

基于分类器融合的中文微博情感倾向性研究

刘志广[†] 董喜双 关毅[†]

哈尔滨工业大学, 哈尔滨, 150001

[†] 通讯作者, E-mail: guanyi@hit.edu.cn

摘要 本文提出一种融合最大熵模型和支持向量机两者预测结果用于识别主观句和褒贬极性分类问题的方法。构建了一个包含基础情感词和网络流行语的情感词典,使用 k-means 和无向图对训练和测试语料进行了优化。针对微博文本的特点提取了微博文本特有的特征向量。在对句子分类之前,将观点句切分成短句并基于规则提取特征,应用最大熵模型和支持向量机共同预测短句的极性,最后根据短句的极性预测长句的极性。在 2012 年 CCF 中文微博情感分析评测的情感倾向性识别任务中微平均的 F 值为 0.741,在 34 个参赛队伍中名列第四位。结果表明分类器融合对褒贬极性判别有效且可以适应文本中噪声。

关键词 情感分析; 观点句识别; 最大熵; 支持向量机

中图分类号 TP391.1

Sentiment Analysis of Chinese Micro Blog Based on Multiple Classifiers Fusion

LIU Zhiguang, DONG Xishuang, GUAN Yi[†],

Harbin Institute of Technology, Harbin, China, 150001

[†]Corresponding Author, E-mail: guanyi@hit.edu.cn

Abstract A Subjectivity extraction and polarity classification method is presented based on the fusion of maximum entropy classifier and support vector machine classifier in this paper. A sentiment lexicon containing basic sentiment words and network catchwords is constructed, and k-means and undirected graph are employed to optimize the training and testing set. Feature vector are extracted considering the specification of micro blog text. Opinion sentences are split up into short sentences in which sentiment features were extracted by rules, then the polarity of short sentences was predicted by maximum entropy model in combination with support vector machine classifier, finally polarities of opinion sentences are predicted according to polarities of short sentences. The micro F-score of this method is 0.741, ranking at the 4th among 34 participants in the polarity detection task of sentiment analysis evaluation of Chinese micro blog, 2012 CCF, which shows that classifier fusion is effective for polarity classification and adapt to text noise.

key words sentiment analysis; subjectivity extraction; maximum entropy; support vector machine

微博情感倾向性分析研究在监控公众舆论对某个人、产品和事件的观点动向等方面获得了广泛的应用。根据维基百科关于新浪微博的最新统计数据显示,新浪微博的注册用户已经超过了 3 亿,用户每日的发帖量超过一亿条。由于微博的开放和即时传播特性使得信息可以在极短时间内大范围的传播,然而作为一种新兴媒体,急需一种情感倾向性监控系统使得微博能够处于舆论安全体系的正确引导和监督,否则很可能会对社会造成负面影响。另一方面,面对微博用户发布的海量信息,企业可以不断的挖掘这些数据迅速理清用户对某些产品的反馈,寻找新的商机、开发新的产品。

可以使用有监督、无监督和半监督三种方法从事句级情感分析研究。Davidiv^[1]等人借鉴 K-最邻近法(k-Nearest Neighbors, kNN)的思想设计了一种有监督分类器,利用 hashtag 和表情符号将 tweet 划分为多种情感类型。Hassan^[2]等人使用监督型马尔科夫模型,利用词性信息和依存关系来确定发布在 Usenet 上的消息极性。Barbosa and Feng^[3]从三个不同的 Twitter 情感分析网站上获取训练数据,然后使用这些数据训练标准的 SVM 分类器,实验结果表明 SVM 分类器的精度可以达到 81.3%。Turney^[4]提出一种无监督的方法用于产品和电影评论的积极和消极分类。A. Meena 等^[5]重点考虑了连词对句子情感极性分析的影响,结合短语和连词分析句子情感极性。但系统依赖人工构建情感词典,并且需要人工构建连词规则,不具有领域

适应能力。王根和赵军^[6]提出情感句分级模型，将句子的主客观分类、褒贬分类以及褒贬强度分类统一处理。上一级的情感分析结果作为下一级情感分析的输入，逐层细化情感分析，但会出现错误从高层向底层传递现象。Socher 等^[7]提出基于递归自动编码的半监督情感句分析模型。将句子中词递归的两两合并构建短语树，提取短语节点特征预测句子情感极性。Tan 等^[8]认为社会关系信息可以提高情感分析精度。假设存在关系的用户可能有相似的观点，提出基于半监督学习和用户关系信息的情感分析模型。Li^[9]等针对语料不平衡问题，提出通用的半监督情感分类模型

