面向微博的社会情绪词典构建及情绪分析方法研究 *

蒋盛益[1]，黄卫坚[1]，蔡茂丽[1]，王连喜[2]

（1.广东外语外贸大学 信息学院，广东 广州 510006；

2. 广东外语外贸大学 图书馆，广东 广州 510420 ）

摘要：本文旨在探索一种面向微博的社会情绪词典构建方法，并将其应用于社会公共事件的情绪分析中。首先通过手工方法建立小规模的基准情绪词典，然后利用深度学习工具 word2vec 对社会热点事件的微博语料通过增量式学习方法来扩展基准词典，并结合 Hownet 词典匹配和人工筛选生成最终的情绪词典。接下来，分别利用基于情绪词典和基于 SVM 的情绪方法对实验标注语料进行情绪分析，结果对比分析表明基于词典的情绪分析方法优于基于 SVM 的情绪分析方法，前者的平均准确率和召回率比后者分别高 13.9%和 1.5%。最后运用所构建的情绪词典对热点公共事件进行情绪分析，实验结果表明本文方法是有效的。

关键词：微博；社会情绪；词典；情绪分析

中图分类号：TP391　　**文献标识码：**A

Building Social Emotional Lexicons for Emotional Analysis on Microblogging

Jiang Shengyi[1], Huang Weijian[1], Cai Maoli[1], Wang Lianxi[2]

(1.School of Informatics, Guangdong University of Foreign Studies, Guangzhou, 510006;

2. Library, Guangdong University of Foreign Studies, Guangzhou, 510420)

Abstract: This paper aims to explore a kind of method to build social emotional lexicons for Microblogging and apply it to analyze social emotions in social public events. First, the small-scale standard emotional lexicons are combined to build the basic emotional lexicon by hand. Then, word2vec, a tool based on deep learning, is used to conduct incremental learning method on the corpus from social events on microblogging increasingly to extend the basic emotional lexicon; and the ultimate emotional lexicon is generated via the filtering of HowNet and manual labor. Further, the paper compares the results of emotional analysis based on the generated emotional lexicon with that based on SVM classification, indicating that the former is superior in both average precision and recall rate, with 13.9% and 1.5% higher than the latter respectively. Finally,the proposed methods are verified according to emotional analysis on different social events with the generated emotional lexicon.

Key words: Microblogging; Social emotions; lexicon; Emotional Analysis

1 引言

互联网成为当下中国社会非理性情绪的集散地，是社会情绪分析的重要数据来源。作为新型的网络交流平台，微博不仅成为人们表达情绪的重要载体，更是民众讨论社会热点事件的重要场所，汇集了大众对社会话题的情绪表达。研究面向微博的社会情绪分析方法具有重要的现实意义，一方面有利于政府或相关部门进行舆情监控和传播引导；另一方面对社会事件的情绪分析有助于危机公关处理、名人形象维护等。

文本情绪分析的本质是对有情绪倾向的主观文本进行分析和处理的过程。现有文本情绪

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目;广东省科技计划项目(2014A040401083)；教育部人文社会科学研究青年项目(14YJC870021)；广东省哲学社会科学" 十二五" 规划项目 (GD14YXW02)

作者简介： 蒋盛益(1963—)，男，博士，教授，主要研究方向为数据挖掘与自然语言处理；黄卫坚(1993—)，男，本科生，主要研究方向为 Web 数据挖掘；蔡茂丽(1992—)，女，本科生，主要研究方向为 Web 数据挖掘；王连喜(1985—)，男，硕士，馆员，主要研究方向为自然语言处理与数据挖掘。

分析方法，以基于词典的匹配方法和基于机器学习的分类方法为主[1]。基于机器学习的分类方法主要有朴素贝叶斯、支持向量机和最大信息熵等。由于基于机器学习的分类方法易受训练语料影响，且部分算法涉及了复杂的参数设置，所以不便于建模。目前最常见的方法是构建高质量的情绪词典，并将其应用于文本情绪识别。由此可见，构建合理、覆盖范围宽泛的情绪词典是基于词典匹配方法的关键。

