情 报 学 报 ISSN 1000－0135
第29卷第5期931－938，2010年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000－0135
Vol.29 No.5，931－938 October 2010

doi：10.3772/j.issn.1000-0135.2010.05.023

# Web文本情感分类研究综述[1)]

王洪伟[1] 刘 勰[1] 尹 裴[1] 廖雅国[2]
（1. 同济大学经济与管理学院，上海 200092；2. 香港理工大学电子计算学系，香港）

**摘要** 对用户发表在Web上的评论进行分析，能够识别出隐含在其中的情感信息，并发现用户情感的演变规律。为此，本文对Web文本情感分类的研究进行综述。将情感分类划分为三类任务：主客观分类、极性判别和强度判别，对各自的研究进展进行总结。其中将情感极性判别的方法分为基于情感词汇语义特性的识别和基于统计自然语言处理的识别方法。分析了情感分类中的语料库选择和研究难点。最后总结了情感分类的应用现状，并指出今后的研究方向。

**关键词** Web文本 情感分类 综述 主观性文本

## Literature Review of Sentiment Classification on Web Text

Wang Hongwei[1], Liu Xie[1], Yin Pei[1] and Liu N. K. James[2]
(1. *School of Economics and Management, Tongji University, Shanghai* 200092;
2. *Department of Computing, Hong Kong Polytechnic University, Hong Kong*)

**Abstract** Analyzing the users' reviews on the Web can help us to identify users' implicit sentiments and find the evolution laws of their emotion. To this end, this paper is a survey about the sentiment classification on the Web text. We divided the process of classification into three categories: subjective and objective classification, polarity identification and intensity identification and respectively summarize the resent research achievements in these fields. We also sorted the methods of polarity identification into two types: one is based on the emotional words with semantic characteristics, while the other statistic methods of natural language processing. What is more, the choice of corpus and potential research problems are discussed. At last, this paper summarized the status quo of application and pointed out the direction of future research.

**Keywords** Web texts, sentiment classification, survey, subjective text

随着互联网的流行，Web文本成为我们获取信息、发表观点和交流情感的重要来源。特别是随着Web2.0技术的发展，网络社区、博客和论坛给网络用户提供了更宽广的平台来交流信息和表达意见。这些文章和言论往往包含有丰富的个人情感，比如对某部大片的影评，对某款手机的用户体验等，其中蕴含着巨大的商业价值。如何从这些Web文本中进行情感挖掘，获取情感倾向已经成为当今商务智能领域关注的热点。所谓情感分析（sentiment analysis），就是确定说话人或作者对某个特定主题的

收稿日期：2009年6月29日

作者简介：**王洪伟**，男，1973年生，博士，副教授/博导，研究方向：本体建模和情感计算，E-mail：hwwang@tongji.edu.cn。**刘勰**，男，1985年生，硕士研究生，研究方向：数据挖掘与情感计算。**尹裴**，女，1986年生，硕士研究生，研究方向：商务智能。**廖雅国**，男，1954年生，博士，教授，研究方向：人工智能与电子商务。

1） 本文得到国家自然科学基金项目（70501024，70971099）；教育部人文社会科学资助项目（05JC870013）；上海市重点学科建设项目（B310）；香港研究资助局项目（polyU5237/08E）资助。

态度。其中,态度可以是他们的判断或者评估,他们(演说、写作时)的情绪状态,或者有意(向受众)传递的情感信息。因此,情感分析的一个重要问题就是情感倾向性的判断,即判断作者的观点是褒义的、积极的,还是贬义的、消极的。这类问题也被称为情感分类(sentiment classification)。

## 1 文本情感分类概述

在已有的研究中,情感分类也被称为意见挖掘(opinion mining)[1,2]。为了表述一致,本文统称为情感分类。情感分类涉及多个领域,如自然语言处理、人工智能、自动文本分类、文本挖掘、心理学等。它不同于传统的基于主题自动文本分类,后者分类的依据是文本的主题,如属于军事类还是体育类,而情感分类主要用来判别自然语言文字中表达的观点、喜好以及与感受和态度等相关的信息[3]。由于Web文本是以非结构化形式存在的,因此对文本进行情感分类是一个复杂的过程,包括:主客观文本分类、情感极性判别、情感强度判别。前者是情感分类的预处理工作,后两者才是真正意义上的情感分类。为了避免混淆,我们将后两者统称为情感识别(见图1)。

图1描述了从原素材到得出情感结果的整个情感分类过程。其中,原素材中的文本可以是句子或者是整篇文章,它们所对应的分类任务分别为句子情感分类和文档情感分类。为了减少干扰,提高情感分类的精度,首先要对文本进行主观性识别,即主客观文本分类。只有带有主观色彩的文本才会蕴含着作者的情感,所以情感识别的对象是主观文本。情感识别分为极性判别和强度判别两个任务。极性分类是识别主观文本的情感是正面的赞赏和肯定还是负面的批评与否定。而强度判别则是判定主观文本情感倾向性强度,比如强烈贬抑、一般贬抑、客观、一般褒扬、强烈褒扬五个类别。

在整个情感分类过程中,还涉及分类前的预处理技术,包括分词、词性标注、平滑、停用词和缩词的处理等语言处理技术,这些技术相对成熟,不再赘述。下面从主客观文本分类和情感识别两个方面来总结情感分类的研究现状。

## 2 主客观文本分类现状

所谓"主观性"是指在自然语言中用来表达意见和评价的语言特性[4]。主观性文本表达的是说话者对某人、某物或某事的态度和看法,包含个人的主观情感色彩。与之相对应的客观性文本则描述客观存在的事实,说话者往往持有中立和客观的情感。在表述上,主客观文本也有明显的差异,客观性文本通常采用比较正式的陈述句,而主观性文本因为强调自我表达,表述上比较自由,偏口语化,比如"这款手机酷毙啦!"。

主客观文本分类研究已经展开,并应用在信息检索和信息抽取等领域[5]。主客观文本分类与其他文本分类类似,可以从篇章、句子和词语三个层面展开,用到的方法主要是机器学习算法。

