基于 SVM 的文本词句情感分析

杨 经 林世平
(福州大学数学与计算机科学学院 福建 福州 350108)

摘 要 近年来,文本情感倾向性分析已成为自然语言处理领域的热点,在垃圾过滤、文本分类、网络舆情分析等领域有广泛的应用。将研究中文文本词句的情感分析问题,重点解决喜、怒、哀、惧四类粒度大的情感分析问题。首先构建喜、怒、哀、惧基准情感词,然后对情感词特征进行分析,进而挖掘潜在情感词,最后使用支持向量机分类的方法融合词特征、词性特征、语义特征等各种特征,对句子进行情感识别及分类。实验表明,在 COAE2009 评测任务情感词句识别此方法是合理和有效的。

关键词 情感词 情感分析 支持向量机 特征选择

中图分类号 TP391.4 **文献标识码** A

EMOTION ANALYSIS ON TEXT WORDS AND SENTENCES BASED ON SVM

Yang Jing Lin Shiping
(*College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, Fujian, China*)

Abstract The analysis on text emotional inclination has received much attention from natural language processing field in recent years, which can be widely used in spam filtering, text classification, network public opinion analysis and other applications. This paper presents a method for analysing the emotions on words and sentences in Chinese texts, which focuses on solving four kinds of emotion analysis with big granule including happy, angry, sad and fear. The seed emotional words including happy, angry, sad and fear are firstly set up, and then we analyse the characteristics of emotional words and mine potential emotional words, finally we employ support vector machine to combine the lexical, part of speech and semantic features to recognise and classify the emotions of sentences. Experiment result shows that the method is reasonable and effective when applied to emotional words and sentences recognition in evaluation task of COAE2009.

Keywords Emotional words Emotion analysis Support vector machine Feature selection

0 引 言

随着互联网的日益普及,网络越来越成为人们获取与发布信息的主要渠道,网络舆情信息的导向作用愈来愈大。文本倾向性分析技术作为网络舆情分析研究的一个重要方向,近几年已成为学术界自然语言处理问题的研究热点。文本倾向性分析就是用户对文本中某个事物的看法或评论的分析,从而得到该看法或评论是属于对该事物的积极或消极意见,在市场预测分析、信息过滤、网络舆情分析、商业决策、电子商务等许多领域有广阔的应用前景。

目前,文本倾向性分析研究一般包括词语、句子、篇章以及海量数据的倾向性分析。大多数学者对于词语级和句子级的倾向性分析研究主要是识别在文本上下文环境中能够表达观点倾向性的词语和句子,即褒贬倾向词句。不同的是,本文将研究文本词句的情感分析,焦点放在识别明确表达人物自身情绪的情感词句,不是表达对外评价的褒贬观点词句,侧重关注四类粒度大、易区分的情感:喜、怒、哀、惧。主要是抽取在一定的上下文环境中抽取出能明确表达人物自身此四类情感词句。

1 相关工作

文本情感倾向性分析研究涉及到计算语言学、人工智能、机器学习、信息检索和数据挖掘等多方面研究内容,因此文本情感倾向性分析具有重要的学术研究价值。

词语的情感倾向研究是文本情感倾向分析的前提。词语情感倾向分析包括对词语极性、强度和上下文模式的分析。目前,对于词语的情感倾向性分析有以下方法:由已有的种子情感词典或词语知识库扩展生成情感倾向词典:英文词语情感倾向信息的获取主要是在 WordNet 和 General Inquirer[1] 的基础上进行的文献,而中文词语情感倾向信息的获取依据主要有 HowNet,朱嫣岚[2] 等通过手工选取少量的基准词,然后利用 HowNet 的语义相似度和语义相关场来计算新词与基准词的相似程度,从而判别新词的情感倾向。这些方法的缺点是对已有的种子情感词的依赖很明显。Yao[3] 等人在计算中文词汇情感倾向性时不仅考虑了词典中词汇的倾向性,而且分析了词语上下文中的情感倾向。无监督机器学习方法:Turney 与 Littman[4] 等利用候选情感词与基准情感词的点互信息(PMI)进行词汇的情感倾向判断。这种方法是目前国外较常采用的方法。缺点是这种方法也在一定程度上对种子情感词的依赖并且处理语料的领域性很强。基于人工标注语料库的学习方法:首先对情感倾向分析语料库进行手工标注。在这些标注语料的基础上利用词语的共现

