# 跨领域中文评论的情感分类研究*

张 莉
(南京大学 计算机科学与技术系 国家重点实验室, 南京 210008)

**摘 要**: 主要对跨领域中文评论句中的各个评价对象所对应的观点表达的情感倾向进行研究。在结合单一领域特别是产品领域中情感分类的常用算法以及结合跨领域评论观点表达的特殊性的基础上,提出了基于词典资源和有监督机器学习这两种方法来对跨领域中文评论句进行情感分类,探讨了跨领域中文评论在算法上与单一领域的异同,同时对两种方法进行了比较。实验结果表明,提出的方法具有较大的实用价值。

**关键词**: 跨领域; 情感分类; 知网; 有监督机器学习方法; 支持向量机
**中图分类号**: TP391 **文献标志码**: A **文章编号**: 1001-3695(2013)03-0736-06
doi:10.3969/j.issn.1001-3695.2013.03.023

## Research on sentiment analysis of cross-domain Chinese comments

ZHANG Li
(*National Key Laboratory, Dept. of Computer Science & Technology, Nanjing University, Nanjing* 210008, *China*)

**Abstract**: This paper aimed the research on opinions' polarity of each object on Chinese comments of universal areas. Based on the freqently-used sentiment analysis algorithm for single area(especially products area) and the particularity of cross-domain opinion, it proposed two methods to determine the polariy of opinions-dictionary resource method and supervised machine learning method. Meanwhile, it also discussed the algorithm difference between cross-domain and single area and compare the two proposed methods. Experimental results show that the proposed method has great practical value.`

**Key words**: cross-domain; sentiment analysis; HowNet; supervised machine learning method; support vector machine (SVM)

## 0 引言

随着互联网技术的快速发展,人们逐渐趋向通过网络获取所需的信息,同时用户在网络上发表自己的观点也成为了一种趋势,如对于某一新闻事件的看法或对某一电子产品的评价。网络上大量的信息能够给人们的学习和生活带来很多的便利和娱乐,但有些虚假的信息和言论也会引起大众的恐慌,造成恶劣的社会影响。因此,快速、准确和全面地从大量的网络评论中自动获取有用的信息十分必要和迫切,它对社会的发展和稳定都起着重要的作用。但是要通过浏览或搜索引擎的方式从网络上海量的信息中获知对某一对象的全面评价则十分困难,而情感分析(也称为情感分类)正是利用计算机自动识别主观文本中的情感倾向(polarity,也称为极性或倾向性,如褒义和贬义)的一项技术。例如,从评论文本中自动抽取出"对于某一酒店推荐的人数和不推荐的人数各占的比例",这种技术可以让用户从繁重的搜索和总结工作中解放出来。

从 1997 年 Hatzivassiloglou 等人[1]利用连词来确定所连接的形容词的倾向性后,情感分析就成为了一个研究热点,随后出现了大量相关的研究方法和研究成果。目前常用的情感倾向判断方法有基于确定极性的种子词/词典资源、基于模板和规则以及基于有监督机器学习这几种。Turney[2]在 2001 年提出了 PMI-IR 算法,该算法基于对象间的共现现象,随后 Turney[3]又将 PMI-IR 算法用于网络评论是推荐或不推荐的分类研究中,选择 excellent 和 poor 分别作为褒义和贬义的种子词,假定与 excellent 经常共现的词具有褒义极性,与 poor 经常共现的词具有贬义极性。Hu 等人[4]利用 WordNet 中的同义词和反义词预测词的极性。朱嫣岚等人[5]在 2006 年首次提出了基于知网计算词汇的语义倾向的基于语义相似度和基于语义相关场两种方法,前者主要利用了刘群等人[6]提出的语义相似度公式,实验表明此方法在汉语常用词中的计算效果较好。孟凡博等人[7]设计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,建立了关键词库和模板库,设计了模板匹配算法及文本褒贬倾向值计算算法。Pang 等人[8]在 2002 年首次使用标准的有监督机器学习方法进行文本的情感分类,比较了朴素贝叶斯(naïve Bayes, NB)、最大熵模型(maximum entropy model, MEM)和支持向量机(SVM)这三种方法在文本语义倾向分类上的效果。唐慧丰等人[9]在 2007 年也利用有监督机器学习方法对中文文本的情感分类进行了对比研究。近年来也出现了其他特殊但效果佳的方法,如 Hassan 等人[10]借助 WordNet 构建了一个词/词性对的同义连接图,利用马尔可夫随机游走模型在连接图上确定词的极性。总的来说,利用种子词或词典的无监督方法实验简单,对种子词和词典的要求较高,在丰富的语料库和精确度高的词典资源基础上能取得较好的效果;基于词典或其他语料库的统计学模型一般能获得更为理想的实验效果;基于模板的方法人工干预较多,移植性不够好;而利用机器学习算法进行训练学习从而获得新对象的倾向性是目前常用的方法。有研究表明,这类算法的情感分类效果要优于手工分类方式。

**收稿日期**: 2012-07-12; **修回日期**: 2012-08-15 **基金项目**: 国家社会科学基金资助项目(11CYY031);国家自然科学基金资助项目(61170180)
**作者简介**: 张莉(1976-),女,江苏宜兴人,讲师,博士,主要研究方向为数据挖掘、情感分析(zhl@nju.edu.cn).

