文章编号:1001 - 9081(2011)12 - 3321 - 03
doi:10.3724/SP.J.1087.2011.03321

中文文本情感分析综述

魏 韡[1,2],向 阳[1],陈 千[1]
(1. 同济大学 电子与信息工程学院,上海 201804; 2. 井冈山大学 电子与信息工程学院,江西 吉安 343009)
(weiweihzkd@163.com)

摘 要:由于主观性文本有很多应用价值,情感分析近年来引起了很多研究人员的兴趣。情感分析是对主观性文本进行挖掘与分析,获取有用的知识和信息。针对中文文本情感分析的研究现状与进展进行总结。首先按粒度层次,从词语级、语句级、篇章级三个不同粒度层次细致地介绍相关的技术,再按文本的类型,分析了产品评论和新闻评论的研究进展。接着介绍了中文文本情感分析的评测和相关资源,最后总结了中文文本情感分析的研究难点与未来的研究方向。

关键词:情感分析;情感极性;中文文本;评测;语料库
中图分类号: TP391.1 **文献标志码**:A

Survey on Chinese text sentiment analysis

WEI Wei[1,2], XIANG Yang[1], CHEN Qian[1]
(1. *College of Electronics and Information Engineering, Tongji University, Shanghai* 201804, *China*;
2. *College of Electronics and Information Engineering, Jinggangshan University, Ji' an Jiangxi* 343009, *China*)

Abstract: The sentiment analysis has aroused the interest of many researchers in recent years, since the subjective texts are useful for many applications. Sentiment analysis is to mine and analyze the subjective text, aiming to acquire valuable knowledge and information. This paper surveyed the status of the art of Chinese sentiment analysis. Firstly, the technique was introduced in detail, according to different granularity levels, namely word, sentence, and document; and the research of product review and news review were presented respectively. Then evaluation and corpus for Chinese text sentiment analysis were introduced. The difficulty and trend of Chinese text sentiment analysis were concluded finally. This paper focuses on the major methods and key technologies in this field, making detailed analysis and comparison.

Key words: sentiment analysis; sentiment polarity; Chinese text; evaluation; corpus

0 引言

随着互联网的飞速发展,尤其是 Web 2.0 技术出现后,越来越多的互联网用户从单纯地获取互联网信息向创造互联网信息转变。互联网中的博客、论坛、讨论组出现了大量的由用户发布的主观性文本。这些主观性文本可以是用户对某个产品或服务的评论,或者是公众对某个新闻事件或国家政策的观点等。潜在的消费者在购买某个产品或服务时获取相关的评论可以提供决策参考,政府部门也可以浏览公众对新闻事件或国家政策的看法来了解舆情。这些主观性文本每天以指数级的速度增长,仅靠人工进行分析需要消耗大量的人力和时间。因此采用计算机来自动地分析这些主观性文本表达的情感,成为目前学术界研究的一个热点,这个热点的研究方向就是文本情感分析或称为意见挖掘。

文本情感分析是指对包含用户表示的观点、喜好、情感等的主观性文本进行检测、分析以及挖掘。文本情感倾向分析作为一个多学科交叉的研究领域,涉及包括自然语言处理、计算语言学、信息检索、机器学习、人工智能等多个领域。文献[1-3]对文本情感分析的目的、主要任务以及主流技术做了简要的介绍,但主要是介绍针对英文的文本情感分析,对中文文本情感分析并没有重点介绍。本文主要介绍针对中文文本情感分析的主流方法与研究进展。

1 不同粒度的中文文本情感分析

1.1 词语的情感极性判别

判别词语的情感极性是文本情感分析的基础。为了定量地判别词语的情感极性,通常用位于区间[-1,1]的某个实数作为情感权重表示词语的褒贬程度。通常如果情感权重大于0,则词语为褒义词;情感权重小于0,则词语为贬义词。情感权重的绝对值越大则意味着词语的褒贬程度越大。词语的情感极性判别主要有基于语料库和基于词典两种方法。