收稿日期:2010 - 05 - 11。**杨经**,硕士生,主研领域:自然语言处理,数据挖掘。

关系、搭配关系或者语义关系,判断词语的情感倾向性。典型的工作如 Wiebe[5] 等利用词语的搭配模式发现在主观性文本中的倾向性词语及其搭配关系。这种方法的缺陷是需要大量的人工标注语料库。

句子的情感倾向分析任务就是对句子中的各种主观性信息进行分析和提取,判断句子的情感倾向以及从中提取出与情感倾向性论述相关联的各个要素,包括情感倾向性论述的持有者评价、对象等。Hu 与 Liu[6] 通过 WordNet 的同义词与反义词关系,获得情感词的情感倾向,然后根据句子中情感倾向占优势的情感词类进行句子极性的判断。Richardson[7] 等人通过使用依存句法树中的依存信息来抽取文本中的各种关系,但目前依存句法关系树的正确性有待提高。王根等[8] 将句子分为主观句和客观句,主观句分成赞扬和贬斥两类,并提出了一种基于多重标记 CRF(Conditional Random Field,条件随机域)的方法判断句子情感倾向性。文献[9]针对的具体任务是抽取评价词和目标对象之间的关联关系,把在同一句子中共现的评价词与评价对象作为候选集合,应用最大熵模型进行关系抽取。

本文将首先构建喜、怒、哀、惧基准情感词,然后对情感词特征进行分析,进而挖掘潜在情感词,最后采用分类的方法判断句子的情感倾向,融合词特征、词性特征、语义特征等各种特征,使用支持向量机从训练语料中学习,并对评测语料中的情感词句进行自动识别。

2 情感词句识别与分类

情感词句识别及分类要求自动识别出测试文本集中包含的情感词句并分类,即在一定的上下文环境中抽取出能明确表达人物自身情感的词语句子并判断出该情感类别,侧重关注喜、怒、哀、惧情感词句。对于情感词句识别及分类,重点在于对基准情感词的构建和对潜在情感词的处理并判断在哪种语境下具备情感。

2.1 基准情感词的构建

基准情感词的构建是中文情感词识别及分类的分析工作的基础,其质量的好坏直接决定了实验结果的效果。所谓的基准情感词就是在中文文本中具有绝对情感语义的词语,这些词语就是在文本语境中能明确表达人物内心的喜、怒、哀、惧情感。

中文情感分析的研究目前还处于起步阶段,可利用的资源还不多,目前尚没有一个完整的喜、怒、哀、惧情感词典库。对基准情感词的构建,我们通过以下三个方面完成:1)挑选出若干个能明确表达的喜、怒、哀、惧情感的词语,使用同义词词典进行扩展。2)通过网络百度百科相关词条进行搜索,人工进一步扩展基准情感词。3)利用 HowNet 提供的极性情感词词典其中包括正面情感词语、负面情感词语、正面情感评价词语、负面情感评价词语中进行人工选择。

通过同义词词典、百度百科、HowNet 三个步骤的构建,一共得出基准情感词喜类 118 个,怒 105 个,哀 111 个,惧 89 个。对于基准情感词,它们具有绝对情感语义,因此可以设定其情感强度为 1.0。然而,基准情感词数量有限,如何挖掘潜在情感词并判断在哪种语境下具备情感将是工作的重点。

2.2 喜、怒、哀、惧情感词的特征分析

针对中文自然语言的特点,对中文情感词特征进行了观察和分析,我们发现以下特征:

1)喜、怒、哀、惧情感词多为形容词和动词、名词。

2)对于中文文本句式结构的分析,发现复句中的一些连接关联词可以用来挖掘候选情感词。在实验中构建了关联词词典,关联词大致可以分为以下 4 大类:

a. 并列 和、跟、同、及、与、并、并且、而;

b. 递进 并、并且、而且、况且、不但…而且…、不仅…并且…;

c. 转折 但、但是、可是、然而、而、不过、虽然…但是…;

d. 假设 如果、假设、假若、倘若、要是、只要、即使、倘使、要不是……。我们可以判定对于具有并列和递进关联词的复句中,出现在同一句子的情感词具有相同的情感倾向,对于具有转折关联词的复句中情感词往往具有相反的情感倾向,而对于具有假设关联词的复句中不表现情感倾向。

3)程度副词对句子中情感倾向性的表达的强弱有很大的影响,同时也是识别情感词句的重要标志。实验中构建了程度词词典,根据程度词的不同程度分为四类:

a. 最高级 最、太、及其、无比、绝对、极度……;

b. 较高级 更、更加、十分、非常、特别、格外……;

c. 比较级 比较、还、还算、较为、还较……;

d. 较低级 有点、稍微、有些、略微、稍许……。

4)句子中出现了否定词,则词语和句子的情感倾向可能发生变化,否定词对句子中情感倾向的表达有直接的影响。实验中构建了否定词词典,共 18 个其中包括不、不能、没、没有、不够、不是、不会等。

