第 40 卷 第 4 期 计 算 机 工 程 2014 年 4 月
Vol.40 No.4 Computer Engineering April 2014

·人工智能及识别技术· 文章编号：1000-3428(2014)04-0159-05 文献标识码：A 中图分类号：TP18

基于证据理论的多分类器中文微博观点句识别

郭云龙，潘玉斌，张泽宇，李 莉

(西南大学计算机与信息科学学院，重庆 400715)

摘 要：随着新技术及社会网络的发展与普及，微博用户数据量剧增，与此相关的研究引起了学术界和工业界的关注。针对中文微博语句特点，通过对比多种特征选取方法，提出一种新的特征统计方法。根据构建的词语字典与词性字典，分析支持向量机、朴素贝叶斯、K 最近邻等分类模型，并利用证据理论结合多分类器对中文微博观点句进行识别。采用中国计算机学会自然语言处理与中文计算会议(NLP&CC 2012)提供的数据，运用该方法得到的准确率、召回率和 *F* 值分别为 70.6%、89.2%、78.9%，而 NLP&CC 2012 公布的评测结果相应平均值分别为 72.7%、61.5%、64.7%，该方法在召回率和 *F* 值 2 个指标上超过其平均值，而 *F* 值比 NLP&CC 2012 评测结果的最好值高出 0.5%。

关键词：微博；观点句；支持向量机；朴素贝叶斯；K 近邻；证据理论

Multiple-classifiers Opinion Sentence Recognition in Chinese Micro-blog Based on D-S Theory

GUO Yun-long, PAN Yu-bin, ZHANG Ze-yu, LI Li

(School of Computer and Information Science, Southwest University, Chongqing 400715, China)

【Abstract】With the development and popularity of the new technology and social network, the data volume of micro-blog users surge sharply. Related research causes increasing attention from both academia and industry. This paper proposes a new statistical method on feature extraction. Classification performances of different schemas such as Support Vector Machine(SVM), Naive Bayes and K-Nearest Neighbour(KNN) are analyzed carefully. It proposes a combined model based on D-S theory to take the advantages of different classifiers. A series of experiments based on the Chinese Micro-Blog data provided by CCF NLP&CC 2012 are conducted, and it gets the average estimate 72.7% in precision, 61.5% in recall and 64.7% in *F*-measure of NLP&CC 2012 as a baseline. Experimental results show that the method can achieve significant enhancement in both recall and *F*-measure with 70.6%, 89.2% and 78.9%, respectively, and *F*-measure is even 0.5% higher than the best result of NLP&CC 2012.

【Key words】micro-blog; opinion sentence; Support Vector Machine(SVM); Naive Bayes; K-Nearest Neighbour(KNN); D-S theory

DOI: 10.3969/j.issn.1000-3428.2014.04.031

1 概述

随着互联网的发展，尤其是 Web2.0 应用的普及，基于用户关系的信息分享、传播及获取平台——微博迅速兴起。微博具有以下特点[1]：(1)内容简短，长度限制为 140 个字符；(2)数据量大，数据的来源丰富，包罗万象；(3)传播速度快，微博用户可以任意转发，评论；(4)实时性，微博可以通过多种终端随时发布。用户可以频繁地使用微博对某产品及热点事件进行评论。产品的评价对于商家及买家都较有价值，而热点事件的评论对政府做出正确决策也至关重要，但巨大的信息量使得用户很难在短时间内准确获取网络群体的兴趣点[2]。

观点挖掘技术已成为国内外研究热点。近年来，ACL、SIGIR、KDD 等国际会议，都有相关议程探讨该领域的发展，NTCIR、COAE 等评测也涉及该研究热点。

中文微博观点句的抽取问题[3-4]，可理解为基于数据短文本的一种二分类的句子级文本分类技术。当前主要方法分为以下 2 类：

(1)基于词典的方法：一般利用预先构建的情感词典(可以人工标注或机器统计)，处理文本中出现的词语及其情感信息，结合制定的规则，进而判断其主客观性(即观点句或非观点句)。文献[5]以 HowNet 情感词语集为基准构建情感词典，计算情感词的极性，从而识别短文本主客观性。文献[6]考虑了连词对句子情感极性的影响，结合短语和连词

