基于情感词识别的BBS情感分类研究

陈锦禾,范 新,沈 闻,沈 洁
(扬州大学 信息工程学院 计算机科学系,江苏 扬州 225009)

摘 要:针对目前BBS网络信息杂乱的现象,提出了一种BBS情感分类方法,能够方便用户准确定位所需信息,辨识评论的极性(肯定还是否定)。根据词语具有语义倾向的概率大小,利用最大熵的特征模型识别文本中具有语义倾向的词语,选择具有一定倾向值的词作为文档的特征表示。通过这些类型特征构造支持向量机分类模型,对BBS文本所表达的情感等主观内容进行分类,判断其是正面还是负面。实验表明,在BBS情感分类中,基于该特征表示的分类精度较好。

关键词:文本分类;情感分类;特征词识别;最大熵;支持向量机
中图分类号:TP311.1 **文献标识码**:A **文章编号**:1673-629X(2009)07-0120-04

Research on Sentiment Classification of BBS Reviews Based on Identifying Words with Polarity

CHEN Jin-he, FAN Xin, SHEN Wen, SHEN Jie
(Dept. of Computer Science, College of Information Engineering,
Yangzhou University, Yangzhou 225009, China)

Abstract: Aiming at the phenomenon BBS network information is mess, present a high performance method to solve BBS sentiment classification problems. It can help people locate the required reviews in the BBS, and identify the comment is affirmatives of negatives. Based on the different probability whether the words have polarity, use maximum entropy to identify the words with polarity as features. Then use SVM classifier to deal with the texts in order to judge it is positive or negative. The experiments show that this method achieves a high performance.

Key words: text classification; sentiment classification; feature recognition; maximum entropy; SVM

0 引 言

近年来BBS的迅速发展,主观性的言论越来越多。如何利用这些丰富的评论资源,对评论的主观内容进行分析与处理,成为研究的问题。基于情感的文本分类是近年来兴起的一个研究方向,主要研究通过文本的情感表达来对文本进行分类,将其分为正面和负面。基于情感的文本分类与基于主题的文本分类不同。基于主题的文本分类主要以主题词为主,这方面的研究已经很成熟,产生了很多有监督的学习方法以及文本特征表示方法和特征选择机制[1~3]。基于情感的文本分类主要以具有主观性的词为主,有监督的学习方法以及文本特征表示方法和特征选择机制在文本的情感分类中不尽人意。针对各种有监督的学习方法以及文本特征表示方法和特征选择机制等因素对情感分类性能的影响,提出了很多有效的解决方法。Bo Pang[4]采用传统的的基于主题的机器学习方法(Naive Bayes, Maximum Entropy, Support Vector Machine)分别对电影评论进行分类,得出结合Unigrams和SVM的分类效果最好。Wei-Hao Lin[5]提出了一种统计模型,通过对词的用法的分析鉴别出文档所要表达的观点。徐军等人[6]利用朴素贝叶斯和最大熵方法进行新闻及评论语料的情感分类研究,取得了较好的分类性能。唐慧丰等人[7]考虑了不同的文本特征表示、特征选择方法和文本分类方法进行情感分类实验。实验表明采用BIGrams特征表示方法、信息增益特征选择方法和SVM分类方法,在足够大训练集和选择适当数量特征的情况下,情感分类能取得较好的效果。但是,这些研究只是从词性的角度考虑特征的表示,由于中文语义的复杂性,引入了很多与表达作者态度不相关的词,同时忽视了一些使用频率较低的词性所表达的情感,如叹词、状态词等。

收稿日期:2008-10-27;**修回日期**:2008-12-15
基金项目:国家自然科学基金资助(60673060)
作者简介:陈锦禾(1965-),男,江苏靖江人,工程师,研究方向为信息管理与安全;沈 洁,教授,硕导,主要研究方向为数据挖掘、信息管理。

文本的情感分类主要基于具有主观性的词,采用基于主题的文本特征表示引入太多的无关特征项,对情感分类的性能具有很大的影响。情感的正面表达和负面表达主要以形容词、名词、副词和少数动词为主。最大熵方法被广泛应用于自然语言处理领域,在词性标注[8]、语义消歧[9]等方面取得了很好的效果。文中主要以词语的语义倾向识别为基础,利用最大熵识别文档中具有语义倾向的词,在此基础上选择合理的具有一定语义倾向值的词作为文档的特征表示,构建 SVM 分类器进行 BBS 文档的情感分类。