5)文本中存在一些情感动作动词如心中、深感、令人、感觉等可以作为判别人物表达自身情感的标志。

2.3 潜在情感词的挖掘

所谓潜在情感词是指不直接表达情感,但潜在的隐含着表达人物自身情感的词语。因此对中文语料进行分词、词性标注、分句等预处理之后,根据上述特征 1 和 2 利用基准情感词典来挖掘识别候选情感词语,采用的是基于 HowNet 的词语语义相似度计算的方法。

基于这种方法,首先计算未知情感词 w 与喜类基准情感词的相似程度 $Happy\text{-}S(w)$ 如下:

$$Happy-S(w)=\frac{\sum_{i=1}^{k} Similar(w,Happy-word_i)}{k} \quad (1)$$

其中 w 为待判定的未知情感词,*Happy-word* 为喜类基准情感词集合,k 为喜类基准情感词集合中词语数量,$Similar(w, Happy\text{-}word_i)$ 是计算词语 w 与 $Happy\text{-}word_i$ 的相似度。

同样方法可以计算得出 w 与其余三类情感词的相似程度 $Angry\text{-}S(w)$,$Sad\text{-}S(w)$,$Fear\text{-}S(w)$,取出最大相似值设为 $Similar(w)$。实验中,对于是否可以作为候选情感词作如下判定:将相似程度临界值设为 0.45,值大于 0.45 作为潜在情感词。归类,进一步扩展情感词集合并将此类情感词的情感强度设为 $Similar(w)$。

2.4 情感句识别

中文自然语言处理中,同一词语在不同的上下文语境中可能出现不同情感倾向,这也是情感词句识别及分类研究的难点所在。例如“我哼着轻快的曲子悠闲地在校园路上散步”和“这款白色的 MP3 小巧轻快,功能强大,便于携带”中的词语轻快表达出了不同的情感倾向。在第一句“轻快”表示出了人物的喜悦情感,而第二句中的“轻快”表达客观,不包含任何的情感倾

向。所以如果单靠情感词是无法准确的识别情感词句,针对这一难点,本文将采用分类的方法判断句子的情感倾向,融合词、词性、程度词、否定词等各种特征,使用支持向量机从训练语料中学习,并对评测语料中的情感词句进行自动识别。

(1) 支持向量机

支持向量机 SVM 是近几年来发展起来的新型分类方法,是在高维特征空间使用线性函数假设空间的学习系统,在分类方面具有良好的性能。支持向量机在模式识别、知识发现等理论研究和计算机视觉与图像识别、生物信息学以及自然语言处理等相关技术研究中取得了广泛的应用。在自然语言处理中,SVM 广泛应用于短语识别、词义消歧、文本自动分类、信息过滤等方面。本文将先对语料进行特征选取,然后采用支持向量机判断句子的情感倾向,识别情感词句。

(2) 特征选择

基于机器学习的方法最重要的是特征项的构造,选择恰当的特征项对实体进行描述,有利于学习效果的提高。根据喜、怒、哀、惧情感词自身的特点及对情感词特征分析,本文选取以下 3 大特征,分别列在表 1 所示。

表 1　情感词句识别及分类特征

特征类别	特征名称	特征描述
词特征	Sentiment	情感词
	Sentiment -1	情感词左边的词
	Sentiment +1	情感词右边的词
词性特征	Pos(Sentiment -1)	情感词左边的词性
	Pos(Sentiment -1)	情感词右边的词性
语义特征	Negative word	否定词
	Degree word	程度副词
	Sentiment action word	情感动作特征词
	Suppose associated word	假设关联词

情感词表达了人物自身所表达的情感。除了情感词特征,情感词左右的词和词性在一定程度上体现了它们在句子中的结构位置。对句子进行句法分析,否定词、程度副词、情感动作特征词都是识别喜、怒、哀、惧情感句的重要标志。对于假设关联词特征的句子,我们可以判定此句判定不具备任何情感倾向。

3　实验结果与分析

实验采用的训练语料是从腾讯网、新浪等国内知名网站上下载来的约 14000 篇文章,涉及财经、房产、人才、体育、汽车、卫生、娱乐等各个领域。本文采用的测试数据是 COAE2009 评测提供的 dataset1 数据集,包括 39976 篇简体中文 txt 文本,主观文本和客观文本混合,包括真实用户评论和新闻报道评论,涉及财经、娱乐、影视、教育、房地产、电脑、手机等领域,文章长度从几个句子到上上百个句子不等。