基于语料库的方法主要是利用词语之间的连词以及统计特征来判别词语的情感极性。由连词连接的词语的情感极性存在某种关联,比如由连词"和"连接的词语的情感极性相同,由连词"但是"连接的词语的情感极性相反。Yuen 等人[4]利用 Turney 的点互信息,用小规模的语料库来判别词语的情感极性。具体算法是将情感极性已确定且情感色彩强烈的词语作为种子词,通过计算需要判断情感极性的词语与这些种子词的互信息。张靖等人[5]建立基于二元语法依赖关系的情感倾向互信息特征模型,利用特征集合描述情感极性,通过机器学习方法训练分类器,自动判别词语的情感极性。

基于词典的方法是利用中文词典 HowNet 提供的语义相似度或者层次结构来判别词语的情感极性。朱嫣岚等人[6]提出了基于 HowNet 的两种词语情感极性判别方法:分别是基于语义相似度和语义相关场的方法。实验表明,基于

收稿日期:2011-05-04;修回日期:2011-07-07。 基金项目:江西省教育厅青年科学基金资助项目(GJJ11178)。
作者简介:魏韡(1983-),男,江西吉安人,讲师,博士研究生,主要研究方向:文本情感分析、数据挖掘; 向阳(1962-),男,重庆人,教授,博士生导师,主要研究方向:决策支持系统、人工智能; 陈千(1983-),男,湖北蕲春人,博士研究生,主要研究方向:数据挖掘、主题检测。

HowNet 语义相似度的方法比基于语义相关场的方法准确率更高,词频加权后的判别准确率可达80%以上。李纯等人[7]利用 HowNet 中的对词语的定义与描述,建立褒贬倾向比较强烈的词语组成种子词,并结合上下文的影响,采用一种计算方法来计算普通词与种子词之间的语义相似度来判别普通词的褒贬极性。杜伟夫等人[8]提出了一个可扩展的词语语义情感极性计算框架,将词语语义情感极性计算问题转化为优化问题。通过基于 HowNet 提供的语义相似度和基于共现率的语义相似度构建词语的无向网,利用以最小切分为目标的目标函数对无向图进行划分,使用模拟退火算法求解目标函数。实验表明该方法有较高的准确率和较好的扩展性。柳位平等人[9]挑选常用的情感词构成一个基础情感词语集,并采用词语相似度方法计算出每个词的情感倾向权值,提出的情感词权值计算方法不要求种子词数量相等。

1.2 语句的情感分析

语句的情感分析主要任务包括对语句的主客观性的区分,对主观句的褒贬性的判别,以及对语句中情感倾向的细粒度提取,包括对与情感倾向表达有关的评论持有者、评论对象、评论的倾向性及强度等。例如,句子"我认为索尼笔记本电脑质量不错而且外观漂亮"。该句中评论持有者是"我";评价对象是"索尼笔记本电脑","质量","外观",其中"索尼笔记本电脑"是间接评论对象,"质量"和"外观"是直接评论对象;"不错"和"漂亮"显示评论倾向褒义,其中"漂亮"的褒义强度要大于"不错"。

叶强等人[10]在 *N*-POS 语言模型的基础上利用卡方(CHI-square)统计方法提取中文主观文本词类组合模式,建立中文双词主观情感词类组合模式2-POS模型来自动地判断中文语句的主观性程度。实验表明采用2-POS模型的分类器对主观句的查准率和查全率接近目前英文同类研究的结果。姚天昉等人[11]利用领域本体来抽取主观句的主题以及它的属性,然后在句法分析的基础上,识别主题和情感描述项之间的关系,从而最终决定语句中每个主题的情感极性。实验结果显示,与手工标注的语料进行比较,用于识别主题和主题极性的改进后的主谓结构极性传递算法的 *F* 度量的性能有所提高。

