语;情感分析
中图分类号: TP301 **文献标志码:** A

A sentiment analysis method based on a polarity lexicon

ZHANG Cheng-gong[1,2], LIU Pei-yu[1,2*], ZHU Zhen-fang[1,2], FANG Ming[1,2]
(1. School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong, China;
2. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan 250014, Shandong, China)

Abstract: A polarity lexicon is the foundation of sentiment analysis and orientation analysis. An overall and effective polarity lexicon was constructed, including base lexicon, domain lexicon, network lexicon and modifier lexicon. The influence of modifiers to polarity words was studied. A sentiment analysis method based on a polarity lexicon was proposed, in which the modifiers and polarity words were combined into polarity phrases, and the phrase was used as the basic unit to compute the polarity of sentences and texts. Experimental results showed that the effect of orientation analysis using polarity lexicon raised in this paper was good.

Key words: polarity lexicon; modifiers; polarity phrases; sentiment analysis

0 引言

随着Web 2.0时代的到来,用户更多地参与到互联网的建设中,网上出现了大量的表达人们情感倾向性的主观性文本,这些海量信息仅仅靠人工很难处理,于是文本情感分析应运而生。文本情感分析就是对带有情感色彩的词语、句子以及文本进行分析、处理、归纳和处置的过程[1],不管是词语、句子,还是文本,对其情感倾向性起关键作用的都是构成句子和文本的极性词语。因此,构建一个完善的、高效率的极性词典,是进行文本倾向性分析和文本情感分析的基础。

目前,国内外众多学者在极性词典的构建方面做了大量工作。在国外,情感分析主要采用的词典资源是General Inquirer(http://wjh.harvard.edu/

收稿日期:2011-11-30; 网络出版时间:2012-03-20 10:58
网络出版地址:http://www.cnki.net/kcms/detail/37.1389.N.20120320.1058.013.html
基金项目:国家自然科学基金资助项目(60873247);山东省自然科学基金重点项目(ZR2009GZ007);山东省高新自主创新专项工程项目(2008ZZ28)
作者简介:张成功(1987-),男,硕士研究生,主要研究信息过滤,情感分析等. Email:zcg870108@163.com
* 通讯作者:刘培玉(1960-),男,教授,博士生导师,主要研究方向计算机网络信息安全、网络系统规划、网络信息资源开发和软件开发技术. Email:liupy@sdnu.edu.cn

~inquirer)。该词典手工标注了词语情感倾向信息,是一部比较完善的极性词典,为英文文本的倾向性分析打下了基础。国内构建极性词典的思路有基于统计的方法和基于语义词典或知识系统的方法。基于统计的方法利用词语之间的共现信息计算词语与基准词之间的相似度,以此来判别词语的语义倾向。文献[2]通过计算给定短语与excellent和poor之间的互信息得到短语的语义倾向性;文献[3]通过计算候选情感词语和种子词间的点互信息之和来得到其情感倾向。基于统计的方法计算简单,但是需要依赖于大规模的语料,且无法区别极性词语的领域差异。基于语义词典或知识系统的方法利用已有的语义词典或知识系统中的词语建立初始词典,采用一定的方法来扩展词典。文献[3]利用HowNet情感词集构造初始词典,构建了3种模型来扩展词典,但是没有考虑到极性词的领域差异以及修饰词对极性词的影响;文献[4]基于HowNet和NTUSD两种方法扩展极性词典,并构建了修饰词词典,但没有对修饰词对极性词的影响做深入的研究;文献[5]将舆情事件分为8个类别,分类别进行极性词的扩充;文献[6]希望由领域词典逐步扩展为通用词典,但却均忽视了基础词典的重要性;文献[7]利用大连理工大学徐琳宏和林鸿飞等构建的情感本体词汇和HowNet评价词构建情感倾向性词典,并将褒义词和贬义词强度分别标注为1和-1,但是没有考虑到修饰词的作用。