目前，中文微博情感分析面临的困难主要有以下几点：（1）缺乏中文微博语料。虽然国外对英文微博的研究比较深入，但语料一般都局限于 Twitter 等英文数据；（2）中英文微博信息量不同。以 Twitter 为例，他只允许用户输入 140 个英文字符，这通常是一句话的长度，而中文微博却可以输入 140 个中文和符号。由于中英文语言信息量的不同，使得中文微博可以表达更加丰富的信息，从而导致情感倾向性的识别更加复杂；（3）微博文本中包含大量俚语和网络用语。与通用语料的情感倾向性分析相比，中文微博的文本内容更加简洁，语言结构不是十分严谨（比如省略句子的主谓宾等），其中充斥着大量的口语和网络流行语。因此，微博情感倾向判别的精度低，单纯依靠情感词典利用模式匹配的方式不能适应微博文本的多样性。

本文提出一种融合最大熵模型和支持向量机两者预测结果用于识别主观句和褒贬极性分类问题的方法：首先构建了一个高质量的情感词典，其中包含大量的基础情感词和网络流行词以及各个情感词所对应的情感极性。预测模型主要采用分类器融合的方法预测句子的主客观和褒贬倾向性，在情感极性的识别中获得了较高的准确率。对微博句子的情感倾向性研究借鉴了 Su^[10]等人的思想，采用最大熵模型和 SVM 相结合的方式预测句子的极性，将句子拆分为短句并使用一些规则提取特征，然后利用模型预测短句极性，最后用短句极性预测长句极性。

本文第一部分描述了如何构建基础情感词典和网络流行语情感词典；第二部分采用分类的思想完成句子极性的预测；第三部分以中文微博情感分析评测的数据为基础分析了本文所使用的方法和实验结果；最后分析了本文所使用方法的不足和今后的研究方向。

1 情感词典的构建和语料的优化

1.1 情感词典的构建

本文使用以下资源构建初始情感词典：（1）由哈尔滨工业大学语言技术研究中心网络智能研究室 WI 提供的中文倾向词典和 WI 情感词典，其中包含细粒度标注的情感词 1428 个；（2）由知网提供的 HowNet 评价词典和 HowNet 情感词典，其中含有正面情感词 836 个，负面情感词 1254 个，正面评价词 3730 个，负面评价词 3116 个；（3）由台湾大学提供的台湾大学情感词典，其中褒义词 2,812 个，贬义词 8,276 个；（4）由清华大学共享的中文褒贬义词典，其中包含 5567 个褒义词，4468 个贬义词^[11]；

为适应互联网用语的多样性，我们又构建了一个常用的网络流行语情感词典。本文从互联网中整理了大量的网络用语，使用从腾讯微博开放平台获取的 40 万条微博语料统计这些网络流行词的词频，选择前 15% 的词纳入网络情感词典。实验表明，抽取观点词总数量的 15% 作为种子词时，情感倾向性判别的准确性最高^[12]。

在完成初始情感词典的构建后，人工过滤掉其中的非情感词并纠正一些极性标注错误的情感词。本文使用了部分情感词构建了一个训练词集，利用 HowNet 扩展同义词和反义词并获取情感词的特征，然后使用这些特征训练最大熵模型^[13]。对于从微博语料中挖掘出的候选情感词，人工标注出其情感倾向性作为种子情感词，同样使用 HowNet 扩展种子情感词生成情感词特征，最后利用最大熵模型预测扩充的情感词情感极性并人工筛选出那些极性不稳定的情感词。情感词典的构建流程如图 1 所示。

