ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2013/07(11)-1033-07
doi: 10.3778/j.issn.1673-9418.1305008

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

中文文本情感词典构建方法*

阳爱民[1+],林江豪[2],周咏梅[1]
1. 广东外语外贸大学 思科信息学院,广州 510420
2. 广东外语外贸大学 国际工商管理学院,广州 510420

Method on Building Chinese Text Sentiment Lexicon*

YANG Aimin[1+], LIN Jianghao[2], ZHOU Yongmei[1]
1. Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510420, China
2. School of Management, Guangdong University of Foreign Studies, Guangzhou 510420, China
+ Corresponding author: E-mail: amyang18@163.com

YANG Aimin, LIN Jianghao, ZHOU Yongmei. Method on building Chinese text sentiment lexicon. Journal of Frontiers of Computer Science and Technology, 2013, 7(11): 1033-1039.

Abstract: Massive Internet text sentiment analysis is currently a hot research topic. This paper describes a method on Chinese text sentiment lexicon construction. This method improves the pointwise mutual information (PMI) algorithm for computing the weights of general sentiment lexicon, by selecting several sentiment seed words and drawing upon the total result numbers from search engine. In order to examine the validity of this method, this paper uses the established sentiment lexicon for text sentiment, and compares the classification effects of the method based on sentiment lexicon with those of naïve Bayesian classifier. The experimental results indicate that the high-quality sentiment lexi-

* The National Social Science Funding Project of China under Grant No. 12BYY045 (国家社会科学基金项目); the New Century Excellent Talents Foundation from Ministry of Education of China under Grant No. NCET-12-0939 (教育部新世纪优秀人才支持计划); the Humanities and Social Sciences Project of Ministry of Education of China under Grant Nos. 10YJCZH247, 09YJCZH019 (教育部人文社会科学研究青年基金项目); the Science and Technology Planning Project of Guangdong Province of China under Grant No. 2010B031000014 (广东省科技计划项目); the Social Sciences Planning Project of Guangdong Province of China under Grant No. 08GJ-08 (广东省社科规划项目); the School Project of Guangdong University of Foreign Studies under Grant No. 12Q22 (广东外语外贸大学校级项目); the Postgraduate Research & Innovation Project of Guangdong University of Foreign Studies under Grant No. 12GWCXXM-12 (广东外语外贸大学研究生科研创新项目).
Received 2013-03, Accepted 2013-05.
CNKI网络优先出版:2013-05-15, http://www.cnki.net/kcms/detail/11.5602.TP.20130515.1608.004.html.

con can effectively choose and classify the sentiment characteristics, and has a stable classification function.
Key words: sentiment lexicon; sentiment classification; pointwise mutual information (PMI); naïve Bayes

摘 要: 互联网海量文本的情感分析是当前的一个研究热点。介绍了一种中文文本情感词典构建方法,该方法选用若干个情感种子词,利用搜索引擎返回的共现数,通过改进的PMI(pointwise mutual information)算法计算情感词的情感权值。将构建的情感词典应用到文本情感分类实验中,在不同的语料环境下,对比基于情感词典和朴素贝叶斯分类器下的文本情感分类效果,实验结果表明,构建的情感词典,可有效用于情感特征选择和直接用于情感分类,并且分类性能稳定。
关键词: 情感词典;情感分类;PMI算法;朴素贝叶斯
文献标志码: A **中图分类号:** TP393