目前对于文档的情感倾向分析已经能获得令人满意的效果,而对于单一领域特别是产品领域对象的情感倾向也常常能获得较好的实验结果。而对于跨领域数据,与产品领域的观点表达常常是一个词或一个短语不同,其观点表达往往较为复杂,常常为一个或多个句子,所以目前的分类结果尚不令人满意。例如对于以下产品领域和非产品领域的两条例句:

例句 1　尼康 D7000 性能很强大,高感很适用,手感较差,套头配置令人失望。

例句 2　Linux 主要由学术界和爱好者们发展起来,并一贯遵循开放、共享的原则,因此非常适合于教育。

例句 1 中产品属性的观点表达(opinion expressions)为“很适用”“较差”和“令人失望”,这种词或短语常常用于产品领域中对象属性的评价,而例句 2 的观点表达为“一贯遵循开放、共享的原则”和“非常适合于教育”,这种以句子形式对对象进行评价的方式常常出现在非产品领域如新闻和博客等领域。因此,必须要寻找适合于跨领域中文评论特征的情感分类方法。

本文考虑用两种方法对跨领域中文评论进行情感分类:a)直接利用情感词典并结合同义词词林,同时考虑转折词和否定词等指引词对于极性的影响;b)利用有监督机器学习方法进行情感分类,探讨跨领域中文评论在文本表示、文本特征压缩和分类模型等方面与单一领域特别是产品领域的异同。

## 1　跨领域评价对象的情感倾向判断

利用如 WordNet 和 HowNet 等词典资源中的已知语义倾向情感词对评价对象的情感倾向进行判断实现方法简单,在单一领域(特别是产品领域)能获得较好的效果,而利用如 SVM 等机器学习模型在单一领域进行情感倾向判断也能获得相当的效果。在这两种方法的基础上针对跨领域评价对象即观点表达的特殊性进行算法改进,并将分类结果与单一领域的分类效果进行比较,同时对这两种方法的实验效果进行分析比较。

### 1.1　基于词典资源

基于词典资源判断评价对象的情感倾向较为简单方便,可以直接利用褒贬义词典,也可以利用褒贬种子词基于通用词典进行进一步学习,从而确定词的情感倾向。本文以原始句子为基础(观点表达部分已标出),基于情感词典,结合《同义词词林》扩展版[11],同时考虑转折词和否定词对于情感倾向的影响。具体流程如图 1、2 所示。其中图 1 为主要算法流程及否定值的计算算法,图 2 为备选部分 SP 计算极性 $P$ 和类型 $t$ 的情感倾向算法。以几个不同形式的例句为例说明算法的执行情况:

a)对例句“8 月 16 日,蓝天白云下的奥运村显得格外美丽”,人工标注的观点表达为“格外美丽”。句子中不包含转折词,根据算法寻找观点表达,句子中只有一个观点表达“格外美丽”,根据情感词典确定“美丽”的情感倾向为褒义(值为 1);另外句子中不包含否定词,所以最后确定句子的情感倾向为褒义。

b)对于例句“[陈春晓],本禹是个很实在的小伙子,我们有过几次通电话,感觉很自然,很真诚”,人工标注的观点表达为“是个很实在的小伙子”和“感觉很自然,很真诚”。句子中不包含转折词,根据算法寻找观点表达,句子中有两个观点表达,分别按照两种计算方式即计算所有的观点表达的极性和只计算第一个观点表达的极性。根据情感词典,第一个观点表达中“实在”的情感倾向为褒义,第二个观点表达中“自然”和“真诚”的情感倾向都为褒义,将值相加得到 2,所以第一种方法的初步计算结果为 2,第二种方法的初步计算结果为 1。另外句子中不包含否定词,所以最后两种方法均确定句子的情感倾向为褒义。