1　基于最大熵的情感词识别

最大熵是一个比较成熟的数学理论,主要用于估计事件的概率分布。最大熵模型根据已知的知识建立模型,对未知的事实不做任何假设,不依赖语言模型,独立于特定的任务。文中将对情感词的识别视为一个分类问题,计算一个词是不是情感词的概率值,通过比较两者的大小,选择概率值大的作为最终的结果。最大熵模型的建立主要基于特征选择和参数估计。特征选择主要是寻找约束条件,选出对模型有表征意义的特征。继而计算出对应的每个特征的参数估计值,建立模型。根据 BBS 评论文档的特点,选择合适的特征集合,标引文档中各种词语的语义倾向值,从而合理选择词语作为文档的特征表示,这里的具有语义倾向的词语的知识就是最大熵所要寻找的特征。通过对 BBS 语料的分析,文中主要考虑如下几种因素:

{词性,位置,长度,出现次数,是否在特殊符号中}

对训练语料中的文档进行手工标注情感词,每个文档由一组情感词集合 $\theta(d_i)$ 组成。同时,由自动生成程序为每个文档生成一个未判别的情感词集合 $\vartheta(d_i)$。每个未判别的情感词具有以上特征,表示成 word{ Part_ of_ speech,Pos,Len,Fre,Style}。将这些词作为一个事件,对其建立特征向量,在此特征限制下,利用最大熵模型计算出每个 $\text{word}_j \in \vartheta(d_i)$ 最优的概率分布 $p(f(\text{word}_j)' \mid \text{word}_j)$。$f(\text{word}_j)'$ 表示 word_j 被标引的结果,它只有两种取值。

$$f(\text{word})' = \begin{cases} 1, \text{word 被标注为情感词} \\ 0, \text{word 未被标注为情感词} \end{cases}$$

word 对应的每个特征表示为一个二值函数 $f(\text{word})$,$f'(\text{word})$,该函数的值属于集合{0,1}。在该特征与情感词所具有的特征一致且 $f'(\text{word}) = 1$ 的条件下,取值为 1,或者特征与情感词所具有的特征不一致且 $f'(\text{word}) = 0$,取值也为 1,其余情况取值为 0。

根据最大熵理论建立模型,满足最大熵的条件用指数形式表示为:

$$\begin{cases} p_\lambda(f' \mid \text{word}) = 1/Z_\lambda(\text{word})\exp\left(\sum_i \lambda_i f_i(\text{word}, f')\right) \\ Z_\lambda(\text{word}) = \sum_{f'} \exp\left(\sum_i \lambda_i f_i(\text{word}, f')\right) \end{cases} \tag{1}$$

Z_λ 称为归一化因子,参数 λ_i 对应 f_i 所代表的特征的权重。采用 GIS(Generalized Iterative Scaling) 算法,计算每个对应的具有最大熵分布的参数值 λ_i,从而得到各个情感词的概率分布。算法如下:

```
Input: document
begin
  利用分词系统分割成词的集合
  for 集合中的所有词
    begin
      {f'(word) = 1
       基于表 1 的特征模板获取该词所对应的特征向量
       根据训练语料所得出的参数值 λi 代入最大熵公式(1)
       计算词属于情感词的概率 p1;
      }
    end
    begin
      {f'(word) = 1
       基于表 1 的特征模板获取该词所对应的特征向量
       根据训练语料所得出的参数值 λi 代入最大熵公式(1)
       计算词属于情感词的概率 p2;
      }
    end
    if p1 > p2
      该词为情感词
    else
      该词为非情感词
  endfor
end
```

2　基于情感词标引的 SVM 分类

对 BBS 评论文档的基于情感的文本分类是一个两类文本分类问题。假设预定义的文本类型集为 $C = \{c_1, c_2\}$,其中 c_1 表示肯定的评论,c_2 表示否定的评论。要进行分类的文本集 $D = \{d_1, d_2, \cdots, d_n\}$,采用 SVM 分类,对每个文档分配一个类型标记 c_1 或 c_2。支持向量机(SVM) 是一种传统的机器学习方法,在基于主题的文本分类中分类性能较好。它的主要思想是:找到一个具有最大间隔的超平面,将两类文本没有错误地分开。在对文本的情感分类,我们的目标是要在带约束的条件下求解一个最优值。

在进行分类之前,需要把文档进行特征表示。采用向量空间模型(VSM) 表示,$V(d) = \{t_1, w_1; t_2, w_2; \cdots, t_n, w_n\}$,其中 $t_i (i \in 1,2,\cdots,n)$ 表示特征词,w_i 表

示 t_i 在 d 中的权重。通过最大熵模型对词的标引，选择具有情感倾向的词作为特征词。

3 实验结果与分析

实验数据使用扬州大学启明星论坛近年的相关评论 2476 篇，这些文档以 1543 篇正面的评论和 933 篇负面的评论组成，不包括中性的评论。语料采集后，处理转换成统一的文本格式。人工标注正面和负面的评论，50% 作为训练集，50% 作为测试集，训练集和测试集不重复。文中采用二值(Binary)[10]作为特征项权重，文本的分词及词性标注采用中科院的 ICTCLAS 系统。与基于主题的文本分类不同，一些否定词、叹词等停用词在基于情感的文本分类中往往具有语义倾向，不在被作为停用词删除。研究特征表示对基于情感的文本分类的影响，分别以基于主题的特征表示(选取全部特征)，基于词性选择的特征表示和文中提出的基于最大熵识别的特征表示进行比较实验。