基金项目：国家自然科学基金资助项目(61170192)。

作者简介：郭云龙(1990－)，男，硕士研究生，主研方向：自然语言处理，语义网络；潘玉斌，本科生；张泽宇(通讯作者)，硕士研究生；李 莉，教授。

收稿日期：2013-05-20 **修回日期：**2013-07-12 **E-mail：**zqlong@swu.edu.cn

分析句子主客观性。文献[7]提出一种可以学习不同数据源，结合上下文自动构建情感词典的算法。此类方法过分依赖情感词典的构建工作，同时，规则的制定需要语言学背景，大大影响了该方法的推广使用。

(2)基于机器学习方法：利用训练集，采用特定的机器学习方法，对测试集进行有效的分类。常用的机器学习分类器有：朴素贝叶斯(Naive Bayes)，最大熵(Max Entropy)，支持向量机(Support Vector Machine, SVM)等。国内这方面的工作有文献[8-9]等。文献[10]提出将情感句分级，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。文献[11]利用相似性、朴素贝叶斯分类和多重朴素贝叶斯分类等统计方法进行观点句识别。文献[12]使用 K-最邻近法(K-Nearest Neighbors, KNN)设计有监督分类器，利用 hashtag 和表情符号将 tweet 划分为多种情感类型。然而，此类方法在训练集的标注、特征值选取及统计、分类器选择等问题上有很大难度，因而有一定的局限性。

本文结合上述 2 类方法的长处，首先对中文微博短文本进行预处理，构建词语词典和词性词典。然后对多种特征进行选取，并改进特征统计方法；从证据理论的角度出发，综合多分类器。

2 整体流程

2.1 观点句定义

NLP&CC2012 评测对观点句的要求为：(句子)只限于对特定事物或对象的评价，不包括对自我情感、意愿或心情的表达，例如“我感到很高兴。”这样的句子是明显的情感句，但不属于评测定义的观点句。而“我真心喜欢 iphone5 的屏幕效果。”，该句属于本评测定义的观点句。

2.2 微博的语言特点

微博的文本结构形式就决定了它的语言具有句子简短、负面倾向多、语句口语化程度强、表达情感强烈而理性评价淡化、评价对象在句中不直接出现、语言不够规范等特点，具体可参见文献[13]，该文的数据分析见表 1。

表 1 评论与微博关于平均句长的比较

文体形式	文本数	汉字数	句子数	平均句长/字
舆情评论文	400	356 511	9 366	38.06
话题型微博	20	68 726	3 416	20.11