Wiebe等很早就对主客观文本分类问题进行了研究[4~11]。Wiebe和Bruce将某些词类(代词、形容词、基数词、情态动词和副词)、标点和句子的位置作为特征值,设计了针对句子级别的NB分类器[6]。在此基础上,Wiebe[5]又将某些词性和基于词典的语义词作为特征项,显著提高了分类器的分类效果。Wiebe和Wilson还针对基于篇章层面的分类方法进行了研究[7]。通过计算每篇文档中出现的主观性词语数量,用KNN分类器来判断篇章的主客观性,取得了较好效果。

Yu等利用三种统计方法进行主客观句的识别研究,包括相似性方法、NB分类和多重NB分类。其中NB分类器在原有研究的基础上采用词、2-gram、3-gram和词类、具有情感倾向的词序列、主语

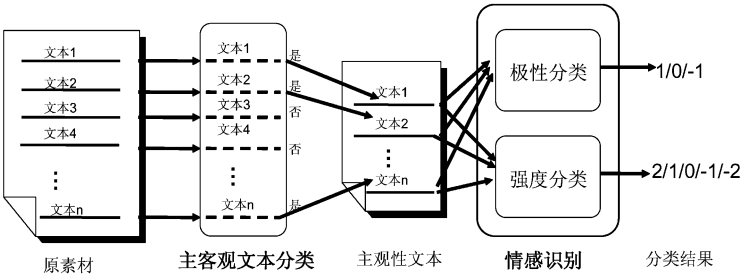


**图1 情感分类的主要过程**

和其直接修饰成分等作为特征项,对主观句识别的查准率和查全率达到了 80%～90%[12]。

Pang 和 Li 将句子间的情感联系作为分类的一个重要因素,用最小图割(Minimum cuts)的方法来寻找上下文语句的关系以提高分类精度。它的划分原理是是使成本公式最小:$\sum_{x\in C_1} ind_2(x)+\sum_{x\in C_2} ind_1(x)+\sum_{x_i\in C_1,x_k\in C_2} assoc(x_i,x_k)$,其中 $x$ 句子,$C_i$ 是类别,$ind_j(x_i)$ 指单根据 $x_i$ 的特征将其划分为 $j$ 类的偏好得分,$assoc(x_i,x_k)$ 指 $x_i$ 和 $x_k$ 属于同一类的得分[13]。

中文语境下主客观文本分类具有一定的复杂性,而且对中文主观性文本的判别起步较晚,大多数情感分析研究都是人为抽取主观性文本。

林斌将影视内容介绍和影视评论分别视为客观文本和主观文本,采用互信息量(MI,mutual information)计算影视评论中每个词语的互信息量,并由大到小排序,取最靠前的 275 个词语,并将它们两两组合,再计算每对组合在影视评论中的互信息量,最后得到“我想”“我应该”等具有主观倾向的 75 个词语组合,并将其用于句子主客观性的判断,总体的准确率达到了 78.42%[14]。

叶强和张紫琼等提出一种根据连续双词词类组合模式(2-POS)自动判别句子主客观性程度的方法。首先在 N-POS 语言模型的基础上,利用 CHI 统计方法提取中文主观文本词类组合模式,利用这些组合模式给每个句子赋以主观性得分,将得分高于设定阈值的句子判定为主观性文本。实验表明,当阈值为 0.12 时,主观文本的分类查准率和查全率能达到 76%[15]。

需要指出,由于中英文语言结构及中西方文化的差异,使得中文的情感流露方式具有特殊性和复杂性,这给中文文本的情感分析带来挑战。与英文文本多都应用机器学习不同,中文文本的主客观分类主要采取语义方法,而且分类效果也不够理想,相比于英文能达到 90% 左右的精度,中文的研究分类精度还不够高。这主要由于影响中文文本主客观判断的因素远远比英文多而复杂,除了词义、词性之外,词语的用法也会影响到文本的主客观性质。因此在今后中文文本主观性判别研究中,除了引入机器学习算法外,还要注意考虑中文词法和句法的特殊功能。

另外,一些研究将主客观分类和褒贬情感分类同时看作三分类问题,将文本分成为“褒义”、“贬义”、“客观”。前两类归为主观文本,后者视为客观文本。王根和赵军指出这种观点忽略了两个任务所用特征的不同,即将主客观和褒贬极性的特征夹杂在一起,影响了分类效果[16]。本文认为,主客观分类中的“客观”类和情感分析中的“客观”类是两个不同概念。比较下面两句话:“这部电影耗资两亿,将于明天在上海万达影城上演首映”;“这部电影整体上还算四平八稳,跟我的预期有点差距,但也不算失望”。前一句是陈述客观事件,是客观文本。而后一句显然是作者的主观评价,却不带有明显的褒或贬。因此对它的分类过程是:首先将其归为主观性文本,然后通过情感分析再归为情感类别中的“客观”(或“中立”)类。所以,非褒非贬并不是作者没情感,而是情感倾向并不明显,持中立态度。如果将双分类任务看成一个多分类问题的话,会错误地把带有主观性但情感倾向不明显的文本分类为客观性文本,影响情感分类的科学性。为了避免混淆,在后面的表述中,本文将情感分类结果中的非褒非贬统称为“中立”类。

## 3 情感识别现状

### 3.1 文本情感极性研究

#### 3.1.1 基于情感词汇语义特性的识别

基于情感词汇语义特性的识别是指利用词语的感情色彩来判断文本的情感极性,主要有两种研究方法:计算词语情感得分[17,18]和构造情感词[19~21]。

(1)计算词语情感得分的方法

Turney 提出基于情感词组的 SO-PMI 语义分类方法[17]。该方法提取出符合一定模式的形容词或副词双词词组作为情感词组。并定义逐点互信息量(PMI,Pointwise Mutual Information)来计算词 $w_1$ 和 $w_2$ 之间的语义相关性:$PMI=\log_2\left[\frac{p(w_1,w_2)}{p(w_1)p(w_2)}\right]$。$p(w_1,w_2)$ 表示 $w_1$ 和 $w_2$ 同时出现的概率。计算抽取出的词组与情感词“excellent”和“poor”的 PMI,并用 SO(semantic opinion orientation)来计算该词组的语义倾向性:$SO(w)=PMI(w,\text{“excellent”})-PMI(w,\text{“poor”})$,这样就确定了每个情感词组的情感倾向。最后通过计算评论中所有提取出的情感词组的平均 SO 值来区分情感极性。如果该值大于零,表示好评,推荐该评论描述的对象,如果小于零则不推荐。