1 极性词典的构建

针对目前极性词典建设过程中存在的以上问题,本文构建了一个包括基础词典、领域词典、网络词词典以及修饰词词典的极性词典。基础词典是一部跨领域的、包含了较全面极性词语的词典;领域词典是从不同领域对词典的扩充;网络词词典主要针对有情感倾向的网络词语;修饰词词典包含了对极性词极性起重要作用的修饰词。

1.1 基础词典

大部分的极性词没有领域差异,因此,构建一个跨领域的基础极性词典是十分重要的。构建的基础词典主要利用了《知网》①提供的情感词语、《汉语褒贬义用法词典》②提供的褒贬义词语以及搜狗实验室提供的互联网词库SougouW③。SougouW统计了SOUGOU搜索引擎一个月内出现的15万条高频词及其词频信息。构建极性词典的流程如图1所示。

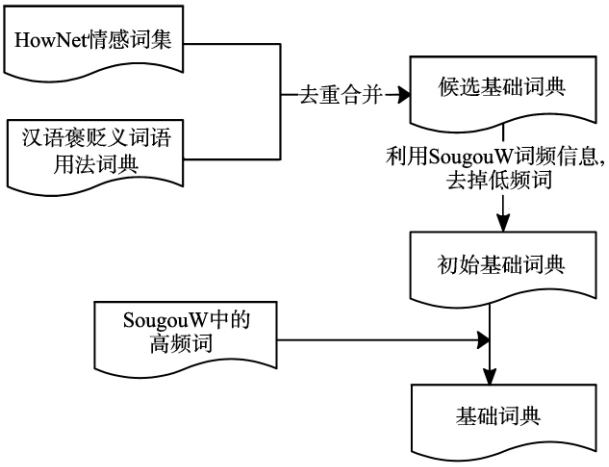

图1 基础词典构建流程图

Fig. 1 Flow diagram of base lexicon

经过整理共得到正面极性词4 324个,负面极性词4 066个。

1.2 领域词典

有些极性词只用在特殊的领域或者在其他领域使用非常少,如"涨停"、"利空"等只用在股票评论中;还有一些极性词在不同的领域、修饰不同的特征时会表现出不同的极性,例如"轻薄"在修饰手机时是褒义的,而在修饰人的时候就是贬义的。因此对每个领域建立一个领域词典来区分词语的领域差异是必要的。本文利用谭松波提供的酒店评论语料构建了一个"酒店评论"领域的领域词典,共搜集到正面领域词汇374个,负面324个。

1.3 网络词词典

网络的日益开放使得人们越来越多地参与到互联网的建设中来,随之而来的是网络上出现了大量的网络词语,如"给力"、"鸭梨山大"等。这些网络词语在一段时间内被大量使用来表达人们的情感,因此,建立一个专门的网络词词典对文本情感分析有重要的作用。

1.4 修饰词词典

许多研究者对修饰词都进行了研究[3-4],但对于修饰词和极性词的相对位置不同导致对极性词产生的不同影响都未做深入研究。本文构建了一个包括否定副词和程度副词的修饰词词典,根据文献[3]中对否定副词范围的界定,选取"白、别、不、不要、没有"等31个否定副词,采用蔺璜[9]对程度副词的分类,选取80多个程度副词,如表1所示。

① http://www.keenage.com

② http://www.docin.com/p-53443925.html

③ http://www.sogou.com/labs/dl/w.html

表 1　程度副词分类表

Table 1　Classification table of degree adverbs

程度描述	相对程度	绝对程度
极量	最 最为	极 极为 极其 极度 极端 至 至为 顶 过 过于 过分 分外 万分
高量	更 更加 更为 更其 越 越发 备加 愈加 愈 愈发 愈为 愈益 越加 格 益发 还	很 太 挺 怪 老 非常 特别 相当 十分 好 好不 甚 甚为 颇 颇为 异常 深为 满 蛮 够 多 多么 殊 特 大 大为 何等 何其 尤其 无比 尤为 不胜
中量	较 比较 较比 较为 还	不大 不太 不很 不甚
低量	稍 稍稍 稍微 稍为 稍许 略 略略 略微 略为 些微 多少	有点 有点儿 有些