2.3 评价标准

本文评价标准采用微平均准确率、召回率与 F 值对测试数据进行评估。

$$Precision=\frac{\#system\_correct(opinion=Y)}{\#system\_proposed(opinion=Y)}$$

$$Recall=\frac{\#system\_correct(opinion=Y)}{\#gold(opinion=Y)}$$

$$F-measure=\frac{2\times Precision\times Recall}{Precision+Recall}$$

其中，*#system_correct* 为模型正确的识别句子数；*#system_proposed* 为模型识别的句子数；*#gold* 为测试集观点句数目。

2.4 整体框架

中文微博观点句识别框架如图 1 所示。

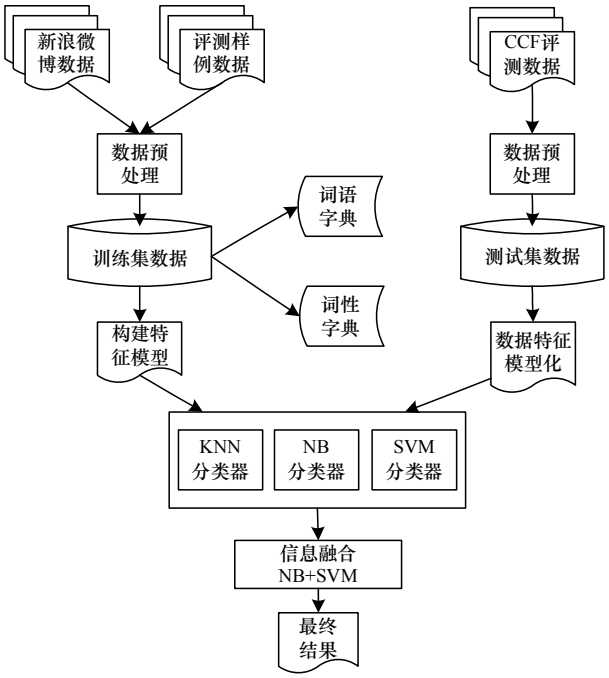

图 1 中文微博观点句识别框架

本文主要工作流程如下：

(1)数据预处理：本文涉及的数据来自新浪微博、CCF 评测样例数据及评测公布的数据；

(2)构建词语字典和词性字典：使用部分数据，形成训练集，并构建相应的词典；

(3)特征提取：根据特征模型进行特征抽取，实现数据的向量化；

(4)在分析单个分类模型的基础上，融合分类效果好的模型，从而实现对中文微博观点句的识别任务。

3 算法设计

本节详细介绍流程涉及的各模块，包括：数据预处理，构建词典，特征模型的选取，特征统计方法的改进，以及根据数据融合思想实现的多分类器。

3.1 文本预处理与词典构建

本文采用中科院汉语词法分析系统(ICTCLAS)以及斯坦福句法工具(Stanford Parser)对文本数据进行分词及词性标注，由于 ICTCLAS 对繁体字支持较弱，先对待评测数据进行简繁体转化。并对文本做如下预处理：

(1)将数据中 Hashtag 之间的话题变为 topic；

(2)将包含网址 url 的部分去掉；

(3)将数据中具体表情都变为 Expression；

(4)去掉数据中的停用词。

根据 2.2 节所述微博的语言特点，有必要针对于微博语言构造词典。一方面，传统的情感词典中收录的词语过于正式，微博中甚少出现，另一方面，目前并没有中文微博网络词语的词典，所以，在实际中根据训练集中出现的所有词语以及词性，构建词语字典及二连词性字典(POS-2)。在本文中，词语字典共收录训练集出现的 12 315 个词语，词性字典共收录 9 120 维两连词性。

3.2　特征模型选取

本文根据文献[14-15]，对中文微博文本特征进行选取。具体特征如表 2 所示。

表 2　SVM 所选特征值

序号	类型	特征内容	描述
F1	情感词	情感词典词语个数	整理 HOWNET 情感词典 8 223 个情感词
F2	指示性动词表	指示性动词表中动词个数	所构建的 23 个指示性动词表中的动词
F3	词性	单一词性	中科院分词系统共 96 种词性
F4	双词性	2 个连续词性组合	中科院分词系统 96 种词性，9 216 种双词性组合，取 CHI 值前 100 的组合
F5	词语	单个词语	中科院分词系统分词后，统计训练集中 CHI 值前 200 的词语

但在实验中发现，表 2 的许多特征并不能够作为中文微博的分类特征。例如，情感词的个数，首先情感词就较为正式，微博文本中很少出现；再者，微博文本长短不一，个数特征难以衡量。在实验中抽取了多种特征，并实验了特征的各种组合，进一步根据文献[16]提出了一种根据连续双词词类组合词语(2-POS)自动判断句子主观性程度的方法，最终选取了单一词语与二连词性作为分类特征，结果如图 2 所示，多特征(参赛)是暑期参加评测时所用的特征，多特征(改进)是在撰写工作报告时的改进工作，削减了词语个数特征，Word–1 以单一词语为特征，Word–1+Pos–2 是以单一词语和二连词性为特征。结果说明，使用单一词语及二连词性作为特征时效果最好。

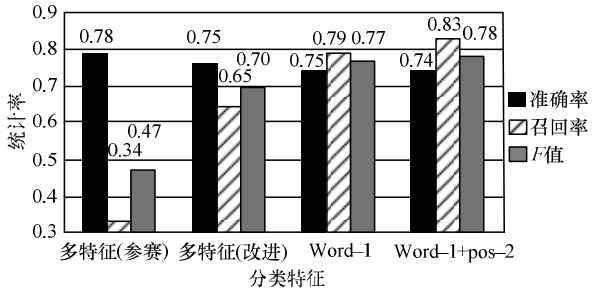

图 2　多种特征分类结果的比较

3.3　特征值统计方法

在对特征值的统计方法中，常用的方法有信息增益方法(IG)、卡方分布(CHI)值统计、文档频率(DF)、词频反文档频率(TF-IDF)等。

下面介绍 3 种常用的统计方法，具体如下：

(1)卡方统计 CHI 值

$$CHI(p,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

其中，p 表示词语；c 代表文本类，在本文中，c 即为观点或非观点；N 表示训练集中的句子总数；A 表示 c 类中词语 p 出现的次数；B 表示非 c 类中词语 p 出现的次数；C 表示 c 类中没有出现词语 p 的句子数；D 表示非 c 类中没有出现词语 p 的句子数。该方法只考虑了词语的宏观情况，并没有顾虑到词语在每一句中的信息，如词频，因而有一定的不足。