1 引言

互联网上的海量文本情感挖掘,有利于产品推荐、观点抽取和舆情监控等研究。现有的文本情感分析方法,以基于机器学习方法的分类方法为主[1],典型的有朴素贝叶斯(naïve Bayes,NB)[2-3]、支持向量机(support vector machine,SVM)[4-5]和最大信息熵(maximum entropy,ME)[6]等方法,但是分类器设计复杂,且受训练语料的影响比较大。然而在实际应用中若利用高质量的情感词典,则采用简单快速的方法就可以得到很好的效果[7]。由于网络不断更新,旧词新用导致语义的转变,都对情感分类造成直接的影响。因此,如何自动构建有效的情感词典引起了国内外学者的广泛关注。中文的HowNet[8]是国内较全面的知识库。在HowNet的基础上,构建特定情感词典的研究也有很多。如柳位平等人[9]在HowNet情感词语集的基础上,利用其提供的义原计算两个词的相似度,根据词与正向和负向种子词的平均相似度的差,来判定词的情感倾向得到一个情感词典。常晓龙等人[10]提出了一种将词语间语素关系融入到图模型中,并结合词语同义关系进行中文褒贬词典半监督构建的方法,构建了融合语素特征的中文褒贬词典。而朱艳辉等人[11]用基础情感词词典、连词词典及词语距离,提出了一种基于多重词典的中文文本情感特征抽取算法。英文的WordNet是经典的知识库,也是认同度较高的词典。Baccianella等人[12]基于WordNet构建了认可度最高的SentiWordNet。Hamouda等人[13]基于机器学习方法,构建了MLBL(machine learning based senti-word lexicon),取得了较SentiWordNet更高的微平均值。Maks和Vossen[14]基于词典模型进行情感分析和意见挖掘。

受语境迁移的影响,现有的情感词典直接应用到特定语料的情感分类中存在情感覆盖面不足,分类效果差的缺点。如名词“垃圾”,在生活中表达废弃物的时候,不具备情感;当用户在评价自己购买的产品时(“这款手机真垃圾,建议不要买”),具有负面的情感。因此,很难构建一部令人满意的情感词典。鉴于此,本文取HowNet、台湾大学的NTUSD(National Taiwan University Semantic Dictionary)和清华大学的褒贬义词典三者的并集作为基础情感词典。在互联网海量信息背景下,利用搜索引擎,获取基础情感词典与种子情感词集(共80个词,包括正向词汇40个,负向词汇40个)的共现次数,改进PMI(pointwise mutual information)算法,计算基础情感词的情感倾向权值。将构建的情感词典应用到电子产品评论和新浪微博两类语料及混合语料的情感分析中,验证所构建情感词典SOSL(sentiment orientation senti-lexicon)的通用性。同时,对比基于情感词典和机器学习方法下的文本情感分类效果,实验发现,随着测试语料数量的增加,基于情感词典的分类性能保持稳定,并优于机器学习方法,说明了本文微博情感词典构建方法的有效性。第2章将介绍情感词典的构建方法;第3章验证本文方法的有效性;第4章总结全文。

2 情感词典构建方法

2.1 情感词选择

直接选用NTUSD、HowNet以及褒贬义词典中情感词并集作为情感词集。考虑到受主观判断的影响，可能存在同一情感词在不同情感词典中的情感极性不同，对其情感极性采用如式(1)所示的投票规则来进行极性选择。

$$Polarity=\begin{cases}-1, \sum P(w)\leqslant -1\\ 0, \text{others}\\ +1, \sum P(w)\geqslant 1\end{cases} \tag{1}$$

其中，$Polarity$为投票结果；$P(w)$为情感词在源词典中的极性(+1表示正向，−1表示负向)。当情感词同时出现在三部源词典中，有两个词典极性相同，则取两个词典的极性；当情感词只出现在其中两部源词典中，可能出现结果为others，此时根据权威性，设定优先权HowNet>NTUSD>褒贬义词典来选择极性。保留原始极性的目的在于对比生成情感词典与源词典的情感词极性。获得情感词24 130个，负向情感词13 861个，正向情感词10 269个。

2.2 基于搜索引擎的情感倾向权值

Wang和Araki[15]改进了SO-PMI(semantic orientation from pointwise mutual information)算法，利用Google搜索引擎，构建日语情感词典。阳爱民等人[16]考虑用户行为，将用户行为因素融合到SO-PMI算法中，构建了中文宾馆评论情感词典，用于宾馆评论情感分类，获得了92.84%的微平均值。本文综合以上两种方法，选用了80个具有情感极性代表性的情感词作为种子词集，如表1所示，其中正向情感词40个，负向情感词40个。