Dave 和 Lawrence 等对语料中的每个词打分：$score(t_i)=\left[\frac{P(t_i \mid C)-P(t_i \mid C')}{P(t_i \mid C)+P(t_i \mid C')}\right]$，$t_i$ 是一个词，$C$ 是一个情感类型。然后计算文档中所有词的得分总和，根据下式来识别新文档的类别：$class(d_j)=\begin{cases} C & eval(d_j)>0 \\ C' & otherwise \end{cases}$，其中文档 $d_j=t_1 \cdots t_n$，$eval(d_j)=\sum_i score(t_i)$[18]。

(2)构造情感词的方法

还有一些工作尝试建立情感词典来匹配文档的情感。Tong 手工建立了一本针对影评分类的情感词典[19]。首先人工抽取出影评相关的情感词汇(比如“great acting”，“wonderful visuals”，“uneven editing”)。同时对每一个情感词汇按其所代表的情感倾向(“positive”或“negative”)进行人工标记，并加到专门的情感词典，最后利用这个词典去判断影评的情感态度。但是该方法建立的情感词典往往是面向特定领域的，每一个分析对象都需要构建一本词典。而 Hu 和 Liu 在手工建立的已知 positive 和 negative 的种子形容词词汇表的基础上，利用 WorldNet 中词间的同义和近义关系来判断新情感词的语义倾向，并以此判断观点的情感极性[20]。

建立情感词来判别文本情感的方法存在两个问题：①基于词典的识别方法以分析词汇情感为基础，但忽略了句子中否定词对情感的影响，造成句子级别和文档级别的分类精确度下降。②所选的情感词往往是情感特征比较强烈的词语(主要是形容词和副词)，而还有一些词汇往往隐含着说话人正面或负面的情绪。比如“爆炸”、“车祸”隐含了较多负面情感，而“舞会”、“打折”往往表达了正面情感。

为了克服单一靠人工建立的词汇在情感解释力上的缺乏性，Liu 等使用 Open Mind Common Sense 对人类通用情感进行学习和解释[21]。Open Mind Common Sense 是一个常识知识库，可以用来对客观世界中的事件、行为、对象进行通用的情感推理。首先从知识库中选出典型的六类情感词汇(高兴、悲伤、愤怒、恐惧、厌恶和惊奇)，然后根据知识库中的概念关系对其他概念进行情感赋值。比如，知识库有这么两句话：“发霉的面包很恶心”，“新鲜的面包很美味”。那么，在“恶心”和“美味”分别被归类为厌恶和高兴的基础上，修饰语言模型(Modifier Unigram Model)可以分别将发霉和新鲜这两个修饰语也判断为表示厌恶和高兴的概念。

(3)中文文本研究现状

在中文文本识别方面，用情感词汇来判断文本情感的方法相对较少。金聪等将 Turney 的 PMI-SO 方法应用到对中文语料的情感判断上，同时用典型文档的语义倾向值的平均值作为阈值来代替零值作为两级情感的分类界限，改善了分类效果[22]。

李钝从语言学角度出发，分析词典中词对语义的特点，采用“情感倾向定义”权重优先的方法计算短语中各词的语义倾向度，然后分析短语中各词组合方式的特点，提出中心词概念来对各词的倾向性进行计算，以识别短语的倾向性和倾向强度。实验表明，该方法对短语的倾向分类识别效果较好，可为更大粒度的文本倾向识别打好基础[23]。

### 3.1.2 基于统计自然语言处理的识别方法

基于统计自然语言处理的方法，是指利用机器学习算法对统计语言模型进行训练，最后用训练好的分类器对新文本进行识别。

一些研究将基于主题的机器分类算法用于情感极性识别。Pang 和 Li 等采用不同的特征选择方法，应用了 NB、ME(Maximum Entropy)、SVM 对电影评论进行分类[24]。在他们的另一项工作中，将文本极性分类问题转换成求取句子连接图的最小分割问题，实现了一个基于 minimum-cut 的分类器[13]。Ni 等利用 CHI 和信息增益进行特征选择，并采用 NB、SVM 和 Rocchio's 算法对情感分类[25]。Mullen 等和 Whitelaw 等都用到 SVM 算法，只是他们在特征的选择和处理上不同[26,27]。Cui 等利用 PA(Passive-Aggressive)、LM (Language Modeling)和 Winnow 分类器，并比较了它们的性能[28]。下面从特征选择和算法性能两个方面对基于机器学习算法的情感识别进行小结：

(1)特征选择

Pang 等在实验中分别使用以词频作为权重的 Unigrams、以布尔值作为权重的 Unigrams、Bigrams、Unigrams+Bigrams、Unigrams+词性、最前面 2633 的 Unigrams、形容词、Unigrams+词语的位置作为其语言特征[24]。实验结果发现，使用布尔值 Unigram 作为特征的分类效果最好，使用 Bigram、词性、形容词和词语的位置作为特征并不能达到预期的分类精度。而 Cui 等指出 Pang 的研究语料较小，无法体现出 n-grams($n \geq 3$)的优势[28]。他们对比了 $n$ 分别等于 1、2、3、4、5、6 时的实验结果，发现当 $n=6$ 时，分类效果最好，特别是识别负面(negative)文本的分类器的

准确度有明显改善,达到 70.03%。Mullen 等将按 Turney 的情感词五种组合模式提取出来的词组称为价值词组(value phrases),然后利用 WorldNet 计算出所有形容词的 EVA、POT 和 ACT 值,并将这三个值和价值词组的 SO 值一起作为特征,最后再用 SVM 分类器进行分类,实验结果表明该方法的分类效果也是好于以前的方法[17,26]。另外,Pang 等的实验结果还表明使用布尔值作为特征值权重的比使用词频为权重的实验精度要高[24]。

(2)分类性能比较

Cui 等的实验对比表明,平均表现最好的是 PA 分类器,Winnow 次之,LM 最差[28]。Pang 等的研究表明,基于机器学习的分类器要比手工分类效果好很多,而在三类分类器中,SVM 分类器的表现比 ME 和 NB 都好,但是实验结果同时还表明,对文本的情感分类效果还是远差于对文本主题的分类[24]。