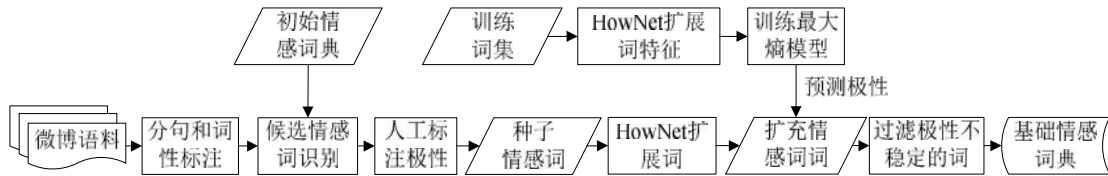


图 1 基础情感词典构建流程图

Fig.1 The flow chart of building sentiment lexicon

将基础情感词典和网络流行词典合并为一个文本文件，作为用户自定义词库加入到 ICTCLAS^[14]中。本文给每个情感词都添加一个形容词或副词的标签，这样重新对训练语料和测试语料进行分词，可以大大提高分词的精准度。另外，由于微博数据中包含很多近期新出现的名词，使得通用分词器不能很好地识别出这些词。为此，本文单独整理了一些近年来新出现的专有名词，比如艺人名字、电视节目名和公司名字等。同样地，给这些专有名词加上名词标签便于分词器更好的识别这些词。

1.2 训练语料的优化

为了使本文所设计的分类器能够适应微博数据，本文从腾讯微博开放平台获取 15 个话题：包含经济、教育和政治等领域的近 5000 条微博数据。然后采用人工标注的形式对每条微博数据的情感倾向进行标注。

另外一个语料来源是 Ren_CECps 中文情感语料库，其数据来源主要为互联网各大中文博客中的文章^[15]。该语料共包含了 1,487 篇中文博客文章，共计 11,255 个段落，35,096 个句子，878,164 个词语。

由于课题组自己标注的语料是按照话题类选取的，而微博中很多话题都可能会出现观点一边倒的情况（即对于某个话题，可能绝大多数的人都倾向于褒义或者贬义），造成了训练语料的不平衡。为了弥补语料不均衡这种现象，本文采用 K-means 聚类的方式将训练语料划分为不同的组，从而得到一组平衡语料。另外，由于 Ren_CECps 中文情感语料库中有很多句子是利用段落的整体情感极性与段落中句子情感极性相同的假设获取到的，因此会存在很多噪声并影响分类器性能。为了尽可能的去除这下噪声，本文使用一种无向图，每个句子都用一个顶点来表示，点与点之间的边表示句子间的关系。当两个句子间存在一个相同的特征时就表示它们之间存在一条边。本文选取的一些特征包括：句子中情感词的数量、特殊句式如反问句和疑问句、微博表情等。使用这种方法可以去除孤立噪声数据、提升训练语料的质量。

1.3 训练和测试语料的预处理

对于从腾讯微博开放平台获取的训练语料，在将他们转换为供分类器使用的提取特征向量之前，需要进行一系列的预处理。表 1 结合一条具体的微博消息，列举了本文所使用的主要预处理步骤。


#法语学习#@阿广： 法语太难学了，学法语的人你伤不起啊，有木有!!! 此处省略三千字
<http://url.cn/9CBmeb>

表 1 微博文本预处理主要步骤

Table 1 The main process of preprocessing step for micro blog text

序号	预处理	举例说明
1	表情符号过滤	#法语学习#@阿广: <ANGRY>法语太难学了，学法语的人你伤不起啊，有木有!!! 此处省略三千字 http://url.cn/9CBmeb
2	URL 过滤	#法语学习#@阿广: <ANGRY>法语太难学了，学法语的人你伤不起啊，有木有!!! 此处省略三千字<URL>
3	标点符号过滤	#法语学习#@阿广: <ANGRY>法语太难学了，学法语的人你伤不起啊，有木有<EXL>此处省略三千字<URL>
4	“#话题#”的过滤	<topic>@阿广: <ANGRY>法语太难学了，学法语的人你伤不起啊，有木有<EXL>此处省略三千字<URL>
5	“@昵称”的过滤	<topic><@><ANGRY>法语太难学了，学法语的人你伤不起啊，有木有<EXL>此处省略三千字<URL>
6	停用词过滤	<topic><@><ANGRY>法语太难学，学法语 伤不起，有木有<EXL>此处省略 字<URL>

2 微博文本倾向性识别

2.1 特征选择

在使用分类器对文本分类前，需要考虑如何表示微博文本信息。本文在特征提取方面，不但抽取了传统信息获取方法中普通文本的特征，还针对微博自身的特点提取了相应的特征。由于微博是一种非常特殊