熊德兰等人[12]提出了基于知网的语义距离和语法距离相结合的句子褒贬倾向性计算方法利用夹角余弦法对语义倾向进行了改进。党蕾等人[13]提出采用否定模式匹配与依存句法分析相结合的方法。该方法分析了修饰词极性以及否定共享模式,确定修饰词以及扩展极性的定量和否定共享范围,提出依存语法距离的影响因素来计算中文语句的情感倾向,并且在否定模式匹配后改进语句极性算法。实验结果表明该方法取得了良好的效果。

李实等人[14]根据中文语言的特点,借鉴关联规则对英文评论产品挖掘的方法,通过构建中文短语提取模式,定义中文评论中的邻近规则和独立概念,提出了面向中文网络评论的产品特征挖掘方法,数据实验证明了该方法的有效性。刘鸿宇等人[15]使用句法分析结果获取主观局中候选评价对象,同时结合基于网络挖掘的点互信息(Pointwise Mutual Information,PMI)算法和名词剪枝算法对候选评价对象进行筛选,再通过分析主观句句型归纳相应的分析规则,使用无指导的方法完成评价对象在主观句中的情感倾向性判断。

1.3 篇章的情感分析

篇章级的情感分析是指将文本从整体上区分为褒义、贬义或中性。谭松波等人[16]使用中文分词及词性标注工具ICTCLAS解析并标注中文文本,分别采用文本频率、CHI统计量、互信息、信息增益四种特征选择方法,以中心向量法、*K* 近邻、Winnow、朴素贝叶斯和支持向量机作为不同的文本分类方法,在不同的特征数量和不同规模的训练集情况下进行了实验,并对实验结果进行了比较。对比结果表明:采用文档频率特征表示方法优于其他特征选择方法和支持向量机分类方法优于其他分类方法。在足够大训练集和选择适当数量特征的情况下,文本的情感倾向分类能取得较好的效果。但是文本的主题不同对分类的结果有影响。孟凡博等人[17]设计并实现了一个基于关键词模板的文本褒贬倾向判定系统。该系统定义了关键词类别、建立了关键词库、关键词模板库,并设计了模板匹配算法和文本褒贬倾向值算法,对测试文本进行关键词及模板匹配进而判断测试文本的褒贬倾向。李寿山等人[18]具体研究四种不同的分类方法在中文情感分类上的应用,并且采用一种基于 Stacking 的组合分类方法,用以组合不同的分类方法。实验结果表明该组合方法在所有领域都能够获得比最好基分类方法更好的分类效果。

2 不同类型的中文文本情感分析

2.1 产品评论的情感分析

文本情感分析的一个重要应用领域是对互联网上出现的大量产品评论进行挖掘与分析,主要目的是能够比较精确地发现产品的优缺点。产品评论的挖掘的主要任务包括:识别并获取产品的特征或属性,定位用户的主观性评论,抽取评论搭配,判别用户评论的褒贬。产品评论的挖掘基本上是基于语句的情感分析。但是由于产品评论的主题就是产品名称,评论的持有者就是默认的使用产品的用户,所以产品评论的挖掘的重点是提取产品的特征及对应的情感词。产品特征分为显示特征和隐式特征:显示特征是指直接在评论中出现描述产品某个特征的名词;隐式特征没有明确出现在评论中但隐含表达了。

黄永文等人[19]首先对产品的规格文档进行挖掘获得产品的特征及其关系,再采用基于 BootStrapping 的弱监督机器学习方法对用户评论抽取产品的描述特征和规格特征的层次关系,先提供少量的产品特征作为种子集合,自动进行文本模式的抽取,再用抽取得到的模式抽取新的产品特征。这种方法可以看成是半自动方法,开始阶段需要人工提供少量的产品特征作为种子。宋晓雷等人[20]提出了一种不依赖外部资源的无指导评价对象自动识别方法。该方法首先综合使用词形模板和词性模板,采用模糊匹配方法和剪枝法抽取候选评价对象;然后从候选对象集中采用双向 Bootstrapping 方法识别出产品评价对象;最后通过采用 *K* 均值聚类方法对产品评价对象进行聚类,实现从评价对象中自动抽取产品名称和产品属性。那日萨等人[21]对产品评论评价和情感进行模糊建模,建立了消费者评价和情感模糊语料库,并结合消费者对产品属性的偏好,提出一种新的产品综合评价和情感计算方法。