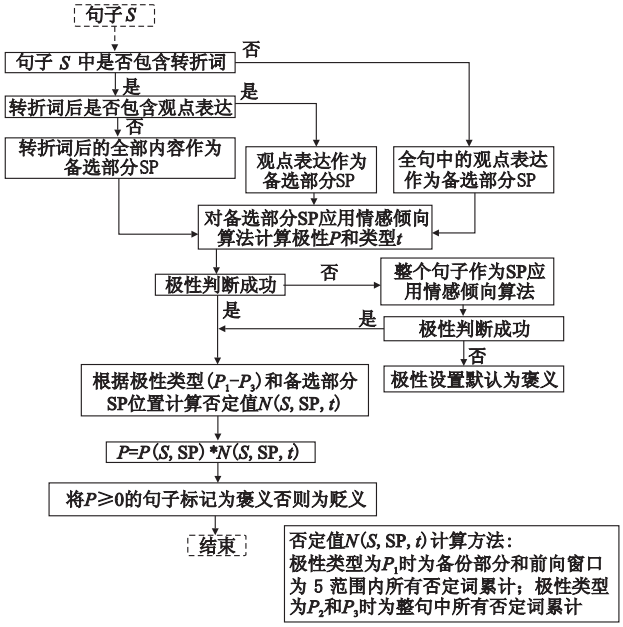


图1　情感倾向判断算法主体

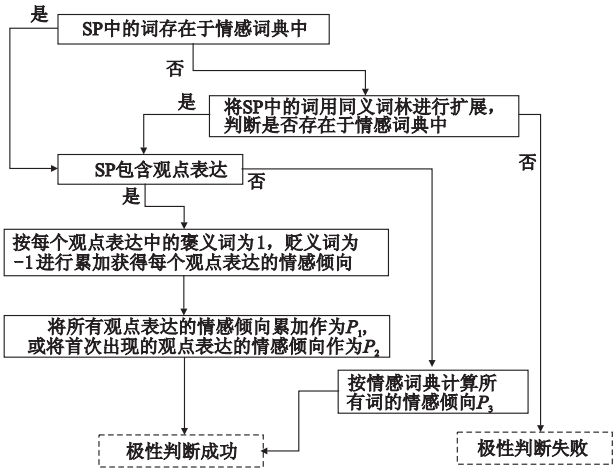


图2　情感倾向判断算法部分

c)对于例句“老师讲得很好,课程也设计得很周密,但他觉得这堂课很失败”,人工标注的观点表达为“很失败”。由于句子中包含转折词“但”,且其后包含观点表达“很失败”,根据算法,基于情感词典判断此观点表达的极性为贬义。另外句子中不包含否定词,所以最后确定句子的情感倾向为贬义。

d)对于例句“小王这个人平时的性格比较狗熊”,人工标注的观点表达为“狗熊”,根据算法需要确定观点表达“狗熊”的极性,情感词典中未包含此词,此时寻找《同义词词林》扩展版可发现与“狗熊”同义的词有“懦夫”和“胆小鬼”等。将这些词在情感词典中进行搜索确定极性,如果这些同义词仍然未出现在情感词典中,则将此观点表达的极性预设为贬义(根据评论文本中褒贬义观点表达的比例确定,可调整)。因为此句子中不包含否定词,所以最后确定句子的情感倾向为贬义。

e)对于例句“绿巨人是全片最不成功之处”,人工标注的观点表达为“是全片最不成功之处”。根据算法由情感词典确定“成功”的极性为褒义,再由于句子中包含否定词“不”,因此

极性要相反,所以最后确定句子的情感倾向为贬义。

以上以五个简单的例句说明了算法的基本执行情况。实际处理中,由于一个评价对象往往包含多个观点表达,所以句子的褒贬义值的绝对值常常大于1,将句子的褒贬义值大于等于1的标为褒义,否则标为贬义。

### 1.2 基于有监督机器学习方法

Pang等人[8]使用标准的有监督机器学习模型NB、MEM和SVM进行英文文本的情感分类并进行比较,实验取得了较好的分类效果,表明机器学习模型是情感分类的一种较为理想的方法;唐慧丰等人[9]也基于不同的特征表示、特征选择和分类方法对中文文本的情感分类进行了对比实验。

1)特征表示 在向量空间模型中,字、词或字的N-Gram等都可以作为文本特征,对于自动形成单词的英文文本来说常常利用单词或基于单词的N-Gram文本特征,而对于中文文本来说,用字作为文本特征在一定程度上忽略了字与字之间的联系,用字的N-Gram作为文本特征则较为粗糙,对于目前分词精度已能达到可接受程度的条件来说,直接用词作为文本特征要相对优于用字的N-Gram。本文采用词作为向量空间模型的文本特征,考虑到数据稀疏问题,可选择部分词性的词作为文本特征。唐慧丰等人选取了名词、动词、形容词和副词这四种词性进行了实验,实验结果表明这四种词性的合集已经能够近似地反映整个文档的情感特征,不过在不同的领域其稳定性还稍差。考虑到中文文本和跨领域文本的特征,将这四种词性进行扩充,基于"863"词性标注集增加名词修饰词、习语和缩略语这三种词性。本文将采用所有词、四种词性、补充词性这三种特征表示方式进行对比。TFI-DF是常用的权重计算函数,其主要思想是如果每个特征在一文本中出现的频率TF高且在其他文本中很少出现的话,则认为此特征具有很好的类别区分能力,而因为本文研究的是跨领域中文评论的观点表达的情感倾向,观点表达中的词频大多数为0或1,因此从理论上来说,布尔函数也会是较合适的权重计算函数。所以本文将采用布尔函数和TFI-DF函数作为权重函数进行比较。