情绪词典是文本情绪分析的重要辅助资源。情绪词典的构建往往需要结合人工标注、语义词典扩展或基于语料库抽取标注等方法。英语中最重要的情绪词典资源是 WordNet-Affect，该词典通过选择和标注代表情绪概念的 WordNet 中的同义词集而获得的 Ekman 六种基本情绪相关的词语，然后利用 WordNet 中定义的关系、情绪标签和领域标签进行扩展，找到情绪同义词所在的 Synset 扩展得到情绪词典[2]。Zaher Salah 等人通过两种方法创建领域情绪词典：1）从已标注的语料中计算词汇极性及程度，生成词典；2）从已有词典出发，融合领域语料中的词汇语义信息、上下文信息、关系信息，学习一个基于分类器的扩展领域词典[3]。Suke 等人[4]认为具有相近或相同情绪的观点词有更高的概率同时出现，采用协同训练框架进行半监督的情绪分类训练以扩展情绪词。

在中文情绪词典资源方面，HowNet 是国内较为全面的知识库。借助于 HowNet，不少学者尝试构建特定领域的情绪词典。柳位平等在 HowNet 情绪词集合的基础上，利用 HowNet 的义原计算词与词间的相似度，并根据词和正向、负向种子词的平均相似度的差来判定词的情绪倾向性，从而得到特定的情绪词典[5]。常晓龙等将词语间的语素关系融入到图模型中、并结合词语同义关系，提出一种构建词典的半监督学习方法，形成了融合语素特征的中文褒贬词典[6]。徐琳宏、林鸿飞等构建了中文情绪词汇本体库，将情绪分为七个基本大类和二十一个小类，并利用相关情绪词典和语义知识库获得候选情绪词，再人工对部分种子词语的情绪类别和强度进行标注[7]。

本文的目的在于构建一个规模大，覆盖范围广的社会情绪词典。首先，根据现有的社会情绪相关文献和分析目标确定社会情绪类别，并整合已有情绪词典，补充典型的微博情绪词，建立规模较小的基准情绪词典。然后采用深度学习工具 Word2vec 对微博平台上的社会热点事件微博及评论等语料进行分析，以增量式的方式扩展基准词典；接下来，再辅以 HowNet 词典和人工筛选，生成最终的情绪词典。最后，利用所构建的社会情绪词典分析微博文本标注语料的情绪倾向，并对比基于情绪词典和基于 SVM 分类的情绪分析结果以验证所构建的词典的有效性；与此同时，利用所构建的情绪词典分析微博平台的社会热点事件呈现的社会情绪倾向，从侧面验证本文构建的情绪词典的有效性。