的媒体，微博中的每句话通常都比较短、相对于新闻中的文本它的结构不是很严谨、甚至有很多省略的现象出现，因此非常有必要寻找出其他的文本特征来辅助微博消息的极性判别。如表 2 所示，本文主要使用词和词序列，情感词强度累加值，句子间的相互关系，句型特征，微博表情符号，口语和网络流行词等特征构成识别观点句的特征向量。表 3 在这些特征的基础上又添加了连词、相邻句子的主客观和情感极性构成了观点句情感倾向性判别的特征集。

表 2 观点句识别特征集

Table 2 Feature set for opinion sentences



特征编号	特征描述	例子
1	情感词和词序列	否定词，如“不”
2	情感词强度累加值	喜，怒，哀，惧的强度值
4	特殊句式	疑问句，反问句
5	微博表情符号	 
6	含有名实体或代词	它，他，她
7	口语和网络流行词	给力，伤不起，
8	重复出现的标点符号	!!!，???

表 3 观点句情感倾向性判别增补特征集

Table 3 Additional feature set for polarity classification of opinion sentences

特征编号	特征描述	例子
1	相邻句子的相互关系	前后句子的主客观性和情感极性
2	情感词附近的连词和转折词	和，但是，却

2.2 褒贬极性分类

本文使用最大熵模型（Maximum Entropy, ME）和支持向量机（Support Vector Machine, SVM）作为基本的褒贬极性分类器。

熵在信息论中是指信息的不确定性，如果将最大熵应用于文本分类中，实际上就是模型将会保留所有的已知条件，最大限度的将文本预测的风险降低。本文使用 SCGIS 算法训练最大熵模型，在文本褒贬极性的识别中取得了很好的效果。

SVM 采用直线模型将数据分类，并且只有处于分界线附近的支撑向量才会对最后的分类结果有贡献，因此特别适合于处理高维向量分类问题，比如说文本分类。通常来说，一个样本点距离分离超平面的远近可以表示分类的置信度。为了防止方差大的随机变量主导分类过程，将每个元素都被缩放到[-1, 1]内。SVM 的输出结果 d_i 是表示与分离超平面之间距离的有向距离向量（正向表示主观句，负方向表示客观），当样本点被分离超平面正确分类时 d_i 就是该样本点到超平面的距离，反之则为负值。本文使用公式(1)将其转换为非负值，其中 $P_{subj}(s)$ 表示句子 s 属于主观句类别的概率，则 $P_{obj}(s)=1-P_{subj}(s)$ 表示句子 s 属于客观句的概率。

$$P_{subj}(s) = \begin{cases} 1 & d_i > 1 \\ (1 + d_i) / 2 & -1 \leq d_i \leq 1 \\ 0 & d_i < -1 \end{cases} \quad (1)$$

通过下一节 SVM 和 ME 的对比实验可以看出，相对于 SVM 来讲，ME 分类器召回率较高，分类精度表现一般；而 SVM 则具有较高的分类精度，可是召回率相对较低。为此，本文提出一种融合这两种分类器的思想，通过计算两种分类器对同一个句子 s 的情感倾向概率和来预测句子真正的极性。当两个分类器对句子 s 的褒义倾向概率和大于贬义倾向概率和时，表示句子 s 为褒义，否则为贬义。联合使用 SVM 和 ME 识别句子情感倾向性的主要流程如图 2 所示。