2　基于极性词典的情感分析

2.1　极性短语的极性计算

文献[10]对副词连用进行了细致的研究,认为程度副词在否定词之后是程度的否定,是将原来的程度下调一个等级;程度副词在否定副词之前是对否定程度的加强。本文借鉴其思想,将极性词与其修饰词构成极性短语,并给出了极性强度的计算公式,如表 2 所示。

表 2　极性短语的极性计算

Table 2　Computation of polarity phrases

极性短语	强度计算公示	例句	强度
S = PW	$E(\mathrm{PW})$	她长的漂亮	0.8
S = NA + PW	$E(\mathrm{PW}) * E(\mathrm{NA})$	她长的不漂亮	−0.64
S = NA + NA + PW	$E(\mathrm{PW}) * E(\mathrm{NA}) * E(\mathrm{NA})$	她长的不是不漂亮	0.512
S = DA + PW	$E(\mathrm{PW}) + (1 - E(\mathrm{PW})) * L(\mathrm{DA})$ 若 PW 是正面的 $E(\mathrm{PW}) + (-1 - E(\mathrm{PW})) * L(\mathrm{DA})$ 若 PW 是负面的	她长的很漂亮 这间房间很旧	0.94 −0.94
S = DA + DA + PW	$E(\mathrm{PW}) + (1 - E(\mathrm{PW})) * L(\mathrm{DA}_1) + [1 - E(\mathrm{PW}) - (1 - E(\mathrm{PW}) * L(\mathrm{DA1})] * L(\mathrm{DA}_2)$	她长的十分很漂亮	0.982
S = NA + DA + PW	$E(\mathrm{PW}) + (1 - E(\mathrm{PW})) * (L(\mathrm{DA}) - 0.2)$	她长的不很漂亮	0.9
S = DA + NA + PW	$E(\mathrm{PW}) * E(\mathrm{NA}) + (-1 - E(\mathrm{PW}) * E(\mathrm{NA}) * L(\mathrm{DA})$	她长的很不漂亮	−0.864

表 2 中,NA 表示否定副词,DA 表示程度副词,PW 表示极性词,S 表示由修饰词和极性词组成的极性短语,$E(\mathrm{PW})$ 和 $E(\mathrm{NA})$ 分别代表极性词和否定词的极性强度。将 $E(\mathrm{NA})$ 设定为 −0.8 而不是 −1,是因为否定副词修饰中心词时不但对极性词极性置反,还会在强度上削弱,例如“不漂亮”和“丑”并不是等价关系,$L(\mathrm{DA})$ 分别定义为 0.9,0.7,0.5 和 −0.5,正、负面极性词语极性分别赋值为 0.8 和 −0.8。

2.2　句子级和篇章级极性计算

以极性短语作为极性计算的基本单位,可以计算句子的极性强度,如公式(1)所示:

$$E(\mathrm{Sentence}) = \frac{1}{n}\sum_{i=1}^{n} E(S_i)。\qquad (1)$$

其中,$E(\mathrm{Sentence})$ 代表一个句子的极性强度,$E(S_i)$ 表示该句子中的极性短语的极性。

篇章极性强度计算可以通过篇章中每个句子的极性计算得到,如公式(2)所示:

$$E(\mathrm{Text}) = \frac{1}{n}\sum_{i=1}^{n} E(\mathrm{Sentence}_i)。\qquad (2)$$

其中,$E(\mathrm{Text})$ 代表篇章的极性,它由篇章中极性句的平均强度决定;$E(\mathrm{Sentence}_i)$ 是每个极性句的极性强度。

3　实验分析

3.1　实验流程

根据本文提出的极性计算方法,设计了 3 组实验对本文构建的极性词典的效果进行了测试,分别利用本文构建的基础词典(Base)、基础词典与领域词典(Base + Domain)以及本文构建的 4 部词典(MSD)对测试集进行倾向性分类,以 0 为界线将结果分为正面、中性和负面 3 类。实验流程如下:

(1) 语料预处理。对语料进行编号、切词、标点标准化等预处理。

(2) 极性词匹配。与词典中的极性词进行匹配,识别出每一句中的极性词 PW。

(3) 识别修饰词。设置一个动态滑动窗口来识别修饰该极性词的修饰词。

(4) 构成极性短语。将识别出的极性词与修饰词构成极性短语作为极性计算单元。

(5) 极性句强度计算。利用公式(1)计算每个

极性句的极性强度。

(6) 篇章的极性计算。利用公式(2)计算整篇的极性强度。

(7) 对评论进行褒贬分类。

3.2 实验分析

本文的测试集分为两部分,一部分是 ChnSentiCorp-Htl-4000 中正、负类各 500 篇,另一部分是从驴评网(http://www.lvping.com/)上采集的酒店评论,3 名同学利用打分的形式进行人工标注,得到正负面评论各 200 篇。实验结果如表 3 所示。

表 3 不同方法对测试集的分类结果

Table 3 Classification results in different methods

测试集	Base				Base + Domain				MSD			
	正面	中性	负面	准确率%	正面	中性	负面	准确率%	正面	中性	负面	准确率%
谭松波——正面 500 篇	408	76	16	81.6	446	47	7	89.2	489	8	3	97.8
谭松波——负面 500 篇	40	93	367	73.4	18	64	418	83.6	12	32	456	91.2
驴评网——正面 200 篇	160	32	8	80.0	182	15	3	91.0	197	2	1	98.4
驴评网——负面 200 篇	12	46	142	71.0	18	29	163	81.5	6	18	176	88.0

只使用基础词典(Base)进行分类,准确率相对较低,主要原因有两种,一是基础词典中不包含“差”、“窄”、“潮”这些单字词以及“臭味”、“吵闹”、“发霉”等酒店评论领域的常用词汇;二是对于正面极性词和负面极性词数量相同评论的不能正确分类,比如“服务很好,就是有点吵”。

将基础词典和领域词典结合使用,基本解决了基础词典中缺少领域词汇的问题,分类的准确率有了较大提高,这说明建立领域词典对特定领域的倾向性分析有很大帮助。加入网络词词典和修饰词词典后,解决了正面极性词和负面极性词数量相同的评论不能正确分类的问题,准确率有了较大幅度的提高,证明将修饰词和极性词结合构成极性短语作为句子极性计算的基本单元可以提高极性计算的准确率。

通过上述几组实验证明,使用本文构建的包含基础词典、领域词典、网络词词典和修饰词词典的极性词典进行倾向性分析取得了不错的效果。

4 结束语与下一步工作

本文构建了一个用于情感分析的极性词典,将修饰词和极性词构成极性短语作为极性计算的基本单元,提出了一种基于词典的情感分析方法,取得了良好的效果。下一步研究工作主要有,利用现有的词典资源发现新的极性词语、动态地扩展和修正词典,自动完成领域词典的构建;综合考虑主题词、评价对象对极性计算的影响,结合上下文语境等信息,提出一种更加全面合理的极性计算方法。

参考文献:

[1] 赵妍妍,秦兵,刘挺. 文本情感分析综述[J]. 软件学报,2010,21(8):1834-1848.

[2] TURNEY P D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL [C]//Proceedings of the 12th European Conference on Machine Learning. Berlin: Springer-Verlag, 2001: 491-502.

[3] 朱力. 中文词语情感倾向研究[D]. 哈尔滨:哈尔滨工业大学,2009.

[4] 杨超,冯时,王大玲,等. 基于情感扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统,2010,04:691-695.

[5] 杨勇涛. WEB 舆情观点挖掘关键技术研究[D]. 西安:西安电子科技大学,2009.

[6] 谭俊武. 面向舆情分析的文本倾向性分类技术的研究与实现[D]. 湖南:国防科技大学,2009.

[7] 吕韶华. 面向中文评论文本的情感倾向性分析[D]. 大连:大连理工大学,2010.

[8] 郝雷红. 现代汉语否定副词研究[D]. 北京:首都师范大学,2003.

[9] 蔺璜,郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报:哲学社会科学版,2003,26(2):71-74.

[10] 尹洪波. 否定词与副词共现的句法语义研究[D]. 北京:中国社会科学院研究生院,2008.

(编辑:许力琴)