2)特征选择 在文本特征选择方法中,由于观点表达大多长度较短,且为多个领域,所以大多数词的词频较低,利用DF进行特征选择要想获得较好的效果则需要设定较低的阈值从而起不到降维的作用。CHI与互信息类似,但是相比互信息来说功能强,因为它同时考虑了特征存在与不存在时的情况,而IG常常能获得最佳的效果。因此本文选择IG和CHI这两种文本特征选择方法进行对比实验。

3)分类模型 本文将利用NB和SVM这两种典型的机器学习分类模型对跨领域中文评论的情感倾向分类进行比较研究。

## 2 实验与分析

### 2.1 数据集和词典

基于词典资源和基于有监督机器学习方法这两个实验虽然都主要考察观点表达的情感倾向,但因为算法的不同,所以数据形式有所不同,其中前者利用了句子的WPCDEKSO形式,即利用句子中的词、词性、主张词、程度词、情感词、候选评价对象、原始句法和观点表达标注这七个语言特征的形式,后者则直接使用句子的观点表达。

1)基于词典资源算法的数据集和词典

(1)数据集 实验采用COAE2009(第二届中文倾向性分析评测)[12]任务4的标注数据集,它共包含四千多条句子,涉及体育、电影和手机等多个领域,标注的答案格式为"观点句 评论对象 倾向性",由三名标注者对数据集进行观点表达的标注,取其共同部分,并由其中一人进行检查得到最终答案,最后获得共同标注部分1 965条评论句,标注方法仍然使用IOB2形式,在标注时借助于最初标注的评价对象,并且在标注时采用相同的原则。第4章中的1 965条跨领域句子的WPCDEKSO形式,例如例句"8月16日,蓝天白云下的奥运村显得格外美丽"的WPCDEKSO形式如图3所示,其中最后一列为观点表达标注列,观点表达为"格外美丽"。数据集共包含1 990个观点表达,其中为褒义的有1 299个,贬义的有691个。数据集除评论句外,还有一个对应每个句子中不同评论对象极性的文本文件,极性分为褒义和贬义两类,褒义用"+1"表示,贬义用"-1"表示。

(2)情感词典 本实验采用知网情感词典,它包含了多个词典文本,本文所使用的褒贬义情感词典文本为中文正面评价词语和中文负面评价词语文本,前者包含3 730个词语,如"安乐"和"百折不回"等,后者包含3 116个词语,如"肮脏"和"抱残守缺"等。

(3)《同义词词林》扩展版 包含同义词词条17 817条,词语77 343个,其形式如下:

Aa01C01 = 众人 人人 人们

Aa01C02 = 人丛 人群 人海 人流 人潮

Aa01C03 = 大家 大伙儿 大家伙儿 大伙 一班人 众家 各户

"="后的词语即为同义词

2)基于有监督机器学习方法的数据集

将观点表达已进行人工标注的1 965条跨领域句子中的多个评价对象的观点表达进行拆分,保证每一条只含一个评价对象的观点表达,并将同一评价对象的多个观点表达合并为一个观点表达处理。每个评价对象的观点表达的情感倾向分为褒义和贬义两类,褒义用"pos"表示,贬义用"neg"表示。例如观点表达"非常的理想化,非常的童话,特别的感动",其情感倾向为"pos",观点表达"仁而不明,昧于识人,不辨忠奸、优劣",其情感倾向为"neg"。

### 2.2 基于词典资源的实验结果与分析

实验利用知网中文褒贬义评价词语词典、哈尔滨工业大学《同义词词林》扩展版以及转折词和否定词等对观点表达进行情感分类。

1)基于知网情感词典 表1所示为所有观点表达以及褒贬义分开的观点表达的分类精度,其中数据第一行为按照所有的观点表达进行情感计算的实验结果,数据第二行为选取第一个观点表达进行情感计算的实验结果。

表1 基于情感词典资源的分类精度 /%

| 观点表达 | 所有 | 褒义 | 贬义 |
|---|---|---|---|
| 所有 | 73.07 | 89.30 | 42.40 |
| 第一个 | 73.27 | 86.91 | 47.61 |

实验结果发现有1/3的句子利用了《同义词词林》,分析发现,由于本文的数据为跨领域网络评论文本,所以句子的领域集中度较低,且句子中观点表达所用的词说法和形式变化大,导致有较多的观点表达中的词并未在情感词典中直接出现,而其同

义词出现在了情感词典中,说明利用《同义词词林》结合情感词典判断句子的情感倾向是有效的。从表 1 可以看到,褒义观点表达的分类精度很高,而贬义观点表达的分类精度很低,导致所有观点表达的情感分类效果不是很好,除贬义词词典中词条不甚完备的原因以外,经分析有如下几个主要原因:

a)情感词典中有部分词的极性确定有问题。例如对于句子"(拉巴特讯)摩洛哥电子环保专家穆罕默德·布达哈姆在《多媒体》杂志撰文指出,电子垃圾正在成为现代社会的灾祸,而美国则是电子垃圾的'最大制造国'",其人工标注的观点表达为"则是电子垃圾的'最大制造国'",情感倾向为贬义。按所有观点表达计算的算法执行情况如下所示:

■处理第 42 行
= = 将检查的区块:[[30, 40]]
= = 处理区块 [30, 40]
- - [ + ] 是
= = 区块[30, 40]发现褒义词或者贬义词,结果为 1
= = 统计结果为 1

算法处理时,寻找观点表达中第一个出现在情感词典中的词,此处"是"出现在褒义词典中,并且句子中不包含否定词,所以最后该句子中评价对象"美国"的情感倾向为褒义。但事实上"是"这个词一般不能作为情感词,否则将出现判断错误。

b)部分否定词未被检出。例如对于句子"T60 的后壳设计不够合理",其人工标注的观点表达为"不够合理",情感倾向为贬义。按所有观点表达计算的算法执行情况如下所示:

■ 处理第 909 行
= = 将检查的区块:[[1, 6]]
= = 处理区块 [1, 6]
- - [ + ] 合理
= = 区块[1, 6]发现褒义词或者贬义词,结果为 1
= = 统计结果为 1

算法处理时在褒义情感词典中找到词语"合理"是褒义词,但未找到否定词"不够",所以最后判断句子的极性为褒义。由于本文算法中选取的均为常见的否定词,如"不""没有"和"非"等,所以导致若干否定句没有被判断出来;另外实际上人们在日常会话中更习惯于否定词搭配褒义词的说话方式,这样就导致了否定句的分类精度偏低。

c)褒义优先原则问题。例如对于句子"西门子在阿根廷涉嫌行贿近亿美元遭调查",其人工标注的观点表达为"涉嫌行贿",情感倾向为贬义。按所有观点表达计算的算法执行情况如下所示:

■ 处理第 494 行
= = 将检查的区块:[[4, 5]]
= = 处理区块 [4, 5]
- -★ 字符串[ 涉嫌 ]在同义词表中,但任一同义词均不能命中褒义词表或者贬义词表
- -★ 字符串[ 行贿 ]在同义词表中,但任一同义词均不能命中褒义词表或者贬义词表
= = 所有区块都不含有褒义或者贬义词
= = 结果强制为 1
= = 统计结果为 1

观点表达中的"涉嫌"和"行贿"在情感词典中均未找到,另外这两个词在《同义词词林》扩展版中的同义词也未出现在情感词典中,所以最后结果强制为 1(褒义),这样同样会导致贬义倾向句子的分类精度较低。而对于句子"为了节省计算机资源,有的用户根本没有打开杀毒软件的邮件监控,这些不好的使用习惯对用户的安全造成了很大威胁",其人工标注的观点表达为"不好的"和"对用户的安全造成了很大威胁",情感倾向为贬义。按所有观点表达计算的算法执行情况如下所示:

■ 处理第 6130 行
= = 将检查的区块:[[17, 18], [21, 29]]
= = 处理区块 [17, 18]
- - - - 词林命中贬义词表,[ 不好 ]匹配到同义词[ 坏 ]
= = 区块[17, 18]发现褒义词或者贬义词,结果为 -1
= = 处理区块 [21, 29]
- - [ + ] 对
= = 区块[21, 29]发现褒义词或者贬义词,结果为 1
= = 统计结果为 0

"不好的"中的"不好"在情感词典中未找到,利用《同义词词林》扩展找到"不好"的同义词"坏","坏"在贬义词情感词典中找到,所以"不好的"的情感倾向被标为 -1;观点表达"对用户的安全造成了很大威胁",词"对"首先在褒义情感词词典中被找到,同时此观点表达本身和窗口 5 之内无否定词,所以其标注结果为 1。综合的统计结果为 0,根据算法将 0 标为 1(褒义)。第二个观点表达在情感词典中匹配了词"对",可以看到"对"在此处并非为评价词,且因为没有对观点表达中各词的情感强弱作判断,所以没法确定每个词的重要程度,这样导致了第二个观点表达被误判为褒义,最终导致整个句子的情感统计结果为 0,最后被错误标注为褒义,同样导致了贬义倾向句子的低情感分类精度。而只利用第一个出现的观点表达计算整个句子的情感倾向与利用所有观点表达进行句子情感倾向的计算实验结果相差无几,说明了转折词后或不含转折词的句子通常只包含一种情感。

从实验结果分析可以看到,利用情感词典资源和同义词计算情感倾向时,情感词典的准确度非常重要。为了说明问题,本文对知网的情感词典进行简单的修正。

2)基于修正的知网情感词典　为了验证情感词典对于情感分类的作用,本文对知网词典作了简单的修改,根据实验结果看到,由于知网情感词典中若干常用词含有多个含义,并不能被明确归类为褒义或贬义。例如单个字的词"是""长""高"和"苦"等,非单个字的词如"根本""经常性""物理"和"名义上"等,简单起见,本文将如上所述类似的明显不能确定情感倾向的词进行删除,删除后的实验结果如表 2 所示。