国内方面,徐军等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 NB 的分类效果明显好[29]。

唐慧丰等对部分基于监督学习的中文情感分类技术做了比较研究,在文本特征方面,采用 N-Gram 以及名词、动词、形容词、副词作为不同的文本表示特征;以互信息、信息增益、CHI 统计量和文档频率作为不同的特征选择方法;以中心向量法、KNN、Winnow、NB 和 SVM 作为不同的文本分类方法;并在不同的特征数量和不同规模的训练集情况下,分别进行了中文情感分类实验[30]。实验结果表明:采用 Bigram 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。

### 3.2 文本情感强度研究

对于某些应用,单纯的褒贬分类是不够的,还需要区别褒贬情感的强弱。这种任务称为情感强度分类,它是一种特殊的分类问题,因为强弱分类的类别是离散且有等级的。文本情感强度分析主要有三类方法:多分类方法,回归方法,序列标注方法。

(1)多分类方法

多分类方法即将文本的每个强度等级当作一个类别,构造分类器对其分类。最常见的处理是将文本强度分成强烈贬抑、一般贬抑、客观、一般褒扬、强烈褒扬五个类别。Lin 等在研究语料的观点问题时,采用 LSPM(Latent Sentence Perspective Model)对未经标注的语句的观点及其五类强度进行判断[31]。但是此类方法得到的结果往往忽略了情感渐变过程,造成训练模型不够准确,影响了分类精度[16]。

(2)回归方法

回归方法即用回归算法来对文本的强度进行拟合。Pang 和 Li 就用了 SVM 回归方法对文本情感强度进行了回归评分[32]。此外,他们还根据相似度越高标记越相近的原理,提出一种基于度量标记(metric labeling)的元算法(meta-algorithm)对文本进行评分,实验表明此方法的效果比多分类方法和 SVM 回归方法都好。

(3)序列标注方法

近年来,条件随机场(Condition Radom Fields,CRFs)模型大量地应用于序列标注任务,比如 Chunking,NER,Parsing 等。同时,CRFs 模型也逐步应用于文本倾向性分析任务,并以此产生出针对特定问题的基于 CRFs 模型的其他图模型方法[33]。Mao 和 McDonald 等把句子的褒贬标记看作一个情感流问题,并利用序列 CRFs 回归模型来给篇章中的每个句子进行打分[34,35]。为了减轻褒贬度分析中信息冗余对强度分类的影响,刘康等在 CRFs 的框架下,考虑句子褒贬度与褒贬强度之间的层级关系,充分利用上下文的信息以及特征的层级特性,提出了基于层叠 CRFs 模型的句子褒贬度分析模型[33]。

总体而言,对文本进行情感强度的研究还不多,但是在电子商务网站中,情感强度的识别对于个性化客户服务来说可能更有意义。

## 4 其他相关问题

### 4.1 语料库的选择

不管是基于语义的方法还是机器算法都需要大规模的情感语料支撑,目前常用的语料库有以下几类:

(1)评论类语料

评论类语料库是目前最常用的语料数据,包括影评[17,24]、产品评论[13,28]、音乐评论[26]等。众多研究选择评论类语料是因为:一是评论类语料数量众多,方便获得。随着电子商务、网络社区的发展,到处都是关于电影、产品的评论,这些数据为文本分类研究提供了充足的素材;二是可近似地将评论类语

料视作主观性文本,无需主客观文本分类即可直接用来情感识别。

(2)词汇知识库

WorldNet是一个按语义关系网络组织的英文词库,多种词汇关系和语义关系被用来表示词汇知识的组织方式。有许多研究就直接利用WorldNet中词汇间的距离来揭示情感倾向的关系[36,37]。Liu等使用的Open Mind Common Sense是常识知识库,它描述了世界上最基本的概念和相关的关系,以此来扩展描述情感的"概念"[21]。作为最大的中文词汇知识库,HowNet为国内研究者提供了进行中文情感研究的渠道[38,39],主要有基于语义相似度的方法和基于语义相关场的方法。

(3)其他语料库

徐琳宏等综合现有的各种情感词汇资源构造情感词汇本体,他们采用手工分类和自动的方法来获取本体的知识,包括词汇的情感类别、强度和极性[40]。在此基础上,对基于本体的情感分类方法进行了研究。路斌等利用中文同义词词林来计算词汇褒贬,该方法利用同义词词林中的同义词词群,将种子词汇扩展得到更大的褒贬义词集合[41]。

### 4.2 研究难点

由于Web文本的表述形式多样,没有统一的规范,给文本挖掘和情感分类带来了许多困难。另外,随着研究的细化,Web文本挖掘和情感分析的任务也不仅仅限于判定情感的倾向和强度。本文将情感分析中的其中几个难点罗列如下:

(1)网络用词

主观性文本往往口语化,甚至会频繁出现时髦的网络用语,比如"做人不能太CNN!",在这里,"CNN"无疑是识别这句话情感倾向的关键。不断涌现的网络用语给情感分类提出了更高的时效性要求。又如"太BS这部手机的性价比了。"BS是贬义词"鄙视"的缩写,这句话表达的是负面情感。可以看出,缩词的使用也给情感分析的准确度带来了很大影响。

(2)表达方式

除了情感词汇之外,句子的表达方式也会对句子的褒贬情感产生巨大影响。比较下面两句话:"这瓶洗发水,适合头发很干的人用。""用了这瓶洗发水,头发会变得很干。"这两个句子的用词差不多,"洗发水","头发","很干"。但是第一句是褒义,第二句则很可能是贬义。还有一类和表达方式有关的问题是"反话"。很多褒义词受论坛文化的影响,有往贬义发展的趋势,比如"您太有才了"等。让机器理解这些表达方式也是情感分析面临的一个挑战。

(3)关系抽取

识别情感和特定主题的关系是情感分析的又一任务。比如这样一条评论:"Sony笔记本的外观蛮好看,就是价格太贵了"。在面向电子商务的情感分析中,就需要识别顾客对Sony外观和价格所持有的不同情感。另外同样的词语由于描述的对象不同,表达的情感也会不同,比如在产品评论中,同样是"少",在描述"价格少"时是一种褒义、积极的情感,而"种类少"却是是一种贬义、消极的情感。这些问题就涉及识别情感所描述的对象问题,即关系抽取。