2 面向微博的社会情绪词典构建

情绪词典的构建流程如下图 1 所示：

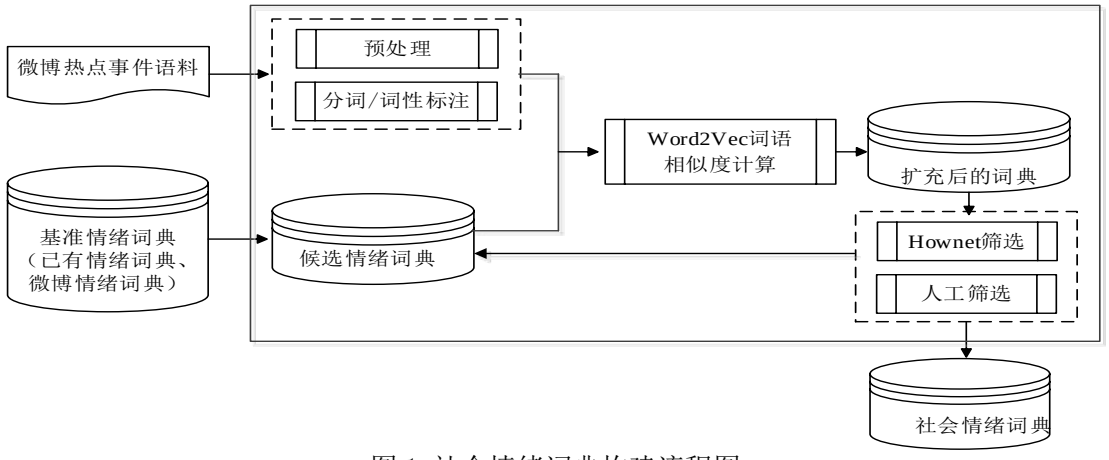

图 1 社会情绪词典构建流程图

2.1 基准情绪词典的建立及微博语料获取

社会情绪不同于个人情绪，个人情绪是指个体对一系列主观认知经验的通称，是多种感觉、思想和行为综合产生的心理和生理状态，其最基本的情绪有喜、愤、哀、惧、恐、爱等。当然也存在不同的看法，美国著名的心理学家伊扎德提出了 12 种基本情绪：兴趣、惊奇、痛苦、厌恶、愉快、愤怒、恐惧、悲伤、害羞、轻蔑和自罪感；社会情绪则侧重于群体成员情绪之间的相互作用和影响，是建立在个体对社会的人和事的认知、心理变化基础上的社会反映。目前，国内关于社会情绪的研究起步较晚，社会情绪分类体系方面的研究相对缺乏。考虑到本文分析的目标是微博平台上社会大众对特定新闻事件的情绪倾向，因此本文重点关注包括喜欢、愉快、愤怒、悲伤等社会大众普遍存在并且广泛关注的情绪类别。结合目前社会情绪的研究成果，同时对已有文献[8][9]和相关情绪词典进行梳理和归纳，最终将社会情绪类别确定为 8 类，分别是：愉快、喜欢、同情、悲伤、焦虑、厌恶、愤怒、怨恨，并以此作为词典扩展以及情绪分析的依据。确定情绪词典的类别之后，依次建立每个类别下所对应的基本情绪词，并以此构建基准情绪词典。基本情绪词是通过心理学、语言学和社会学专家从大连理工大学林鸿飞教授等建立的情绪词典以及微博语料中为每个类别挑选约 40 个词语构成。

本文实验的微博语料来自新浪微博平台上社会热点事件的微博文本及其转发和评论的文本。通过模拟登陆的方式，对新浪微博平台上的特定事件进行爬取。首先获取新浪微博媒体中粉丝数大于 100,000 的权威媒体的微博账号，然后针对某一特定社会公共事件，爬取这些微博账号发表过的带有该公共事件关键字的微博及其所有转发与评论文本，最终由这些微博以及其转发评论文本共同构成该公共事件的语料集合。目前共获取了 200 多个社会公共事件的微博语料，包括厦门纵火案、昆明火车站暴恐案、上海外滩踩踏事故等备受关注的事件。在获取微博语料后，需要对语料进行预处理，包括去除重复微博文本、去除停用词和分词等，为后续的分析提供有效的语料。

2.2 基于 word2vec 的词语相似度计算

word2vec 是谷歌一款基于 Deep Learning 的开源学习工具，它通过有效的连续词袋模型和 skip-gram 语言模型实现了词语的向量化，最大化地利用了词的上下文信息以丰富词语的语义信息，以文本向量空间的相似度来表示文本语义相似度。本文利用 word2vec 在大规模语料中计算词语相似度，并将其应用到情绪词典的构建过程中，然后通过迭代实现增量式的词典扩展。

word2vec 提供了 distance 函数，用以发现所查询词语的相似词语列表。根据初始构建的基准情绪词典，依次输入基准词典各个类别下的词语，利用 word2vec 得到的词向量进行相似度计算，分析得出最为相关的词语。实验设置如下：

1)迭代的次数为 8，即将上一次迭代的输出词语作为下一次迭代的输入词语，从而使输入词语的规模更大；

2)考虑到前后迭代词语的重要性不同，给迭代前和迭代后的词语赋予不同的权重。第一轮迭代后扩展出来的词语按照与情绪类别相似度从高到低排列，取相似度最大的前 10 个词语作为候选词扩充到词典，原基准词语与新扩充的候选词作为下一轮的基准词语;往后的迭代则取相似度最大的前 2 个词语扩充到词典。

3)迭代过程中，首先要去除停用词，但是由于停用词表具有一定的局限性，所以另外制定一些规则，过滤无意义的词语,包括：

a)纯数字的词语；

b)非表情符的纯标点符号组成的字符串；

c)在前面迭代过程中已经筛选掉的词语；

d)根据词性去除一部分词语，保留名词、动词、形容词、副词等。

由于微博语料是动态获取和扩展的，因此构建词典时采用增量式的扩充方式。具体的做

法是将上一轮语料扩展输出的词典用作下一轮扩展的候选情绪词典，同时加入新的微博语料以扩展语料规模，进一步有效地扩大词典规模。

2.3 情绪词的筛选

在语料规模不大的情况下，通过 word2vec 扩展得到的词语可能存在着准确度不高的问题，因此我们对扩展后的词典进行基于 HowNet 词典的自动筛选和独立的人工筛选。