表 2　删除后的实验结果　/%

| 观点表达 | 所有 | 褒义 | 贬义 |
|---|---|---|---|
| 所有 | 74.62 | 89.07 | 47.47 |
| 第一个 | 74.37 | 86.45 | 51.66 |

从表 2 看到,仅仅对知网情感词典进行简单删除,不管是利用所有观点表达还是第一个出现的观点表达,判断句子中评价对象的情感倾向的分类精度均有所提高,表明情感词典对于情感分类有重要的作用。

### 2.3　基于有监督机器学习方法的实验结果与分析

实验主要对跨领域评论文本的观点表达的特征表示、权重函数、特征选择和分类方法这几个方面进行比较研究。

1)基于所有词和部分词性词的特征表示结果与分析

首先将观点表达利用 LTP2.0 进行分词。三种不同的文本特征为:

a)所有词。所有词作为文本特征,不需要考虑词性。

b)四种词性。按照“863”词性标注集,四种词性(名词、动词、形容词和副词)中名词的标注形式包括 n、nd、nh、ni、nl、ns、nt、nz 和 r,动词的标注形式为 v,形容词的标注形式为 a,副词的标注形式为 d。

c)补充词性。名词修饰词的标注形式为 b,习语的标注形式为 i,缩略语的标注形式为 j。

权重利用布尔函数和 TFI-DF 函数分别计算,分类算法为 SVM,使用台湾大学林智仁教授开发的 LIBSVM,核函数使用线性核函数(Linear),相应参数均使用默认形式。与 Pang 等人[8]一样,本文也采用 3 折交叉验证的方式,分类精度如表 3 所示。

表 3 不同特征表示的分类结果 /%

| 函数 | 所有词 | 四种词性 | 扩充词性 |
|---|---|---|---|
| 布尔 | 75.58 | 74.32 | 75.38 |
| TFI-DF | 75.68 | 75.13 | 75.93 |

由表 3 可以看到:

a)与单一产品领域一样,选择部分词性作为特征表示进行训练和学习的效果与选择所有词的效果相当,所以跨领域数据可以与单一领域一样只选择部分能表示文本情感特征的词作为文本特征,起到降维的作用。

b)对四种词性进行扩充后分类精度均有所提高,且其能获得与所有词作为文本特征相当的效果,表明新扩充的三种词性能够帮助表示文本的情感特征。为了进一步实验部分词性词与所有词的效果,以下实验的特征表示使用所有词和扩充词性这两种形式。

c)从实验结果来看,两种权重计算函数 TFI-DF 和布尔函数效果相差不大,表明在跨领域数据的情感分类中利用布尔函数计算文本特征的权重是可行的。而 TFI-DF 函数的效果略好于布尔函数,所以以下实验中的权重计算均使用 TFI-DF 函数。

2)基于所有词和扩充词性的特征选择结果与分析

利用 CHI 和 IG 进行特征选择,其中 IG 选择 500 和 2 000 两种不同的特征数目,实验结果如表 4 所示。

表 4 不同特征选择的分类结果 /%

| 分类 | CHI(500) | CHI(2000) | IG(500) | IG(2000) |
|---|---|---|---|---|
| 所有词 | 70.70 | 73.17 | 70.70 | 73.17 |
| 扩充词性 | 70.40 | 74.07 | 70.40 | 73.97 |

从表 4 可以看到,CHI 和 IG 的效果非常接近,且扩充词性数据利用 CHI 选择特征数目是 2 000 时效果略好于 IG,这与以往大多实验表明的 IG 的效果最佳有所不同。产生这种现象的主要原因是因为本文的实验数据为跨领域观点表达,特征词的频率很低,所有 CHI 这种对低频词偏袒(一个特征词条在文档中出现一次和多次同样对待)的特征选择方法反而更为合适,而 IG 也考虑了低频词对于分类的影响,因而也能取得相当的效果。从表 4 显示的所有词和扩充词性的实验效果来看,可以进一步证明部分典型的词性在情感上的确能代表整个文本集。

另外本文发现如唐慧丰等人[9]提出的单一领域的分类精度随特征的增加而增大,但特征数达到一定值时分类精度达到最佳这样的现象同样出现在跨领域数据中,如图 4 所示即为所有词利用 CHI 特征选择方法在选择某几个不同的特征数目时分类精度的变化情况。

如图 4 所示,所有词利用 CHI 选择特征数目越多其分类精度在总的趋势上越高,但在特征数目到达某一个值附近(图中特征数目为 4 150 个)时分类精度到达最高,说明跨领域数据与单一领域一样,并非特征数量越多越好,当其达到一定值时分类精度将趋于平稳,而如果特征数目大于该值时分类精度反而会有所下降。