(2)词频反文档频率(TF-IDF)

对某给定的文件，词频(Term Frequency, TF)指的是某一个给定的词语在该文件中出现的次数。

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

上式中 $n_{i,j}$ 是该词在文件 d_j 中的出现次数，而分母则是该词在文件中 d_j 所有字词的出现次数之和。

逆向文件频率(Inverse Document Frequency, IDF)是一个词语重要性的度量。某一特定词语的 IDF，可以由总文件数除以包含该词语之文件的数目，再将得到的商取对数得到：

$$idf_i=\log\frac{|D|}{\left|\left\{j:t_i\in d_j\right\}\right|}$$

其中，$|D|$ 为语料库中的文件总数；$\left|\left\{j:t_i\in d_j\right\}\right|$ 为包含词语 t_i 的文件数目(即 $\left\{n_{i,j}\neq 0\right\}$ 的文件数目)，如果该词语不在语料库中，就会导致被除数为 0，因此，一般情况下使用 $1+\left|\left\{j:t_i\in d_j\right\}\right|$。

$$tfidf_{i,j}=tf_{i,j}\times idf_i$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。但此种方法没有考虑到不同的类别信息。

(3)词频分类文档频率(TF-Classify)

逆向文档频率倾向于保留出现次数少的词语，这并不符合微博用于情况。本文改进的统计算法为：

$$Classify_i=\log\left(\frac{|Y-Sentence|}{j:t_i\in Y-Sentence}\right)-\log\left(\frac{|N-Sentence|}{j:t_i\in N-Sentence}\right)$$

其中，$|Y-Sentence|$ 是观点句的个数；$\{j:t_i\in Y-Sentence\}$ 是该词语出现在观点句中的次数；$|N-Sentence|$ 是非观点句的个数；$\{j:t_i\in N-Sentence\}$ 是该词出现在非观点句

中的次数。

$$tfClassify_{i,j} = tf_{i,j} \times Classify_i$$

利用上述 3 种统计方法的实验结果如图 3 所示，相比之下，TF-Classify 方法在 2 个指标上有优势，因此本文最终选取 TF-Classify 统计方法计算实验数据的特征值。

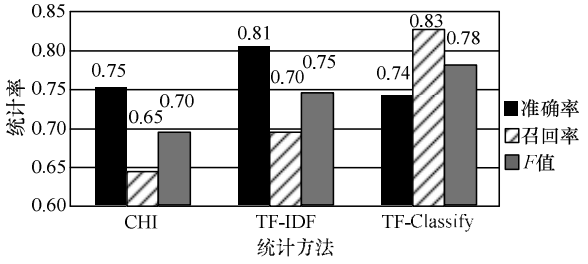

图 3 多种统计方式结果比较

3.4 多模型分类结果比较

常用的机器学习分类器算法包括：朴素贝叶斯(NB)，支持向量机(SVM)，K-最邻近(KNN)等。根据文献[17]，本文选取了上述 3 种分类器进行比较。

(1)K-最邻近法(KNN)

假设存在某样本，KNN 的目的就是从训练样本空间中找出 K 个与其最相近的样本，然后观察这 K 个样本中哪个类别的样本多，则待判定的值(即抽样)就属于这个类别，计算相似度的公式如下：

$$Sim(d_i,d_j)=\frac{\sum_{k=1}^{M} W_{ik}\times W_{jk}}{\sqrt{\left(\sum_{k=1}^{M} W_{ik}^{\ 2}\right)\left(\sum_{k=1}^{M} W_{jk}^{\ 2}\right)}}$$

在 K 个邻居中，依次计算每类的权重：

$$p(\bar{x},C_j)=\sum_{\bar{d}_i\in KNN} Sim(\bar{x},\bar{d}_i)y(\bar{d}_i,C_j)$$

其中，$\bar{x}$ 为测试数据的特征向量；$Sim(\bar{x},\bar{d}_i)$ 为相似度计算公式，与式(10)类似；$y(\bar{d}_i,C_j)$ 为类别属性函数，也就是说，如果 $\bar{d}_i$ 属于类 C_j，那么函数值为 1，否则为 0。

(2)朴素贝叶斯(NB)