借助 HowNet 词典计算扩展出来的词语与基准词语的相似度，通过排序方式筛选相似度高的词语。因为 HowNet 里面的词语更新具有一定的滞后性，HowNet 中并不一定包括扩充出来的词语，因此不能通过 HowNet 来计算该词语与种子词语的相似度来筛选候选词，所以我们采用的方法是：如果 HowNet 词典没有包含某词语，则默认其为新词保留；如果 HowNet 词典中包含该词但相似度小于指定阈值，则剔除该词。经过筛选后，进一步通过人工判断其类别。经过外部词典辅助筛选和人工筛选，最终得到包含 6,887 个词语的基于微博语料的社会情绪词典。本词典的各个情绪类别的情绪词数量分布如表 1 所示。

表 1 扩展后的词典

类别	个数	类别	个数
喜欢	396	焦虑	448
愉快	1019	厌恶	2795
同情	132	愤怒	687
悲伤	1309	怨恨	101

3 基于情绪词典的微博社会情绪分析

3.1 单条微博的情绪分析

微博文本体现出来的情绪倾向可认为是微博用户对于某一社会事件发表的主观看法，主要由两个方面来体现：情绪类别及其强弱程度。情绪类别即所构建的情绪词典定义的 8 个类别中的一个或多个；情绪类别的程度由情绪词的权值来体现。为验证本文所构建社会情绪词典的有效性，利用本文词典对微博平台上的社会热点事件进行情绪分析。由于文本的情绪强度更取决于句法结构、语境等整体因素，为了减少单个词语对整个文本情绪强度的影响，本文对情绪词典的每个词语赋予 1 的权值，如果出现多个同类别的词语，则将对应的向量维度值进行叠加或加权计算。考虑到情绪词可能被特殊词语（否定词和程度副词）修饰而改变情绪倾向，因此本文对这些特殊词语做进一步处理：一方面，被否定词修饰的情绪词通常会改变情绪倾向，所以考虑搜索并判断情绪词前后 3 个词内是不是含有否定词。如果是，则将该情绪词的权值乘以-1。另一方面，程度副词使情绪倾向在强弱程度上发生变化，类似于否定词的处理，搜索并判断情绪词前面 1 个词是不是程度副词，将程度副词的强度分为 5 个等级并赋予相应的权值。

单条微博情绪分析方法具体描述如下：

1)文本预处理。首先过滤噪声文本，如广告、重复的文本等；然后使用中科院分词系统导入本文构建的情绪词典，对微博文本进行分词，去掉停用词。由于 word2vec 是根据词共现的原理计算两个词语之间的相似度，而预处理的过程中去掉的停用词大部分为没有实际意义的介词，代词等，所以去掉停用词不会对实验结果产生太大的影响；

2)情绪特征词提取。通过导入情绪词典对评论文本分词后，选取出当前情绪词典里面的词作为该条评论的情绪特征词，利用情绪特征词构建文本情绪特征向量。

3)如果情绪特征词前有程度词，则情绪特征词的权重应该为程度词与特征词的权重之积(情绪特征词的权重设为 1)；

4)如果情绪特征词前有否定词，则统计否定词的个数 N，每个否定词的权重设为-1，最终情绪特征词的权重应该是 N 个-1 与特征词权重之积；

5)通过计算该条评论文本属于每一个情绪类别的对应情绪特征词的权值之和，选取权值最大的那个情绪作为该条评论的最终情绪类别。

按照以上处理步骤得出每条微博文本的特征向量后，选取出权值最大的特征项作为该微博文本的情绪倾向，并与事先人工标注的进行对比评价，分别计算出准确率和召回率。

本文从厦门纵火案和呼格吉勒图冤案两个热点社会公共事件的微博语料中随机抽取了10,000条微博文本进行不同情绪类别的人工标注，最后确认了7,629条有效微博文本作为实验的数据。实验结果如表2所示：

表 2　基于情绪词典的情绪分析结果

类别	准确率	召回率	F 值
喜欢	63.4%	78.4%	0.820
愉快	53.2%	73.3%	0.702
同情	57.8%	46.0%	0.523
悲伤	56.9%	26.9%	0.626
焦虑	67.8%	64.2%	0.663
厌恶	76.2%	52.2%	0.626
愤怒	93.9%	64.9%	0.772
怨恨	56.9%	26.9%	0.374
平均值	76.9%	61.5%	0.694