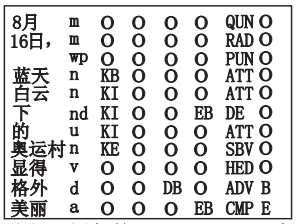


图3 例句的WPCDEKSO形式

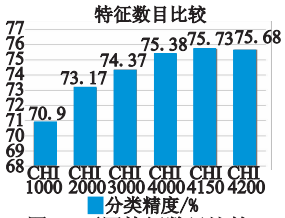


图4 不同特征数目比较

3)基于 SVM 和 NB 的分类结果与分析

利用扩充词性词作为特征表示方法,CHI(选择 50% 的特征数目)作为特征选择方法,分别使用 SVM 和 NB 这两种机器学习模型进行情感分类,其中 SVM 仍然使用 LIBSVM,核函数选用线性核,NB 使用 Weka 中的 NaiveBayes 分类器,两者均使用默认参数并采用 3 折交叉验证方式。SVM 和 NB 两种算法的分类精度结果分别为 74.07% 和 51.31%,SVM 的分类精度比 NB 高出 22.76%。可见,不管是在混合领域还是在单一领域,SVM 算法均具有较大的分类优势。

4)基于 SVM 的不同核函数及其参数分类结果与分析

SVM 常用的核函数有四类,包括线性核函数、多项式核函数、径向基核函数和 *S* 形核函数,不同的情况下核函数的性能不一。在核函数类型确定后则需要确定其参数,参数的选择非常重要,如果选择得过大或过小有可能出现“过拟合”或只能分出一类等问题。各类核函数具有的参数个数不一,惩罚系数 *C* 是各类核函数都有的参数,*C* 越高表示越重视离群点。寻找最优参数可采用网格验证和交叉搜索等方法。本文旨在比较各类核函数对于跨领域数据的影响,所以只在其他参数合理的前提下考察各类核函数在不同的惩罚系数 *C* 下的分类精度并进行比较。比较结果如表 5 所示,用 10 折交叉验证方式。

表 5 不同惩罚系数的各类核函数分类结果 /%

| 函数 | $C=1$ | $C=10$ | $C=50$ | $C=100$ | $C=500$ | $C=1000$ |
|---|---|---|---|---|---|---|
| Linear | 75.93 | 75.03 | 72.86 | 70.96 | 70.10 | 65.80 |
| Polynomial | 65.28 | 65.28 | 65.28 | 65.28 | 65.28 | 65.28 |
| RBF | 65.28 | 65.28 | 65.28 | 65.28 | 70.50 | 75.78 |
| Sigmoid | 65.28 | 65.28 | 65.28 | 65.28 | 67.24 | 70.55 |

从表 5 的分类结果来看,线性核函数在惩罚系数较小的情况下表现较好,而径向基核函数在惩罚系数较高时表现最佳,这与林智仁教授等人认为的对于核函数及其参数的认识即线性核函数和径向基核函数在文本分类上的性能较好是一致的,也表明了跨领域数据在核函数以及参数的选择上与单一领域类似。另外值得一提的是,表 5 中出现的分类精度 65.28% 为只判断出正类(褒义)的结果,可见惩罚系数在分类性能上具有重要的作用。

## 2.4 两种方法比较

利用情感词典和同义词词林并借助于转折词和否定词的方法与利用机器学习模型 SVM 方法确定句子中评价对象的情感倾向的方法实验效果差别不大。前一种方法经典易实现,如果有准确度高的情感词典和否定词词典,则其能获得较高的分类精度;后一种方法不需要构建情感、否定词以及其他如《同义词词林》这样的词典,人工参与少,只要有足够的语料库和寻找到最优的 SVM 参数同样,可以获得较高的分类精度。

另外从实验结果也可以看到,本文所用实验数据其评价对象的情感倾向的分类精度不高,这主要是因为所用的数据是跨

领域的且为网络评论,观点表达的结构复杂且用词变化大,所以情感词典的命中率往往不会特别高,导致分类精度也不会特别高。而唐慧丰等人[9]提出的利用有监督机器学习模型 SVM 进行情感分类的方法与本文提出的第二种方法类似,由于其数据主要为单一产品领域,其分类精度能达到 90% 以上;而另一方面,对于本文所研究的跨领域评论文本在用有监督机器学习方法进行训练时比单一领域需要更大的语料库,但是由于本文的训练集较小,也同样导致了分类精度不高。

## 3　结束语

本文在对情感分类常用算法进行综述及确定跨领域评论观点表达的特殊性的基础上提出了基于词典资源和有监督机器学习这两种方法来对跨领域中文评论句的情感进行分类。基于词典资源的方法为利用知网评价词词典和《同义词词林》,并结合转折词和否定词确定观点表达的极性,同时还根据语料的特征确定了褒义优先的原则,对知网评价词词典进行了简单的修正并进行了实验以证明情感词典在情感分类中的重要作用。基于有监督机器学习方法则采用所有词、四种词性(名词、动词、形容词和副词)和补充词性这三种文本表示方法,TFI-DF 和布尔函数这两种权重计算函数,IG 和 CHI 这两种文本特征选择方法以及 SVM 和 NB 这两种文本分类模型进行对比实验。从实验结果可以看到,补充词性的文本表示方法起到了缩小数据集的作用,并且其性能与采用所有词进行特征表示相当,同时布尔函数适合作为跨领域评论文本分类的权重计算函数。在文本特征选择上,CHI 比 IG 在跨领域数据上相对更有优势,而 SVM 比 NB 在跨领域评论句观点表达的情感分类上具有明显的优势。最后对两种方法进行了比较。

**参考文献:**


[1] HATZIVASSILOGLOU V, McKEOWN K R. Predicting the semantic orientation of adjectives[C]//Proc of the 8th Conference of European Chapter of the ACL. 1997:174-181.