利用百度搜索引擎，设定搜索条件为“$w+P\_set[i]$”或“$w+N\_set[i]$”，即只返回两个词共现的页面数。计算词w的SO值，如式(2)所示。

$$SO(w)=\text{lb}[A] \tag{2}$$

$$A=\frac{\sum\limits_{P\_set[i]\in P\_set} hits(w \text{ and } P\_set[i])}{\sum\limits_{N\_set[i]\in N\_set} hits(w \text{ and } N\_set[i])}\times \frac{\sum\limits_{N\_set[i]\in N\_set} hits(N\_set[i])}{\sum\limits_{P\_set[i]\in P\_set} hits(P\_set[i])}$$

Table 1 Basic sentiment word set

表1 基础情感词集

词集	极性	情感词
N_set	负向	不良，弊病，痴呆，莫名其妙，粗暴，功利主义，固步自封，愁，惨不忍睹，悲惨，诡异，狠毒，假冒，吹毛求疵，粗鄙，尖酸，浑浑噩噩，暴殄天物，自不量力，抱佛脚，薄情，保守，饱食终日，崩溃，畸形，画蛇添足，委屈，坏，变态，失败，呆板，游离，走火入魔，痛苦，毛病，煽情，嘘头，支离破碎，郁闷，扭曲
P_set	正向	好，光环，得体，震动人心，巅峰，福分，感激，魅力，别具匠心，大师，昌盛，出色，淳朴，甘之如饴，独到，绚丽，恰到好处，优秀，才华横溢，创造力，积极，如火如荼，著名，力透纸背，逼真，舒服，灿烂，纯真，飞扬，青春，美好，和谐，宽容，自由自在，欢愉，成熟，诚实，善良，和平，文明

其中，$hits(*)$为搜索引擎的返回页面数。考虑微博网络用语词典可能存在中性词的情况，对式(2)加入以下条件场，Ta、Tb和Tc为阈值。

$$\begin{cases}\dfrac{HP}{H}>Ta \text{ and } \dfrac{HN}{H}>Ta\\ HP<Tb \text{ and } HN<Tb\\ \left|\dfrac{HP}{H}-\dfrac{HN}{H}\right|<Tc\end{cases} \tag{3}$$

$$HP=\sum_{P\_set[i]\in P\_set} hits(w \text{ and } P\_set[i])$$

$$HN=\sum_{N\_set[i]\in N\_set} hits(w \text{ and } N\_set[i])$$

$$H=HP+HN$$

当满足式(3)中的一个条件，则词汇归入中性词集，即$SO(w)=0$。因此，基于搜索引擎的情感词典SL(sentiment lexicon)构建算法描述如下：输入情感词典SL，正向情感种子词集P_set和负向情感种子词集N_set；输出带情感权值的情感词典$SOSL$。$length(*)$为数组的长度，sum(*)为求和函数，abs(*)为求绝对值。

Input: SL & P_set & N_set

Output: Sentiment lexicon(SL) with SO weight

Step 1. For i=1 to $length(P\_set)$

$Hits\_P[i]=hits(P\_set[i])$;

$Hits\_N[i]=hits(N\_set[i])$;

//$length(P\_set)=length(N\_set)$

End For

Step 2. Calculate *A_basic*=sum(*Hits_P*)/sum(*Hits_N*);

Step 3. For *j*=1 to *length*(*SL*)

For *i*=1 to *length*(*P_set*)

hits_p[*i*]=*hits*("*P_set*[*i*]+*SL*[*j*]");

hits_n[*i*]=*hits*("*N_set*[*i*]+*SL*[*j*]");

//*length*(*P_set*)=*length*(*N_set*)

End For

HP=sum(*hits_p*);*HN*=sum(*hits_n*);*H*=*HP*+*HN*;

if ((*HP*/*H*>*Ta* and *HN*/*H*>*Ta*) or (*HN*<*Tb* and *HN*<*Tb*) or abs(*HP*/*H*−*HN*/*H*)<*Tc*) *SO*[*j*]=0;

Else

SO[*j*]=lb((sum(*hits_p*)/sum(*hits_n*))**A_basic*);

Save word and *SO* weight to *SOSL*;

End if

End For

Step 4. Output *SOSL*.