为了对比基于情绪词典和基于SVM分类的情绪分析，利用开源工具liblinear对微博语料进行情绪分类：

1)对于标注了的7,629条微博文本，按2:1的比例将数据集划分为训练集和测试集。

2)对文本数据进行文本预处理、特征表示和选择，实现文本向量化。

a) 去除重复文本和无意义的符号，进行中文分词。

b) 根据Chi公式计算词语的特征权重。

首先，计算每个词t与类别c之间的相关程度（假设t和c之间符合具有一阶自由度的Chi分布）。词语t对于类别c的Chi值由公式(1)计算：

$$X_2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1)$$

其中，N表示语料库中微博文本的总数目；A表示包含词语t且属于类别c类的微博数目；B为包含词语t但是不属于类别c的微博数目；C表示属于类别c但不包含词语t的微博数目；D表示既不属于类别c也不包含词语t的微博数目。

接着，根据公式(2)计算特征词语t对于整个语料的Chi值

$$X_{\max 2}(t)=\max_{1<=i<m}\{X_2(t,c_i)\} \qquad (2)$$

其中，m为类别数，该式子表示选取特征与各类别的Chi值中的最大值。

然后，对每个词语t的Chis值进行排序，选取前K个词语作为特征项。

最后，通过公式(3)的词语权重计算公式得到每个特征项的权值，将文本向量化。具体地，针对语料统计每个特征项在该文本数据中的个数m，记为词频TF；统计每个特征项在不同文档中出现的次数，记为文档频率DF，其逆文档频率log（DF）记为IDF。根据公式(3)计算每个t的权重值。

$$TF*IDF=m*\lg\frac{N}{\sum_{0<i<m}\text{词t在类别i中出现次数}} \qquad (3)$$

3)利用开源项目 liblinear 对训练数据进行建模，建立分类器，参数设置为默认值。

4)构建分类器后，在测试集上预测分类，并计算准确率和召回率以评价分类的结果(如下表 3 所示)。

表 3　基于 SVM 分类的分析结果

类别	准确率	召回率	F 值
喜欢	72%	75%	0.62
愉快	85%	70%	0.77
同情	54%	58%	0.56
悲伤	74%	42%	0.54
焦虑	79%	70%	0.74
厌恶	50%	51%	0.50
愤怒	58%	39%	0.47
怨恨	60%	67%	0.63
平均值	63%	60%	0.61

由表 2 所示结果可以看出，基于本文所构建的情绪词典的分类器的平均准确率为 76.9%，平均召回率为 61.4%；F 值是 0.694，而 SVM 分类器的平均准确率为 63%，召回率为 60%，F 值是 0.61。这初步验证了本文所构建的面向微博的社会情绪词典的性能，说明该词典能够比较准确且高效地反映微博文本中不同的情绪倾向。

3.2 基于情绪词典的微博热点事件情绪分析

进一步，将本文所构建的情绪词典应用到微博平台上的社会公共事件的情绪分析，通过典型的社会事件例子从侧面反映情绪词典的有效性。对特定社会公共事件的语料进行整体的情绪分析，以判断该事件反映出来的公众社会情绪倾向。本文选取受到广泛关注的“厦门纵火案”事件和“呼格吉勒图冤案”事件。把同一个事件的微博语料当成整体，通过分词、情绪词典匹配和特征权重计算，得到微博语料对应的不同情绪倾向的比重。分析结果如图 2 和图 3 所示：