## 5 研究展望

情感分类过程主要有上下承接的两个任务:主客观文本分来和情感识别。总体上,随着自然语言处理技术的发展,国内外的研究已经取得了不小的成果,在评论分析、个性化推荐和舆论监控等方面也得到了应用,如Dave研发的Review Seer是世界上第一个情感分析工具,也是第一个针对产品评论区别其褒贬性的系统[42]。Gamon等开发的Pulse系统可自动挖掘网上用户对汽车评价中的贬褒信息和强弱程度[43]。Liu等开发的Opinion Observer系统可以处理网上顾客对产品的评价,对涉及产品各种特征的优缺点进行统计,并采用可视化方式对产品特征的综合质量进行比较[44]。但这些应用仍不尽如人意,今后还需在以下几方面展开深入研究:

1)相比于面向主题的文本分类,情感分类的精度还比较低,主要由用词习惯、表达方式等问题造成的。如何提高情感分类的准确率和召回率,是今后的研究重点。

2)利用文本的更多特征来提高情感分类效果。如标点符号对文本的情感表达有特殊作用,"!"号和"?"的情感强度明显强于其他标点,但目前的研究中还鲜有考虑。

3)将情感分析分成主客观文本分类和情感分类两个步骤,甚至在此基础上还要进行强度分类,这会造成冗余问题:前一步骤的误差同时会影响到后面步骤的准确度。考虑冗余问题的情感分析建模也是今后的研究方向之一。

4)除单纯地识别情感态度和强度,文本情感分类还需与其他文本挖掘技术结合,实现情感和情感

对象的关系抽取,挖掘出比单独的褒或贬的情感倾向更有价值的信息,以提高情感分类的应用价值。

## 参 考 文 献


[1] Pekar V, Ou S. Discovery of subjective evaluations of product features in hotel reviews[J]. Journal of Vacation Marketing, 2008, 14(2): 145-155.

[2] 姚天,程希文,徐飞玉,等. 文本意见挖掘综述[J]. 中文信息学报, 2008, 22(5): 71-80.

[3] 陈博. Web 文本情感分类中关键问题的研究[D]. 北京:北京邮电大学博士论文, 2008.

[4] Wiebe J. Tracking point of view in narrative [J]. Computational Linguistics, 1994, 20(2): 233-287.

[5] Wiebe J. Learning subjective adjectives from corpora[C]// Proc. of the 17th National Conf. on Artificial Intelligence (AAAI-2000). Texas, USA, 2000.

[6] Wiebe J, Bruce R, O'Hara T. Development and use of a gold standard dataset for subjectivity classifications[C]// Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), Seattle, USA, 1999: 246-253.

[7] Wiebe J, Wilson T, Bruce R, et al. Learning subjective language[R]. Technical Report TR-02-100, Pennsylvania, USA, 2002.

[8] Bruce R, Wiebe J. Recognizing subjectivity: A case study in manual tagging [J]. Natural Language Engineering, 1999, 5(2): 187-205.

[9] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase level sentiment analysis[C]// Proc. of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP2005). Vancouver, Canada, 2005: 347-354.

[10] Hatzivassiloglou V, Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity[C]// Proc. of 18th Int'l Conf. on Computational Linguistics (COL ING-2000). NJ, USA, 2000.

[11] Wiebe J, Wilson T, Bell M. Identifying collocations for recognizing opinions[C]// Proc. of the ACL-01 Workshop on Collocation. Toulouse, France, 2001.

[12] Yu H, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences[C]// Proc. of the 2003 Conf. on Empirical Methods in Natural Language Processing. Sapporo, Japan, 2003: 129-136.

[13] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based onminimum cuts[C]// Proc. of the 42nd Meeting of the Association for Computational Languages. Barcelona, Spain, 2004: 271-278.

[14] 林斌. 基于语义技术的中文信息情感分析方法研究[D]. 哈尔滨:哈尔滨工业大学硕士论文, 2007.

[15] 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[J]. 信息系统学报, 2007, 1(1):79-91.

[16] 王根,赵军. 基于多重冗余标记 CRFs 的句子情感分析研究[J]. 中文信息学报, 2007, 21(5): 51-55.

[17] Turney P. Thumbs Up or Thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]// Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. NJ, USA, 2002: 417-412.

[18] Dave K, Lawrence S, Pennock D. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews[C]// Proc. of the 12th Int'l World Wide Web Conf.: ACM Press. Budapest, Hungary, 2003: 519-528.

[19] Tong R M. An operational system for detecting and tracking opinions in on-line discussion[C]// SIGIR Workshop on Operational Text Classification. NY, USA, 2001:1-6.

[20] Hu M, Liu B. Mining and summarizing customer reviews[C]// Proc. of Knowledge Discovery and Data Mining, NY, USA, 2004: 168-177.

[21] Liu H, Lieberman H, Selker T. A model of textual affect sensing using real-world knowledge[C]// Proc. of the 11th Int'l. Conf. on Intelligent User Interface, 2003: 125-132.

[21] 金聪,金平. 网络环境下中文情感倾向的分类方法[J]. 语言文字应用, 2008,5(2): 139-144.

[23] 李钝. 基于短语模式的文本情感分类研究[J]. 计算机科学, 2008, 135(14): 231-233.

[24] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques[C]// Proc. of the Conf. on Empirical Methods in Natural Language Processing. Philadelphia, US, 2002: 79-86.

[25] Ni X, Xue G, Ling X, et al. Exploring in the Weblog space by detecting informative and affective articles[C]// Proc. of the 16th Int'l. Conf. on World Wide Web, 2007: 281-290.

[26] Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources[C]// Proc. of EMNLP-2004, Barcelona, Spain, 2004: 412-418.

[27] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis[C]// Proc. Of the 14th ACM Int'l. Conf. on Information and Knowledge Management, 2005: 625-631.

[28] Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews[C]// Proc. of the 21st National Conf. on Artificial Intelligence

(AAAI-06), Boston, USA, 2006.

[29] 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.

[30] 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.

[31] Lin W H, Wilson T, Wiebe J, et al. Which side are you on? Identifying perspectives at the document and sentence levels[C]// Proc. of the 10th Conf. on Computational Natural Language Learning (CoNLL-X), NY, USA, 2006：109-116.

[32] Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales[C]// Proc. of the 43rd Annual Meeting of the Association for Computer Linguistics. Morristown, NJ, USA, 2005：115-124.

[33] 刘康，刘军. 基于层叠CRFs模型的句子褒贬度分析研究[J]. 中文信息学报，2008，22(1)：123-128.