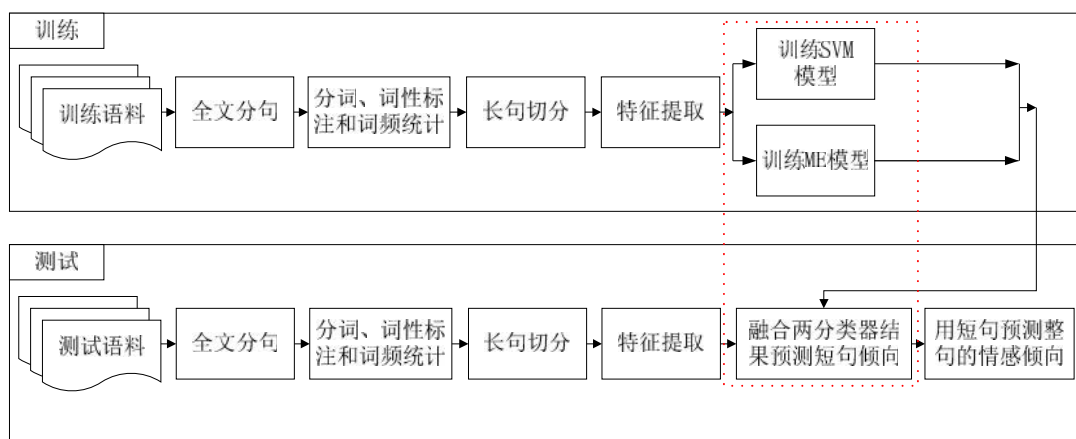


图2 句子情感倾向性识别流程图

Fig.2 The flow chart of polarity classification for opinion sentences

2.3 主客观句分类

虽然可以使用基本的褒贬极性分类器直接对测试文本的褒贬情感倾向分类，但是如果预先将客观句过滤掉之后再行褒贬分类则可以大大提高分类的准确率。本文重新使用主客观句训练语料训练褒贬极性分类器，从而使褒贬分类器也可以作为基本的主客观分类器。图3显示了识别主观句的具体流程。通过主客观分类器的处理后可以得到一组新的数据集，该集合中只含有主观句。接下来使用上一节设计的褒贬倾向性分类器，以主观句集合为输入数据，识别这些句子的情感倾向性。

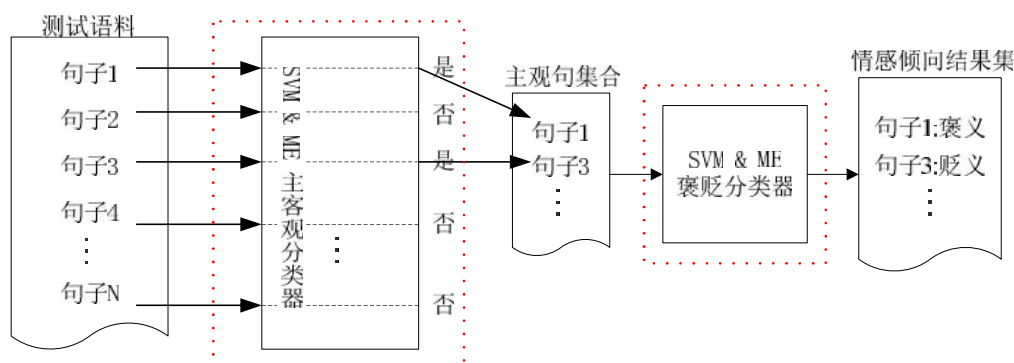


图3 使用主客观分类器辅助句子褒贬倾向性的识别

Fig.3 Subjectivity classifier assists with polarity detection

2.4 自我训练的学习方法

自我训练算法是一种半监督学习方法，它利用模型对未标注数据的极性进行预测，将“最确定”（置信度较高）的分类结果加入训练集重新训练模型，从而可以从未标注数据中获取训练语料、扩展语料规模。本文在主观句识别和褒贬句倾向性识别中都使用了自我训练的思想，在模型的迭代过程中不断的将置信度较高的分类结果添加到训练集中，这样分类模型的精度会有所提高。

3 实验结果与分析

训练语料 为了训练主客观分类器，本文从腾讯微博开放平台抓取 30,000 条微博信息，选择文本内容的长度大于 10 的句子并人工标注了主观句 2500 句、客观句 1000 句。为了平衡语料，采用 K-means 聚类的思想从 Ren_CECps 中文情感语料库中抽取客观句 1500 条，从而得到 5000 条主客观平衡语料。采用同样的方法获取 4000 条褒贬平衡语料。

测试数据集 测试数据采用 2012 年 CCF 自然语言处理与中文计算会议提供的中文微博情感分析评测数据集，其中共有 3416 个句子，主观句 2207 个，正面倾向的句子 407 个，负面倾向的句子 1766 个，混合倾向的句子 44 个。

观点句抽取和情感倾向性识别结果如表 4 和表 5 所示。

表 4 观点句识别结果

Table 4 Results of subjectivity extraction

结果编号	微平均			宏平均		
	正确率	召回率	F 值	正确率	召回率	F 值
SVM&ME	0.738	0.726	0.732	0.743	0.717	0.726
平均值	0.727	0.615	0.647	0.727	0.607	0.634
最高值	0.835	0.959	0.784	0.836	0.96	0.783