假设 S 为句子，C 表示类别，NB 算法如下：

$$P(c|s)=\frac{P(s|c)P(c)}{P(s)}$$

通过假设在给定类别的条件下，句子中的每个词 x_i 相互条件独立，$P(s|c)$ 分解为：

$$P(s|c)=\prod_{x_i\in s} P(x_i|c)$$

其中，$P(x_i|c)$ 为每个词在某一类中出现的频率。

$$\min\frac{1}{2}\|\boldsymbol{w}\|^2 \text{ s.t.} y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)\geqslant 1,\ i=1,2,\cdots,n$$

其中，$y_i\in\{-1,1\}$ 为数据点 x_i 的类别；$\|\boldsymbol{w}\|$的意思是 $\boldsymbol{w}$ 的二范数，最小化这个式子即找到间隔最大平面。s.t 是加入的限制条件，可以看做是一个凸多面体求最优解。上式可转化为：

$$\boldsymbol{w}=\sum_{i=1}^{m} a_i y_i x_i$$

其中，a_i 是拉格朗日算子。求解这个式子的过程需要拉格朗日对偶性的相关知识。

(3)基于单个分类器的实验结果

实验结果如图 4~图 6 所示。测试数据由 NLP&CC 2012 主办方提供，对比实验选取“菲军舰撞击”等 7 个话题共 1 337 个句子进行测试。结果显示，KNN 准确率相比略高，但召回率与 F 值较低，NB 算法召回率比较高，SVM 结果相比不错。从图中不难看出，KNN 分类器的召回率与 F 值偏低，所以本文将不再考虑该分类方法。

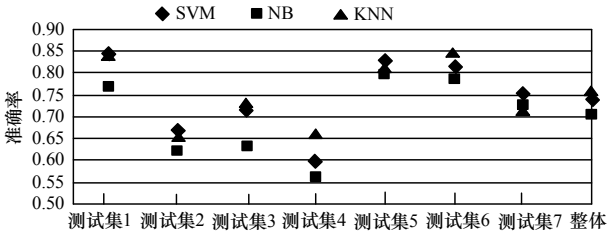

图 4 多分类器结果准确率比较

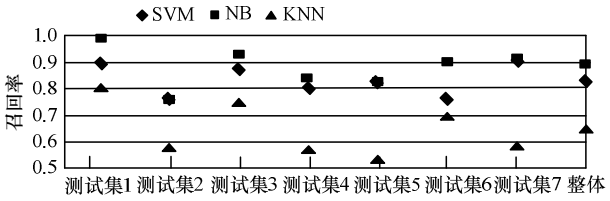

图 5 多分类器结果召回率比较

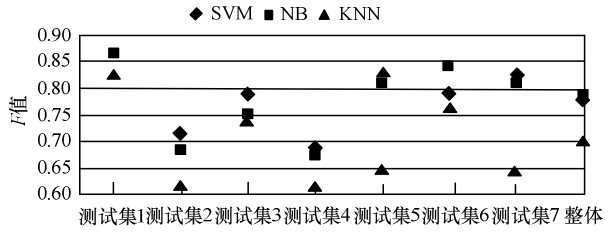

图 6 多分类器结果 F 值比较

(4)基于 D-S 理论的多模型融合

为结合多分类器的优势，本文引入证据理论，即根据不同分类器给出的概率结果进行融合。根上面的实验，选用 NB 和 SVM 进行融合。根据证据理论[18-19]，有：

$$m(A)=\frac{1}{1-K}\sum_{B\cap C=A} m_1(B)m_2(C)$$

$$K=\sum_{B\cap C\neq\varnothing} m_1(B)m_2(C)$$

其中，m_1 和 m_2 可以理解为利用 SVM 以及 NB 的分类结果，A、B、C 为基本事件，本文分类中只有观点句与非观点句 2 个基本事件，因此 $m(A)$ 可以理解为融合后事件 A 的最终概率。结合 D-S 理论，重新进行了实验，具体内容将在下一节叙述。

4 实验结果

测试数据由 NLP&CC 2012 主办方提供。整个数据集

包括 20 个话题，每个话题包含大约 1 000 条微博，共约 20 000 条微博。数据采用 XML 格式，句子已预先切分好，共 31 675 句。

利用本文的方法，对数据集进行观点句提取。本文通过测试样例数据以及网上相关微博数据标注，训练集共 8 050 句，其中观点句及非观点句各 4 025 句，并随机选取 7 个话题共 1 337 句，以评测方平均水平为基线，与评测方已知结果进行比对，如表 3 所示。准确率、召回率以及 *F* 值均有提高，结果如图 7 所示，最终的融合结果准确率为 70.6%，召回率为 89.2%，*F* 值为 78.9%，比评测的平均结果有明显的提升，其中，*F* 值甚至超过了评测的最好结果。同时，信息融合技术的使用，召回率 *R* 和 *F* 值均较 SVM 和 NB 方法有所提高，融合后的准确率 *P* 比单独使用 NB 有一定提高，但不如 SVM 的准确率高。