对于实验数据集,首先进行分词、去除停用词、分句等预处理工作,然后通过上述方法对情感词的构建,利用情感词对句子进行情感过滤,对词特征、词性特征、语义特征三大类特征进行特征选取,应用支持向量机分类预测情感类别,识别情感句。

3.1　实验结果

我们参加了 COAE2009 任务 2 情感句的识别及分类评测,共有 7 支队伍提交了实验结果,其中包括北京大学、哈尔滨工业大学等。此任务是取各提交结果的前 500 条记录组成评测池,组织裁判员小组对评测池结果进行人工评判和正确答案标注,由评测池中所有正确结果组成标准答案;再根据标准答案,对各提交结果的前 1000 条记录(P@ 1000、R-accuracy 指标)和全部结果准确率、召回率、F1 指标进行自动评测打分。测试结果如表 2 所示。

表 2　情感句识别及分类的评测结果

	Precision	P@ 1000	Recall	F1	R-accuracy
Angry	0.497015	0.333	0.202925	0.288187	0.202925
Fear	0.63142	0.209	0.127052	0.211538	0.127052
Happy	0.164175	0.309	0.150658	0.090009	0.098976
Sad	0.290039	0.297	0.154206	0.201356	0.149533
Average	0.370662	0.287	0.15871	0.197772	0.144622

表中 Angry、Fear、Happy、Sad 分别为喜、怒、哀、惧情感句识别的测试结果,Average 为四类情感句识别的平均值。

北京大学计算语言学实验室徐戈、蒙新泛等人基于多模态的流行学习方法识别喜、怒、哀、惧情感词句,哈尔滨工业大学董双关、关毅等人基于最大熵模型识别情感词句,评测平均结果如表 3 所示。

表 3　北京大学、哈尔滨工业大学评测结果

	Precision	P@ 1000	Recall	F1	R-accuracy
Peking	0.0167274	0.6805	0.374071	0.032013	0.181369
HIT	0.2864017	0.3505	0.192923	0.221881	0.176617
Median	0.158407	0.366	0.201223	0.124652	0.15746
Max	0.370662	0.6805	0.374071	0.232405	0.189788

表中 Peking、HIT 分别为北京大学与哈工大的评测平均结果,Median、Max 分别为参加 COAE2009 评测结果七支队伍各项指标的平均结果和最大值。

3.2　实验分析

通过本次的测试结果和与其他大学的评测结果比较,总体而言本次在 COAE2009 情感句识别任务的评测中达到了较好的准确率,但在召回率上结果不甚理想。北京大学的评测结果在 P@ 1000、Recall 指标上取得了不错的实验测试结果,但在 Precision、F1 测试指标上偏低。另外,从喜、怒、哀、惧四个测试结果来看,喜类情感句的识别无论在准确率还是召回率上都低于其他三类,我们分析原因可能是在于喜类情感词典的构建,挖掘了一些表达正面情感但并不表达人物自身情绪的潜在情感词如祝贺、喜欢、幸福、喜庆等。另外,基于支持向量机分类方法对情感识别分类在很大程度上依赖于训练语料,情感训练语料库的数量及领域也在一定程度上影响了测试结果。

4　总结及进一步工作

本文重点研究了中文词句级的情感分析问题,抽取出表达人物自身情感即喜、怒、哀、惧四类情感粒度大的句子。通过对基准情感词的构建,潜在情感词的挖掘,采用支持向量机分类对句子进行情感分类识别,解决同一词语在不同语境的不同情感倾向问题,并取得较好结果。对目前的研究还有许多提高和改进的余地,进一步工作将以词句级情感分析为基础研究短语以

及篇章的情感分析。

参 考 文 献

[1] General Inquirer. http://wjh.harvard.edu/~inquirer.

[2] 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,2006,20(1):14-20.

[3] Yao T F, Lou D C. Research on semantic orientation distinction for Chinese sentiment words[C]//The 7th International Conference on Chinese Computing. Wuhan, 2007.

[4] Peter D Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002:417-424.

[5] Janyce Wiebe, Rebecca Brucey, Matthew Bell, et al. ACorpus Study of Evaluative and Speculative Language[C]//Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. 2001:1-10.

[6] Ming Hu, Bin Liu. Mining and summarizing customer reviews[C]//Proceedings of the 10th international conference on Knowledge discovery and data mining (KDD). 2004:168-177.