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

[3] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]//Proc of the 40th Annual Meeting Association for Computational Linguistics. 2002:417-424.

[4] HU Ming-qing, LIU Bing. Mining and summarizing customer reviews [C]//Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004:168-177.

[5] 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.

[6] 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.

[7] 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,2008,7(7):1458-1461.

[8] PANG Bo, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning teehniques[C]//Proc of ACL Conference on EMNLP. 2002:79-86.

[9] 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):55-94.

[10] HASSAN A, RADEV D. Identifying text polarity using random walks [C]//Proc of the 48th Annual Meeting of the ACL. 2010:395-403.

[11]《同义词词林》扩展版[EB/OL]. http://www.ir-lab.org/.

[12] COAE[EB/OL]. http://www.ir-china.org.cn/information.html/.

[13] LIBSVM[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[14] KEERTHI S S, LIN C J. Asymptotic behaviors of supportvector machines with Gaussian kernel[J]. Neural Computation, 2003, 15(7):1667-1689.


(上接第 735 页)学习状态和协作能力多维指标的学习者特征模型,并以优秀度、活跃度和影响力对其进行量化分析,应用模糊相似关系演算和模糊等价演算,对学习者进行多层次的模糊动态聚类,识别候选学习领袖,并在其基础上通过融合多类型学习者,以迭代方式完成学习小组的划分。实验分析表明,本文方法可显著提高虚拟学习社区中学习小组划分的准确性和自适应性,有利于所有学习者在知识水平、学习积极性、协作能力上的整体提升。

**参考文献:**


[1] 袁磊,黄道鸣. CSCL 环境中的社会交互[J]. 现代教育技术,2004,14(4):46-49.

[2] CHEN C, CHANG C. Mining learning social networks for cooperative learning with appropriate learning partners in a problem-based learning environment[J]. Interactive Learning Environments, 2012, 20(1): 1-28.

[3] GREER J, McCALLA G, COOKE J, *et al*. The intelligent helpdesk: supporting peer-help in a university course[EB/OL]. [2012-07-03]. http://julita.usask.ca/Texte/Jim-html/I-Help.htm.

[4] 程向荣, 周竹荣, 邓小清. 计算机支持的协作学习的伙伴模型[J]. 计算机应用,2007,27(7):1763-1766.

[5] 程艳, 许维胜, 何一文. 虚拟学习社区的绩效评估模型[J]. 计算机工程与应用,2011,47(4):17-21.

[6] CHEN Rong-chang, CHEN Shih-ying, FAN Jyun-you, *et al*. Grouping partners for cooperative learning using genetic algorithm and social network analysis[J]. Procedia Engineering, 2012, 29(1): 3888-3893.

[7] WANG Dai-yi, LIU Yen-chun, SUN Chuen-tsai. A grouping system used to form teams full of thinking styles for highly debating[C]// Proc of the 10th WSEAS International Conference on Systems. 2006: 721-726.

[8] 黄伟. 社会网络分析视角下的虚拟学习社群研究[J]. 电化教育研究,2011(12): 53-58.

[9] 赖文华, 叶新东. 虚拟学习社区中知识共享的社会网络分析[J]. 现代教育技术,2010,20(10): 97-101.

[10] 张豪锋, 李瑞萍, 李名. QQ 虚拟学习社群的社会网络分析[J]. 现代教育技术,2009,19(12): 80-84.

[11] 张舸, 周东岱, 葛情情. 自适应学习系统中学习者特征模型及建模方法述评[J]. 现代教育技术,2012,22(5):77-82.

[12] 王万森, 龚文. E-Learning 中情绪认知个性化学生模型的研究[J]. 计算机应用研究,2011,28(11):4174-4177.

[13] 陈淑洁, 叶新东, 邹文才. 社会网络分析在网络课程评价中的应用研究[J]. 现代教育技术,2009,19(3): 26-30.

[14] 林聚任. 社会网络分析:理论、方法与应用[M]. 北京: 北京师范大学出版社,2009.

[15] 刘军. 整体网分析讲义——UCINET 软件实用指南[M]. 上海: 北京师范大学出版社,2009.

[16] 李士勇. 工程模糊数学及应用[M]. 哈尔滨: 哈尔滨工业大学出版社,2004.

[17] 张红宇, 王坚强, 高阳. E-learning 2.0 环境下高校教学资源的构建及应用[J]. 计算机教育,2012(14):52-55,59.