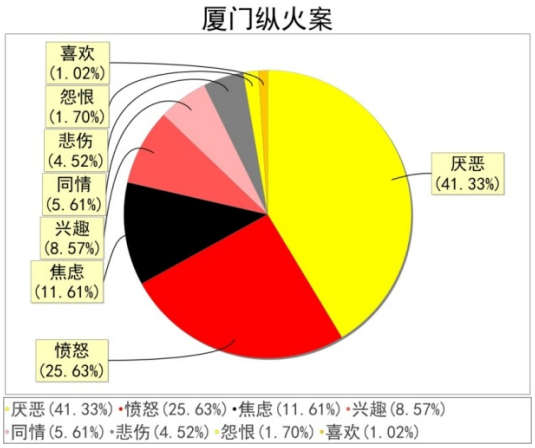

图 2 “厦门纵火案”事件分析结果

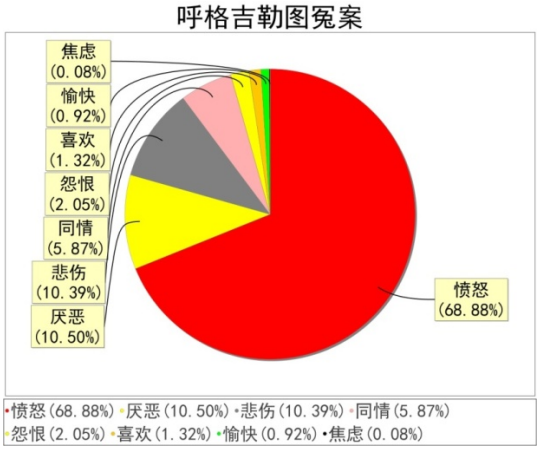

图 3 “呼格吉勒图冤案”事件情绪分析结果

如图 2 和 3 所示，不同的情绪倾向按照权值占比从高到低排序，可以直观地观察到占比排在前 3（总和超过 78%）的情绪倾向。对于“厦门纵火案”事件：厌恶，愤怒，焦虑三者的权值占比高，表明社会情绪偏向于厌恶、愤怒和焦虑；而对于“呼格吉勒图冤案”事件：愤怒、厌恶和悲伤权值占比高，表明社会情绪偏向于愤怒、厌恶和悲伤。针对这两个事件，从社会民众普遍的心理感知角度来看，符合上述提到的几种情绪倾向，这在一定程度上也说明本文构建的情绪词典和分析方法的有效性。

4 总结与展望

本文基于微博平台上社会热点事件的微博语料，建立了面向微博的社会情绪词典，该社会情绪词典包括 8 个类别共 6,887 个词条。应用该词典对公共事件进行社会情绪分析，并通过基于情绪词典的微博情绪分析和基于 SVM 的情绪分析结果的对比，验证了本文构建的情绪词典及情绪分析方法的有效性。最后，通过对微博平台上的社会热点事件的整体情绪分析，从另一个侧面表明所构建的情绪词典的有效性。

本文的研究工作还存在一些不足，后续将从以下几个方面进行深入研究：

1)邀请更多心理学、语言学等领域专家等对词典进行校验，提高词典的质量。

2)获取更多公共事件，扩大语料库规模，同时借助维基百科等外部数据源，融合多种词语相似度计算方法，进一步扩大词典规模。

3)在情绪词典的应用方面，增加微博评价对象识别，以更准确反映公共事件的社会情绪。

致谢：在本文的研究过程中，郑漫丽、陈丽云、陈东沂等同学作了大量探索性实验，丘心颖、谢柏林、李霞等老师给出了一些建设性的建议。

参考文献：

[1] Zhang Jianfeng, Xia Yunqing, Yao Jianmin. A review towards microtext processing[J]. Journal of Chinese Information Processing, 2012, 26(4):21-27.

[2] Carlo Strapparava, Alessandro Valitutti. WordNet-Affect: an Affective Extension of WordNet [J]. ITC-irst,Istituto per la Ricerca Scientifica e Tecnologica I-38050 Povo Trento Italy:1083-1086.

[3] Salah Z, Coenen F, Grossi D. Generating domain-specific sentiment lexicons for opinion mining[M]//Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2013: 13-24.

[4] Suke Li, Jinmei Hao, Yanbing Jiang, Qi Jing. Exploiting Co-occurrence Opinion Words for Semi-supervised Sentiment Classification. ADMA2013_book1[C]. 36-47.

[5] 柳位平,朱艳辉,栗春亮等.中文基准情感词词典构建方法研究[J].计算机应用, 2009.10(29): 2875 – 2877.

[6] 常晓龙,张晖.融合语素特征的中文褒贬词典构建[J]. 计算机应用, 2012, 32(7): 2033 – 2037.

[7] 徐琳宏,林鸿飞,潘宇,等.情感词汇本体的构造[J].情报学报,2008,27(2): 180-185.

[8] 桂守才.基础心理学[M].北京:人民教育出版社, 2007.

[9] 林传鼎.社会主义心理学中的情绪问题[J].社会心理学科,2006,21(83):37-62.

作者联系方式：蒋盛益，广东外语外贸大学信息学院，510420

电话：15915869428　邮箱：jiangshengyi@163.com