2.2 新闻评论的情感分析

新闻评论大部分是对新闻人物或新闻事件的看法。通过对新闻评论的情感分析可以了解民众对新闻人物和新闻事件的总体评价,掌握当前的舆情信息,特别是热点事件的舆情信息。

Tsou 等人[22]在 Yuan 等人研究工作基础上对汉语报刊上有关四位政治人物褒贬性的汉语新闻报道进行了分类研究。在研究中,首先通过标记语料库获得文本中的极性元素(Polar Elements),然后主要采用了三个度量指标,即极性元

素的分布(Spread)、极性元素的密度(Density)和极性元素的语义强度(Intensity)来对每个文本进行统计,得出文本褒贬分类和强度大小的结果。徐军等人[23]用朴素贝叶斯和最大熵模型分别对新闻及评论语料进行了情感分类研究,发现选择具有语义倾向的词汇(特别是形容词和名词)对情感分类效果具有决定性作用,采用二值作为特征项权重相比采用词频作为权重的方法更能提高分类的准确率。并且最大熵模型比朴素贝叶斯的分类效果明显好。周杰等人[24]选取不同的特征集、特征维度、权重计算方法和词性等因素对网络新闻评论进行分类测试,并对实验结果进行分析比较。陶富民等人[25]构建了一个面向话题的新闻评论的情感特征提取框架,通过对那些热门话题构造对应的情感特征表来达到改善情感分析的效果。

3　中文文本情感分析评测及资源

随着中文文本的情感分析得到了越来越多的学者和研究机构的关注,为了推动中文情感分析技术的发展,国内第一个情感分析方面的评测(Chinese Opinion Analysis Evaluation, COAE)[26]于2008年举办第一届。COAE目的在于推动中文情感分析理论和技术的研究和应用,同时建立相关的分析语料库。COAE共设置6个任务,可分为3个方面:一是中文评价词语的识别和分析,属于词语级的情感分析评测;二是中文文本倾向性相关要素的抽取,主要是抽取句子中的评价对象,以及对于其观点的倾向性判别,属于语句级的情感分析评测;三是中文文本主客观性及倾向性的判别,属于篇章级的情感分析评测。

除了COAE提供了产品类的评价语料库,中国科学院计算技术研究所的谭松波博士提供的较大规模的中文酒店评论语料,约有10000篇,并标注了褒贬类别,可以为中文的篇章级的情感分类提供一定的平台。

中文的评价词词典资源有NTU评价词词典(繁体中文)和HowNet评价词词典。NTU评价词词典由台湾大学收集,含有2812个褒义词与8276个贬义词。HowNet评价词词典包含9193个中文评价词语/短语,9142个英文评价词语或短语,并被分为褒贬两类。而且该词典提供了评价短语,为情感分析提供了更丰富的情感资源。

4　结语

文本的情感分析与传统的文本分类有着特殊的挑战,主要体现在自然语言表达的丰富多变使得要计算机自动理解其中蕴含的情感语义比较困难。而中文比英文在语言结构以及句式类型更加复杂,导致针对英文文本情感分析的一些方法在对中文文本情感分析的应用并没有取得理想的结果。文本情感分析作为文本挖掘的一个新的研究方向还有很多值得深入研究的课题,尤其是中文文本情感分析近几年才开始吸引研究者的注意。未来需要深入研究的问题有以下一些:1)对于词语的情感倾向判别不应该局限在形容词,一些名词和动词也具有情感倾向,而且应该结合具体的语境和领域来判别词语的情感倾向;2)针对语句和篇章的情感分析还比较粗粒度,应该更精确地更细粒度地对某一个具体的评价对象进行分析来满足用户的需求;3)需要在自然语言处理等相关领域取得新的突破,开发新的技术和方法来更好地进行文本情感分析。

参考文献:

[1] 周立柱, 贺宇凯, 王建勇. 情感分析研究综述[J]. 计算机应用, 2008, 28(11): 2725 - 2729.