主题的特征表示，选取全部特征作为文本的表示方法，构建 SVM 分类器进行分类。这种方法在基于主题的文本分类中，具有较高的分类精确度。基于词性选择的特征表示中，从不同词性具有情感程度的强弱考虑，分别选择单一词性作为特征表示。通过对评论语料的分析，发现表达态度的肯定与否定主要以名词、动词、形容词、副词为主，对选取其中的一种和选取它们的全部分别进行了实验。根据文中提出的基于最大熵识别的特征表示的分类，在构建分类器之前，通过训练样本实例化最大熵模型。根据最大熵模型的特征模板，获得参数值。对于新文档，首先获得一个特征集合，再根据最大熵模型，选择具有情感倾向的词作为文档的特征表示。图 1 是最大熵情感词识别模型，从图中可以看到情感词的识别由训练过程和测试过程组成。

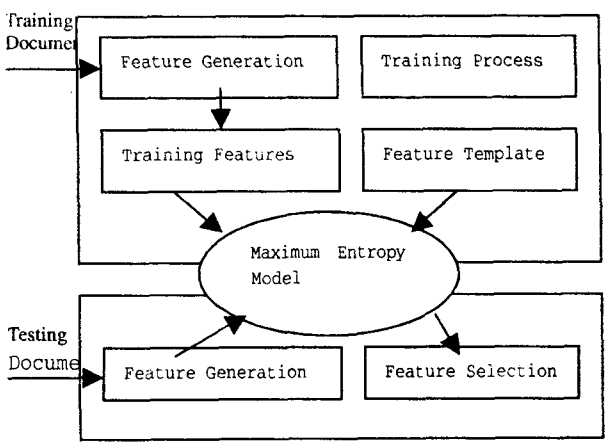

图 1 基于最大熵的情感词识别

从文档抽取 6740 个词作为样本，其中包括了 2659 个正例和 4081 个负例，正例表示该词具有语义倾向，负例表示该词不具有语义倾向。通过五次交叉测试，得出最大熵模型对于情感词识别的准确率见表 1。

表 1 最大熵识别的准确率

	第一组	第二组	第三组	第四组	第五组
准确率	0.6471	0.6025	0.6214	0.6138	0.6347

从 BBS 语料中发现很多的词本身不具有情感倾向，但是当结合其他的词共同出现的时候，具有情感倾向，以及由于训练数据量的限制，导致对特征参数的估计时有欠缺。从而影响了最大熵在对特征词进行情感鉴别的准确率。基于该最大熵模型对情感词的识别，进行文档特征表示的情感分类结果见表 2。

表 2 不同特征表示方法的分类精确比较

	特征表示	特征数量	准确率
基于主题的特征表示	选取全部特征	8254	69.58%
基于词性选择的特征表示	名词	2120	85.42%
	动词	2016	83.26%
	形容词	1928	74.56%
	副词	1864	73.48%
	以上词性的全部	7928	92.35%
基于最大熵识别的特征表示	识别为具有情感倾向的词作为特征	5432	96.73%

从表 2 中可以看到，主题的特征表示方法分类效果较差。基于情感的文本分类主要以一些主观性的词为主，基于主题的特征表示，选取全部特征作为文档的表示，没有将情感词和普通词区分开来，引入了大量无关的特征项，干扰了文本所表达的情感倾向。例如以下是某一作者对电影《美国田园下的罪恶》的评论：

例：很久没有看到一部震撼人心的电影了。所幸在那个毫无准备的深夜，看到今年新片《美国田园下的罪恶》，算是弥补了长久的苍白。此片最打动我的是，几乎每一个角色都那样普通，在现实生活中随处可见。

从上面的例子可以看到作者表达了明显的肯定态度，但是能够表达作者这一情感的只有像“震撼”、“打动”等小部分具有情感倾向的词，引入了很多与表达作者态度不相关的词。在进行文本的情感分类的时候，影响了分类的准确率。基于情感的文本分类主要以具有语义倾向的词为主，选择具有情感倾向的词作为特征项，能够有效地提高分类的精度。

选择以单个词性作为特征，大量地缩小了文本特征项的数量，去除了很多干扰文本情感的主题词，改善了分类的效果。同时，名词和动词作为特征的分类精度要比副词和形容词的结果好。但是以四种词性作为特征的分类精度比任何一种单一词性的结果要好。选择单一的词性作为特征时，特征数量较少，不能反映整