[34] Mao Y, Lebanon G. Isotonic Conditional Random Fields and Local Sentiment Flow[C]// Proc. of the Neural Information Processing Systems, 2007.

[35] McDonald R, Hannan K, Neylon T, et al. Structured models for fine-to-coarse sentiment analysis[C]// Proc. of ACL, 2007：432-439.

[36] Kamps J, Marx M, Mokken R J, et al. Words with Attitude[C]// Proc. of the 1st Int'l Conf. on Global WordNet. Mysore, India, 2002.

[38] 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.

[39] 徐琳宏，林鸿飞，杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报，2007，21(1)：96-100.

[40] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 中文信息学报，2008，22(2)：180-185.

[41] 路斌，万小军，杨建武，等. 基于同义词词林的词汇褒贬计算[C]//第七届中文信息处理国际会议，武汉，中国，2007：17-23.

[42] Dave K, Lawrence S, Pennock D M. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews[C]//Proc. of the 12th Int'l. WWW Conf. Budapest, Hungary, 2003.

[43] Gamon M, Aue A. Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms[C]//Proc. of the ACL-2005 Workshop on Feature Engineering for Machine Learning in NLP. Michigan, USA, 2005：57-64.

[44] Liu B, Hu M, Chen J. Opinion Observer: Analyzing and Comparing Opinion on the Web[C]//Proc. of the 14th Int'l. Conf. on World Wide Web. Chiba, Japan, 2005：343-351.

(责任编辑 芮国章)

# Web文本情感分类研究综述

万方数据 WANFANG DATA 文献链接

作者： 王洪伟， 刘勰， 尹裴， 廖雅国
作者单位： 王洪伟,刘勰,尹裴(同济大学经济与管理学院,上海,200092)， 廖雅国(香港理工大学电子计算学系,香港)

刊名： 情报学报 ISTIC PKU CSSCI
英文刊名： JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC ANDTECHNICAL INFORMATION
年，卷(期)： 2010，29(5)
被引用次数： 0次

## 参考文献(43条)

1. Pekar V,Ou S.Discovery of subjective evaluations of product features in hotel reviews[J].Journal of Vacation Marketing,2008,14(2):145-155.

2. 姚天,程希文,徐飞玉,等.文本意见挖掘综述[J].中文信息学报,2008,22(5):71-80.

3. 陈博.Web文本情感分类中关键问题的研究[D].北京:北京邮电大学博士论文,2008.

4. Wiebe J.Tracking point of view in narrative[J].Computational Linguistics,1994,20(2):233-287.

5. Wiebe J.Learning subjective adjectives from corpora[C]//Proc.of the 17th National Conf.on Artificial Intelligence (AAAI-2000).Texas,USA,2000.

6. Wiebe J,Bruce R,O'Hara T.Development and use of a gold standard dataset for subjectivity classifications[C]//Proc.of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99),Seattle,USA,1999:246-253.

7. Wiebe J,Wilson T,Bruce R,et al.Learning subjective language[R].Technical Report TR-02-100,Pennsylvania,USA,2002.

8. Bruce R,Wiebe J.Recognizing subjectivity:A case study in manual tagging[J].Natural Language Engineering,1999,5(2):187-205.

9. Wilson T,Wiebe J,Hoffmann P.Recognizing contextual polarity in phrase level sentiment analysis[C]//Proc.of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing(HLT/EMNLP2005).Vancouver,Canada,2005:347-354.

10. Hatzivassiloglou V,Wiebe J.Effects of adjective orientation and gradability on sentence subjectivity[C]//Proc.of 18th Int'l Conf.on Computational Linguistics(COL ING-2000).NJ,USA,2000.

11. Wiebe J,Wilson T,Bell M.Identifying collocations for recognizing opinions[C]//Proc.of the ACL-01 Workshop on Collocation.Toulouse,France,2001.

12. Yu H,Hatzivassiloglou V.Towards answering opinion questions:separating facts from opinions and identifying the polarity of opinion sentences[C]//Proc.of the 2003 Conf.on Empirical Methods in Natural Language Processing.Sapporo,Japan,2003:129-136.

13. Pang B,Lee L.A sentimental education:Sentiment analysis using subjectivity summarization based onminimum cuts[C]//Proc.of the 42nd Meeting of the Association for Computational Languages.Barcelona,Spain,2004:271-278.

14. 林斌.基于语义技术的中文信息情感分析方法研究[D].哈尔滨:哈尔滨工业大学硕士论文,2007.

15. 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[J].信息系统学报,2007,1(1):79-91.

16. 王根,赵军.基于多重冗余标记CRFs的句子情感分析研究[J].中文信息学报,2007,21(5):51-55.

17. Turney P.Thumbs Up or Thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]//Proc.of the 40th Annual Meeting of the Association for Computational Linguistics.NJ,USA,2002:417-412.

18. Dave K,Lawrence S,Pennock D.Mining the peanut gallery:Opinion extraction and semantic classification of product reviews[C]//Proc.of the 12th Int'l World Wide Web Conf.:ACM Press.Budapest,Hungary,2003:519-528.

19. Tong R M. An operational system for detecting and tracking opinions in on-line discussion[C]//SIGIR Workshop on Operational Text Classification. NY, USA, 2001: 1-6.

20. Hu M, Liu B. Mining and summarizing customer reviews[C]//Proc. of Knowledge Discovery and Data Mining, NY, USA, 2004: 168-177.

21. Liu H, Lieberman H, Selker T. A model of textual affect sensing using real-world knowledge[C]//Proc. of the 11th Int'l. Conf. on Intelligent User Interface, 2003: 125-132.

22. 金聪, 金平. 网络环境下中文情感倾向的分类方法[J]. 语言文字应用, 2008, 5(2): 139-144.

23. 李钝. 基于短语模式的文本情感分类研究[J]. 计算机科学, 2008, 135(14): 231-233.

24. Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques[C]//Proc. of the Conf. on Empirical Methods in Natural Language Processing. Philadelphia, US, 2002: 79-86.

25. Ni X, Xue G, Ling X, et al. Exploring in the Weblog space by detecting informative and affective articles[C]//Proc. of the 16th Int'l. Conf. on World Wide Web, 2007: 281-290.

26. Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources[C]//Proc. of EMNLP-2004, Barcelona, Spain, 2004: 412-418.

27. Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis[C]//Proc. Of the 14th ACM Int'l. Conf. on Information and Knowledge Management, 2005: 625-631.

28. Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews[C]//Proc. of the 21st National Conf. on Artificial Intelligence (AAAI-06), Boston, USA, 2006.

29. 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.

30. 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.

31. Lin W H, Wilson T, Wiebe J, et al. Which side are you on? Identifying perspectives at the document and sentence levels[C]// Proc. of the 10th Conf. on Computational Natural Language Learning (CoNLL-X), NY, USA, 2006: 109-116.

32. Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales[C]// Proc. of the 43rd Annual Meeting of the Association for Computer Linguistics. Morristown, NJ, USA, 2005: 115-124.

33. 刘康, 刘军. 基于层叠CRFs模型的句子褒贬度分析研究[J]. 中文信息学报, 2008, 22(1): 123-128.

34. Mao Y, Lebanon G. Isotonic Conditional Random Fields and Local Sentiment Flow[C]//Proc. of the Neural Information Processing Systems, 2007.

35. McDonald R, Hannan K, Neylon T, et al. Structured models for fine-to-coarse sentiment analysis[C]//Proc. of ACL, 2007: 432-439.

36. Kamps J, Marx M, Mokken R J, et al. Words with Attitude[C]//Proc. of the 1st Int'l Conf. on Global WordNet. Mysore, India, 2002.

37. 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.

38. 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1): 96-100.

39. 徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造[J]. 中文信息学报, 2008, 22(2): 180-185.

40. 路斌, 万小军, 杨建武, 等. 基于同义词词林的词汇褒贬计算[C]//第七届中文信息处理国际会议, 武汉, 中国, 2007: 17-23.

41. Dave K, Lawrence S, Pennock D M. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews[C]//Proc. of the 12th Int'l. WWW Conf. Budapest, Hungary, 2003.

42. Gamon M, Aue A. Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms[C]//Proc. of the ACL-2005 Workshop on Feature Engineering for Machine Learning in

NLP.Michigan,USA,2005:57-64.

43.Liu B,Hu M,Chen J.Opinion Observer:Analyzing and Comparing Opinion on the Web[C]//Proc.of the 14th Int'l.Conf.on World Wide Web.Chiba,Japan,2005:343-351.

## 相似文献(3条)

1.学位论文 陈博 WEB文本情感分类中关键问题的研究 2008

随着计算机技术和互联网的迅猛发展，网络在线的文档成为现代主要的信息载体，是人们生活中不可或缺的主要信息来源。而随着互联网进入web2.0时代，人们从被动的接受门户网站发布信息，转变为主动的获取、发布、共享、传播信息。同时，由于用户参与到信息的产生，网络信息的内容形式也变得多样化，越来越多的具有个人观点性的内容充斥着网络。这些观点性内容对于网络电子商务、网络社区发掘、网络信息安全、网络信息检索等多方面都具有重要的意义和实用价值。对网络文本观点性内容的自动情感分析成为近期web信息处理的一个研究热点，而其中的核心技术就是文本情感分类。

在这样一个背景下，本文对面向web文本的中文分词、文本情感分类以及Weblog观点检索问题进行了下述创新性研究工作：

首先，研究了面向web文本的中文分词问题。根据web文本环境的特点，研究重点在于中文分词中的未登录词识别问题，同时兼顾切分歧义消解、整体切分准确率和高效处理海量文本的能力。在未登录词识别方面，提出了POC-NLW字符标记模板，从字符级别的粒度来表征中文词汇的构成机制，并结合隐马尔可夫模型，实现了基于字符序列标注的中文分词方法。此外，分别使用了基于规则匹配的预处理、基于词典匹配的初级全切分、基于词语级别的N-Gram统计切分模型，并通过级联方式将上述各模块有效组合，构成了多模型混合的层叠系统。实验结果表明，本文提出的基于POC-NLW模板标注的切分方法具有较强的未登录词识别能力；而多模型混合的层叠系统在整体切分精度和未登录词识别方面都达到了较高的实用水平。另外，本文提出的系统还具有高效的建模和切分处理速度，具有面向海量web文本切分的实用性能。

第二，研究了web文本情感分类问题，主要包括文本的主客观分类和正负面极性分类两个子问题。在语言特征表示方面，对比研究了基于多种N-Gram语言特征模板的文本特征表示方式；在文本特征加权方面，对比了布尔、绝对词频、归一化词频以及基于TFIDF的特征加权方式；在特征选择方面，提出了全局TFIDF显著指数，引入“全局过滤．局部加权”的特征抽取方式；在情感分类模型方面，以朴素贝叶斯模型做对比，详细研究了最大熵模型的最大似然估计问题，采用高斯先验和指数型先验，对传统的最大熵模型进行改进。通过在真实网络电影评论数据集上的详细对比试验，以及对语料样本的分析，证实了采用高阶语言特征模板、基于TFIDF的特征选择和过滤方法、加入指数型先验的最大熵模型较好的适用于文本情感分类问题。

第三，研究了Weblog观点检索问题，以TREC Blog Track评测为主线，主要研究了面向blog文档的主题检索和文本情感分类技术在观点检索中的应用。首先，针对Weblog文档的特点以及观点检索的特殊性，在Weblog文档的HTML解析、噪声标签过滤、文本内容提取、词形还原等预处理方面作了技术改进；之后，以Indri检索系统为研究平台，利用结构化查询语言和web搜索引擎进行查询扩展和结构化查询主题构造，并采用基于文档标题字段的域查询，有效的提高了基本的ad-hoc主题检索的性能；在Weblog观点检索方面，使用基于最大熵的主观性内容判别模型，并提出了分类器自学习的策略，实现不同数据集之间的知识传递，在Weblog数据集上有效建模；同时，分别构建了句子级别和文档级别的最大熵模型，并将两者组合构成层叠式的Weblog文档观点性内容判别模型。在Blog Track数据集上的评测指标表明，本文构造的。Weblog观点检索系统达到了较高的性能水平。

2.学位论文 何慧 WEB文本挖掘中关键问题的研究 2009

随着互联网和通讯网的迅猛发展，网络文本成为信息的主要载体及人们生活中不可或缺的主要信息来源，文本挖掘技术的研究意义和实用价值越来越突出。另一方面，随着Web2.0时代的到来，出现了越来越多的由用户创作的网络数字内容。用户数字内容的大量产生和传播使得短文本计算、Web文本信息抽取、文本情感分析等逐渐成为Web文本挖掘研究的热点问题。针对这些问题，本文进行了以下研究：