[2] 姚天昉, 程希文, 徐飞玉, 等. 文本意见挖掘综述[J]. 中文信息学报, 2008, 22(3): 71 - 80.

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

[4] YUEN R W M, CHAN T Y W, LAI T B Y, *et al*. Morpheme-based derivation of bipolar semantic orientation of Chinese words[C] // Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 1008 - 1014.

[5] 张靖, 金浩. 汉语词语情感倾向自动判断研究[J]. 计算机工程, 2010, 36(23): 194 - 196.

[6] 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.

[7] 李纯, 乔保军, 曹元大, 等. 基于语义分析的词汇倾向识别研究[J]. 模式识别与人工智能, 2008, 21(4): 482 - 487.

[8] 杜伟夫, 谭松波, 云晓春, 等. 一种新的情感词汇语义倾向计算方法[J]. 计算机研究与发展, 2009, 46(10): 1713 - 1720.

[9] 柳位平, 朱艳辉, 栗春亮, 等. 中文基础情感词词典构建方法研究[J]. 计算机应用, 2009, 29(10): 2875 - 2877.

[10] 叶强, 张紫琼, 罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报, 2007, 1(1): 79 - 91.

[11] 姚天昉, 娄德成. 汉语语句主题语义倾向分析方法的研究[J]. 中文信息学报, 2007, 27(5): 73 - 79.

[12] 熊德兰, 程菊明, 田胜利. 基于HowNet的句子褒贬倾向性研究[J]. 计算机工程与应用, 2008, 44(22): 143 - 144.

[13] 党蕾, 张蕾. 一种基于知网的中文句子情感倾向判别方法[J]. 计算机应用研究, 2010, 27(4): 1370 - 1372.

[14] 李实, 叶强, 李一军. 中文网络客户评论的产品特征挖掘方法研究[J]. 管理科学学报, 2009, 12(2): 142 - 152.

[15] 刘鸿宇, 赵妍妍, 秦兵, 等. 评价对象抽取及其倾向性分析[J]. 中文信息学报, 2010, 24(1): 84 - 88.

[16] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6): 88 - 94.

[17] 孟凡博, 蔡莲红, 陈斌, 等. 文本褒贬倾向判定系统的研究[J]. 小型微型计算机系统, 2009, 30(7): 1458 - 1461.

[18] 李寿山, 黄居仁. 基于Stacking组合分类方法的中文情感分类研究[J]. 中文信息学报, 2010, 24(5): 56 - 61.

[19] 黄永文, 何中市, 伍星. 产品特征的层次关系获取[J]. 计算机工程与应用, 2009, 45(22): 235 - 240.

[20] 宋晓雷, 王素格, 李红霞. 面向特定领域的产品评价对象自动识别研究[J]. 中文信息学报, 2010, 24(1): 89 - 93.

[21] 那日萨, 刘影, 李媛. 消费者网络评论的情感模糊计算与产品推荐研究[J]. 广西师范大学学报: 自然科学版, 2010, 28(1): 143 - 146.

[22] T'SOU B K Y, YUEN R W M, KWONG O Y, *et al*. Polarity classification of celebrity coverage in the Chinese Press[EB/OL]. [2011 - 03 - 20]. https://analysis.mitre.org/proceedings/Final_Papers_Files/109_Camera_Ready_Paper.pdf.

[23] 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报, 2007, 21(6): 95 - 100.

[24] 周杰, 林琛, 李弼程. 基于机器学习的网络新闻评论情感分类研究[J]. 计算机应用, 2010, 30(4): 1011 - 1014.

[25] 陶富民, 高军, 王腾蛟, 等. 面向话题的新闻评论的情感特[J]. 中文信息学报, 2010, 24(3): 37 - 43.

[26] 赵军, 许洪波, 黄萱菁, 等. 中文倾向性分析评测技术报告[R]. 北京: 中文信息学会, 2008.