表 3　评测方最终实验结果

评测值	准确率	召回率	*F* 值
评测平均值	0.727	0.615	0.647
评测最优值	0.835	0.449	0.584
评测最优值	0.645	0.959	0.772
评测最优值	0.671	0.944	0.784
本文结果	0.706	0.892	0.789

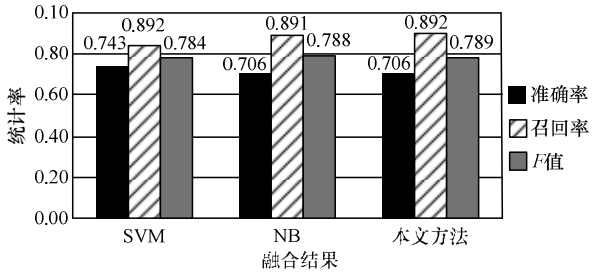

图 7　融合结果对比

本文的贡献如下：

(1)区别于传统的情感词典，构建出新的词语词典及词性词典。

(2)对比多种特征，选取最为有效的特征作为文本分类依据。

(3)对比 CHI 及 TF-IDF 等特征统计方法，设计并改进特征统计方法。

(4)对比支持向量机(SVM)、朴素贝叶斯(NB)、K 最邻近(KNN)等分类器，结合证据理论的知识，进一步提高分类的指标。

5　结束语

随着微博的崛起，有关中文微博的研究引起了各方的极大兴趣。本文以训练集数据构建词语字典和词性字典，实验对比多种特征，改进了现有的特征量化方法，并利用信息融合的思路对中文微博观点进行识别。基于新浪微博数据、CCF 评测样例数据及评测公布数据，本文的准确率、召回率、*F* 值分别到达了 70.6%、89.2%和 78.9%，比评测公布的平均值有明显提高。

目前对于中文微博的研究还处于探索阶段，下一阶段的工作主要包括：(1)关注微博特有的网络词汇和表情特征，充实特征集合；(2)以句子为单位进行分析过于单一，以单个微博为单位，结合句子的上下文关系，找出观点微博，具有更广泛的应用价值；(3)深入挖掘 D-S 技术在多模型融合中的应用。

参考文献

[1] Kwak H, Lee C, Park H, et al. What Is Twitter, a Social Network or a News Media?[C]//Proceedings of the 19th International Conference on World Wide Web. Seattle, USA: ACM Press, 2010: 597-600.

[2] Hu Minqing, Liu Bing. Mining and Summarizing Customer Reviews[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Stroudsburg, USA: ACM Press, 2004: 168-177.

[3] Wilson T, Wiebe J, Hoffmann P. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis[C]//Proceedings of Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACM Press, 2005: 347-354.

[4] Jiang Long, Yu Mo, Zhou Ming, et al. Target-dependent Twitter Sentiment Classification[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: [s. n.], 2011: 151-160.

[5] 何凤英. 基于语义理解的中文博文倾向性分析[J]. 计算机应用, 2011, 31(8): 2130-2133.

[6] Meena A, Prabhakar T V. Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis[C]// Proceedings of the 29th European Conference on IR Research. Berlin, Germany: Springer-Verlag, 2007: 573-580.

[7] Lu Yue, Castellanos M, Dayal U, et al. Automatic Construction of a Context-aware Sentiment Lexicon: an Optimization Approach[C]//Proceedings of the 20th International Conference on World Wide Web. [S. 1.]: ACM Press, 2011: 347-356.

[8] 姚天昉, 彭思崴. 汉语主客观文本分类方法的研究[C]//第三届全国信息检索与内容安全学术会议论文集. 苏州: [出版者不详], 2007.

[9] 刘志明, 刘　鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1): 1-4.

[10] 王　根, 赵　军. 基于多重标记 CRF 句子情感分析的研究[C]//全国第九届计算机语言学学术会议论文集. 大连: [出版者不详], 2007.

(下转第 169 页)