表 5 观点句情感倾向性识别结果

Table 5 Results of polarity classification for subjectivity

结果编号	微平均			宏平均		
	正确率	召回率	F 值	正确率	召回率	F 值
SVM&ME	0.881	0.640	0.741	0.878	0.632	0.733
ME	0.863	0.626	0.726	0.860	0.619	0.718
平均值	0.704	0.460	0.546	0.702	0.454	0.536
最高值	0.902	0.850	0.850	0.899	0.854	0.854

从表 5 的数据可以看出，虽然在识别主观句时的准确率只有 73.8%，低于最高值 83.5% 近 10 个百分点，但是在观点句的倾向性判别中准确率达到 88.1%。可见本文所使用的分类器融合方法非常适用于观点句情感极性判别，如果能引入其它方法将观点句识别的精度和召回率提高，一定会提高情感倾向性判别的总体性能。另外，从表 4 和表 5 中 F 值的微小变化可以分析出本文所采用的方法鲁棒性好，对文本中的噪声有较好的适应性。

4 结论与展望

本文融合了最大熵模型（ME）和支持向量机（SVM）两者预测结果用于处理句子主客观和褒贬极性分类问题。使用这种方法在任务 1 和任务 2 中取得了较好的预测效果，尤其是任务 2 在主观句识别率较低的情况下仍然获得了 88.1% 的正确率。在微博情感倾向性分析中，将观点句切分成短句并提取短句的特征，然后联合使用 ME 和 SVM 来预测短句的褒贬极性，用短句情感极性预测长句的情感极性。通过对比主客观分类器识别出的主观句和褒贬极性分类的结果，发现使用主客观分类器可以提高褒贬分类的整体性能。下一步工作主要从以下几方面入手：（1）继续丰富情感词典和提高微博标注语料的规模；（2）重新筛选一些更加适合描述微博文本的句子特征；（3）采用适当的方法优化分类器的预测结果。

参考文献

- [1] Dmitry Davidiv, Oren Tsur and Ari Rappoport. Enhanced Sentiment Learning Using Twitter Hash-tags and Smileys. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. 2010: 241-249.
- [2] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. What's with the attitude? identifying sentences with attitude in online discussions. In Proceedings of the 2010 Conference, 2010: 1245-1255.
- [3] Barbosa, L., and Feng, J. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of COLING. 2010: 36-44.
- [4] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In proceedings of ACL. 2002: 417-424.
- [5] A. Meena, T. Prabhakar, G. Amati, et al. Sentence Level Sentiment Analysis in the Presence of Conjunctions Using Linguistic Analysis. Advances in Information Retrieval, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 573-580.
- [6] 王根, 赵军. 基于多重标记CRF句子情感分析的研究. 全国第九届计算机语言学学术会议. 2007.
- [7] R. Socher, J. Pennington, E. H. Huang, et al, Semi-supervised Recursive Auto-encoders for Predicting Sentiment Distributions.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011: 151-161.

- [8] C. Tan, L. Lee, J. Tang, et al. User-level Sentiment Analysis Incorporating Social Networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 1397-1405.
- [9] S. Li, Z. Wang, G. Zhou, et al. Semi-Supervised Learning for Imbalanced Sentiment Classification, in Proc. IJCAI, 2011: 1826-1831.
- [10] Fangzhong Su; Katja Markert. Subjectivity Recognition on Word Senses via Semi-supervised Mincuts. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, Boulder, Colorado, 2009: 1-9.
- [11] 李军. 中文评论的褒贬义分类实验研究[D]. 北京: 清华大学, 2008.
- [12] 柳位平, 朱艳辉, 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.
- [13] 董喜双, 关毅, 李本阳, 等. 基于最大熵模型的中文词与句情感分析研究. 第二届中文情感倾向性分析会议, 2009: 1-8.
- [14] ICTCLAS. ICTCLAS汉语分词系统[EB/OL]. [2012-01-30] . <http://ictclas.org/>.
- [15] Changqin Quan, F. R.. A blog emotion corpus for emotional expression analysis in Chinese.. Computer Speech and Language. 2010, 24(4): 726-749.