[7] Harabagiu S M, Bejan C A, Morarescu P. Shallow Semantics for Relation Extraction[C]//Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05). Edinburgh, Scotland: 2005,1061-1066.

[8] 王根,赵军. 基于多重标记 CRF 的句子情感分析研究[C]//全国第九届计算语言学学术会议. 清华大学出版社,2007.

[9] 章剑锋,张奇,黄萱菁,等. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报,2008,22(2):63-67.

(上接第 213 页)

最后,如果一个图像的马赛克宏块大于域值 MosaicThresh,标记其为马赛克图片。算法流程见算法 3。

5 马赛克检测算法的比较与实验

对于整个监测方案来说,RTP 延迟、抖动和 TR. 101-290 的各项指标的监测都是很成熟的技术,在工业界得到了广泛的使用。对于图像层监测来说,与黑场、静帧相比,马赛克的图像特征比较复杂,比较容易误检,检测的难度和复杂度都较高。

在文献[6]中,采用了 SVM 的方法,对马赛克宏块的直角进行匹配以实现马赛克检测。在文献[7]中,改进了文献[6]的直角匹配模板。在文献[2]中,采用了双域值二值化的方法来计算边缘,然后通过计算马赛克内的灰度等信息与相邻宏块的差别来进行马赛克检测。

文献[2,6,7]的计算都是以边缘检测的结果为前提的。文献[6,7]中使用的是 Canny 边缘检测[8],文献[2]使用的是差分边缘检测。因为 Canny 边缘检测的算法复杂度较高,同时 SVM 需要进行训练,具有一定局限性,所以文献[6,7]的方法不适用于实时检测。文献[2]的方法算法复杂度较低,总的时间复杂度在 $O(2\times n)$,n 为一帧的像素个数。本文的马赛克检测算法需要计算一遍宏块边的二阶梯度,时间复杂度为 $O(0.5\times n)$。

使用实际的马赛克故障视频进行实验,结果如图 6 所示。二阶梯度图很好地消除了自然边缘信息的干扰,马赛克宏块与非马赛克宏块之间的边缘梯度差明显。从单帧图片内部来说,马赛克宏块的检出率在 85% 以上,误检率为 6.7%,达到了较高的准确度。

(a)故障图 (b)一阶梯度图 (c)二阶梯度图 (d)马赛克检测结果

图 6 马赛克检测

在实验中,将马赛克宏块域值(MosaicThresh)设置为 4,测试了 100 个马赛克故障帧,其中检测出马赛克的数目为 94 帧,未检出数目为 5 帧,误检的数目为 1 帧。

文献[2]的检出率为 91%,误检率为 5%。文献[6]的检出率为 96%,误检为 5.6%。对比前人算法,本算法更好地利用了马赛克故障的特点,大幅度降低了误检率,并保持了较好的检出率。

6 结 语

本文介绍了对“全球眼”系统进行视频质量监测的方案。方案通过网络层和应用层的监测来进行服务质量的分析与故障的定位。网络层监测对中间链路各节点两端的网络质量进行分析,生成丢包抖动等告警日志。应用层监测在视频终端对图像数据进行处理,报告黑屏、静帧和马赛克等告警日志。两层监测的结果汇总至告警服务器。告警服务器就可以即时地发布故障报告,帮助运维人员提高“全球眼”系统服务质量。

参 考 文 献

[1] 张怡群. 全球眼视频监控系统架构设计[J]. 安防科技,2007(6):68-72.

[2] 司文丽,朱镇林. 马赛克故障图像的分析与监测[J]. 在线电视技术,2008(5):113-116.

[3] RFC1889. RTP: A Transport Protocol for Real-Time Applications[S]. 1996.

[4] ETSI TR 101 290 V1.2.1. Technical Report. Digital Video Broadcasting(DVB)[R]. 2001.

[5] 周锋,魏蚊龙. 电视信号中黑场和静帧的监测[J]. 有线电视技术,2005(16).

[6] Huang Xiaodong, Ma Huadong, Yuan Haidong. Video mosaic block detection based on template matching and SVM[C]//ICYCS,2008:1082-1086.

[7] 孙水发,雷帮军,刘军清,等. 基于 OpenCV 的数字视频缺陷检测快速算法[J]. 计算机工程与应用,2010,46(32).

[8] Forsyth D A, Ponce J. Computer vision: A modern approach[M]. Prentice Hall,2002.

[9] RFC4445. A Proposed Media Delivery Index(MDI)[S]. 2006-04.