个文档的情感倾向。结合名词、动词、形容词、副词作为特征,提高了分类精确率。但是单纯考虑名词、动词、形容词、副词的效果也不理想,主要原因是论坛语料的非常规性导致了其他词性的重要性加强,如一些叹词、状态词也具有感情色彩。语料的特殊性需要更多的词性词语作为特征,选取更多的可能带有情感倾向的特征能够提高分类的精度。但是,汉语的词汇量很丰富,同一词性的词有的具有情感倾向,有的只是客观性的陈述。选择更多的词性词语,引入大量的无关项,影响最终的分类精度。尽管最大熵模型在对情感词的识别中精确度不是很高,主要是对特征参数的估计欠缺。但是在对文档的整个情感分类中,分类效果较好。基于最大熵识别的特征表示,通过最大熵模型,选择具有语义倾向的词作为特征项,排除了大量无关的主题词,大大地缩小了特征项的规模。同时,解决了单一词性不能完全反映整个文档的情感特征问题。具有情感倾向的词都可以作为特征项,综合多个词性的词。根据文中的情感词分类模型,能够针对 BBS 相关评论做出较好的判断,区分哪些是正面的哪些是负面的,态度是赞成还是反对,从而决定是否值得推荐此文。

4　结束语

针对基于情感的文本分类的特点,通过最大熵模型识别文本中具有情感倾向的词,从而优化了文本的特征表示,提高了情感分类的精度。但是由于训练数据数量的限制,对于特征的选择以及特征参数的估计时有欠缺,影响了最大熵模型对情感词的识别。尽管如此,通过最大熵识别情感词作为特征表示的基础上,对 BBS 语料进行分类取得了较好的效果。单从词是否具有语义倾向考虑,未考虑词的共现对于情感倾向的改变,未考虑带有否定倾向词的疑问句表达的是肯定的情感,以及同一词语在不同语境中的语义倾向不同,影响了对文档最终的分类。下一步的工作,不断积累训练语料,改善最大熵模型。考虑不同词的共现组合对于情感倾向的改变,以及各种特殊句型所表达的情感差别。进一步完善训练语料,挖掘更多的最大熵特征。

参考文献:

[1] 董　梅,胡学钢. 基于多特征选择的中文文本分类[J]. 计算机技术与发展,2007,17(7):117 - 119.

[2] 马忠宝,刘冠蓉. 基于支持向量机的中文文本分类模型研究[J]. 计算机技术与发展,2006,16(11):70 - 72.

[3] 谢春丽,崔志明. web 文本挖掘中特征提取的设计与实现[J]. 微机发展(现更名:计算机技术与发展),2003,13(2):72 - 77.

[4] Pang Bo, Lee Lillian, Vaithyanathan S. Thumbs up Sentiment classification using machine Learning techniques [C]//In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing(EMNIJP). NJ, USA: Association for Computational Linguistics Morristown, 2002:79 - 86.

[5] Lin Wei - Hao, Wilson T, Wiebe J, et al. Which side are you on? Identifying Perspectives at the Document and Sentence Levels[C]//In : Proceedings of the 10th Conference on Computational Natural Language Learning(CoNLLX). New York: [s. n.], 2006:109 - 116.

[6] 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.

[7] 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.

[8] 林　红,苑春法,郭树军. 基于最大熵方法的汉语词性标注[J]. 计算机应用,2004,24(1): 14 - 16.

[9] 陈笑蓉,秦　进. 基于最大熵原理的汉语词义消歧[J]. 计算机科学,2005,32(5):174 - 176.

[10] Lan Man, Tan Chew - Lim, Low Hwee - Boon, et al. A Comparative study on Term Weighting Schemes for Text Categorization with support vector machines [C]//International World Wide Web Conference. New York, NY, USA: [s. n.], 2005:1032 - 1033.

(上接第 119 页)

2008 International Conference on Internet Computing in Science and Engineering. Washington, DC, USA: IEEE Computer Society, 2008:195 - 200.

[5] 袁　柳,李战怀,陈世亮. 基于本体的 Deep Web 数据标[J]. Journal of Software, 2008, 19(2):238 - 245.

[6] 苏志华,杨冬青,唐世渭,等. 基于结构分析和实体识别的信息集成[J]. 计算机研究与发展,2004,41(10):1823 - 1827.

[7] Raghavan S, Garcia - Molina H. Crawling the Hidden Web [C]// the International Conference on Very Large Data Bases (VLDB). Rome: Morgan Kaufmann Publishers, 2001:129 - 138.

[8] 郑冬冬,崔志明. Deep Web 查询接口选择[J]. 计算机应用,2006,26(9):2025 - 2027.