(1)基于统计语言模型的短文本计算。针对短文本包含字符少、文本语言不规范、文本数量巨大的特点，本文提出了一种基于N-gram的特征提取和RPCL(Rival Penalized Competitive Learning)的短文本聚类算法。首先进行基于字符级的N-gram特征提取，即从未分词的语料中抽取中文块。中文块可以是一个汉字、一个词或者字符串，这样，中文块不但可以表达短文本的语义信息，而且能够保留语序结构和字符之间的依赖。然后通过统计子串约减和互信息过滤得到候选中文块集合。最后，使用一种神经网络聚类算法RPCL对短文本进行聚类。实验结果表明，这种基于N-gram的特征提取和RPCL的短文本聚类算法能够有效的对短文本聚类，并能有效的降低特征的维度。

(2)面向广告推荐和情感分析的Web文本信息抽取。针对广告推荐中的复合词抽取问题，本文提出了基于隐马尔科夫模型的半监督中文复合词抽取算法。从少量种子复合词出发，通过设定一个BEMI(Begin，End，Middle，Independent)模板，使用隐马尔科夫模型识别与种子复合词具有相同或相似信息的复合词。算法采用Bootstrapping的学习方法，通过自学习不断增大复合词列表的规模。实验结果表明，本算法可以满足广告系统关键词推荐的信息抽取需求，并具有较高的准确率和可以接受的召回率。

针对文本分析问题中情感词抽取的问题，本文提出了基于最大熵和LMR(Left，Middle，Right)模板的中文情感词抽取算法。通过对文本设定一个滑动窗口，使用LMR模板标记词的位置信息，使用词、词的先后位置信息、词性信息作为特征，对情感词进行识别和抽取。实验结果表明，本算法具有较高的召回率和准确率，同时在某些特征组合的情况下，情感词抽取具有良好的鲁棒性。

(3)基于监督和半监督的文本情感分类。针对网络上大量流行音乐、网友原创、改编的音乐，本文提出了一种对音乐歌词的情感分类方法。首先，通过对歌词语料库的词进行统计发现其分布基本符合齐夫定律，但与中文分类通用语料库(863计划文本分类测试数据)中词语分布略有差异。由于对歌词表现的情感进行的分类不同于按照主题对普通文本的分类任务，所以需要抽取更多表现情感色彩的特征。本文在N元模型的框架下采取了三种不同的预处理方法(不同N-gram模板、消去停用词、按词性过滤)抽取更多的歌词情感语义特征，并提出了带有高斯先验和指数先验的最大熵模型的分类算法对歌词的情感特征进行建模。实验结果表明，具有高斯先验和指数先验的最大熵模型非常适合用于歌词情感分析问题。

针对实际的情感分类中标注数据不足的情况，本文提出了一种基于半监督学习的文本情感分类算法。假设空间中存在一个情感流形结构，将待分类文本看作是这个情感流形上抽样的点。首先，利用这些点的邻域信息进行构图，每个点与它近邻的边的权重使用它的近邻线性加权表示；然后，将该图看作是一个概率转移矩阵，各类别的标签在此矩阵上扩散完成情感分类过程。在电影评论和中文歌词语料集上的实验结果表明，该算法在文本情感分类上具有良好的性能。

(4)文本观点检索。以本文作者2008年参加的COAE2008中的面向主题的中文文本观点检索任务为主线，介绍了本文参评系统PRIS-SAS。本系统采用两阶段处理方式，在经过编码转换、分词等预处理后，PRIS-SAS首先使用Indri检索系统对语料集建立索引，使用任务中的主题词进行ad-hoc检索，然后使用本文中文本情感分类算法建立倾向性模型和极性模型，对检索得到的相关文本进行文本倾向性判断，并对检索结果重新排序。在COAE2008数据集上的评测指标表明，本文设计的文本观点检索系统达到了较高的性能水平。

3.学位论文 熊德兰 中文网页褒贬倾向性分类研究 2006

文本自动分类是一种有效的信息处理方法，广泛应用于信息检索、信息过滤、信息管理、数据组织等领域。随着计算机和网络通信技术的发展，Internet迅速成为海量的、动态的全球信息服务中心，如何在浩若烟海而又纷繁芜杂的Web文档中掌握最有效的信息成为信息处理技术遇到的新的挑战。Web文本自动分类技术是目前Web数据挖掘的研究热点之一，它能够有效地组织和管理Web资源，提高信息检索的效率。网页自动分类技术与主题搜索、个性化信息检索、信息过滤、信息主动推送服务等技术相结合，可以有效地提高了信息服务的质量。

传统的Web文本分类是根据网页所涉及的主题来进行分类，如将网页分为政治类、军事类、经济类等等，而根据网页中作者对所描述内容的看法、观点等主观感情色彩进行分类的研究较少，我们称后者为情感分类。网页内容的褒贬性就是明显反映作者观点、态度的感情色彩之一，网页褒贬倾向性分类是未来多角度、立体性、个性化文本分类的研究内容之一。

本文探讨了网页褒贬色彩的客观性和褒贬倾向性分类的可行性，提出了名人网页褒贬感情色彩的综合评价方法。作者通过构建褒贬义词典和褒贬评价模板，提取出网页文本中具有情感取向的褒义词、贬义词及语法结构等褒贬特征，结合情感计算和层次分析法的相关理论，建立褒贬评价模型，实现对名人网页褒贬感情色彩的综合度量。同时，针对褒贬倾向的局限性，文中还提出了一种领域褒贬词典的构建方法，并探讨了使用模板自动更新褒贬词典的可行性方案。

在上述研究的基础上，结合自动分类技术，本文进一步探讨了名人网页褒贬倾向性分类的工作原理和实现方法，提出了LSI和KNN相结合的褒贬分类模型。根据网页的褒贬评价结果，提出了一种新的文本相似度计算方法，并给出了有关特征提取和分类过程的具体算法。最后，在名人网页数据集上，对上述理论进行了实验验证，取得了较好的成效。

本文链接：http://d.g.wanfangdata.com.cn/Periodical_qbxb201005023.aspx
授权使用：武汉理工大学(whlgdx)，授权号：adad6cf2-6175-44bc-abdf-9e6400fac413

下载时间：2011年1月7日