采用以上算法进行情感词典的构建，获得情感词24 127个，其中正向词14 922个，负向词9 165个。与情感词在源词典中的极性进行对比，情感词的识别率达到99.99%，正向情感词的微平均为82%，负向情感词的微平均为80%。观察情感倾向不一致的情感词发现，大部分属于弱强度情感词，如"不是味儿"、"酸溜溜"等词。而在基于词典的情感分类中，采用对情感权值进行加权求和作为分类标准，因此这些弱情感词对分类结果的影响不大，所构建情感词典可用于实际的文本情感分类。

3 实验结果及分析

3.1 语料采集

采集太平洋网的电子产品（手机、相机、笔记本）评论和新浪微博构建语料库。语料由3位标注人员进行人工标注，为保证标注的可靠性，筛选了3个结果均一致的评论。构建的语料库如表2所示。

Table 2 Corpus

表2 语料库

语料库	语料数	语料库	语料数
M_train_set	1 500	M_test_set	4 000
P_train_set	1 000	P_test_set	1 500
ALL_train_set	2 000	ALL_test_set	4 000

表2中，M_train_set为用于微博情感分类的训练语料集，M_test_set为对应的测试语料集；P_train_set和P_test_set分别为产品评论情感分类的训练语料集和测试语料集。本文基于情感词典和机器学习方法，采用表2的语料集比较在不同领域下的情感分类效果。另外，还将微博和产品评论混合，随机抽取其中2 000条语料作为训练语料集ALL_train_set，对测试语料集ALL_test_set进行情感分类。实验目的在于检验基于情感词典和朴素贝叶斯方法在无领域语料情感倾向分类中的性能。

3.2 评价标准

在对分类器的性能进行评测时，本文采用了微平均（*F*1）作为评价指标，其计算公式如式（4）所示：

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall}\times 100\% \tag{4}$$

其中，*Precision*为查准率；*Recall*为召回率。

3.3 情感分类实验

3.3.1 基于朴素贝叶斯的文本情感分类实验（实验1）

采用最大匹配算法对文本D进行分词，分词结果为$D=\{w_1, w_2, \cdots, w_n\}$。基于情感词典SOSL选择$D$中的情感词作为情感特征$V_D=\{sw_1, sw_2, \cdots, sw_m\}$，$V_D$是$D$的子集，并且有$sw_i \in SOSL$。记情感词$sw_i$在文本$D$中的情感权值为$v_{sw_i}$，则情感权值向量$\boldsymbol{V}=\{v_{sw_1}, v_{sw_2}, \cdots, v_{sw_m}\}$。对于$v_{sw_i}$，采用词频、BOOL权值、TFIDF（term frequency-inverse document frequency）、TFIDF-X四种情感权值计算方法。将$\boldsymbol{V}=\{v_{sw_1}, v_{sw_2}, \cdots, v_{sw_m}\}$输入到朴素贝叶斯分类器中，对文本$D$进行情感分类[1]。分类器的先验概率从训练中获取，将M_train_set、P_train_set作为训练语料，训练得到分类器NB_M、NB_P。测试语料库M_test_set、P_test_set中分别输入NB_M、NB_P，进行文本情感分类。分类评价指标为*F*1值，结果如图1所示。

从朴素贝叶斯分类结果看，产品评论的情感分类效果优于微博，原因在于产品评论的用词较微博更规范化，评论的样式也比较单一。从总体性能上看，本文生成的情感词典可用于文本情感分类中的特征选择，也说明了情感词的选择是有效的。

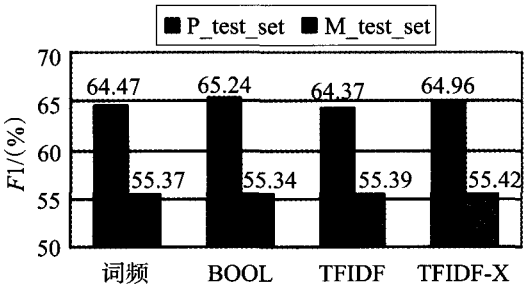

Fig.1 Result of text sentiment classification based on NB

图1 基于NB的文本情感分类结果

3.3.2 基于情感词典的情感分类实验(实验2)

利用SOSL进行微博情感分类,不需训练分类器,不用考虑分类器中的前向概率和后向概率的重新估计问题。使用词的情感倾向值进行文本情感分类,一般是对特征词的情感倾向值使用线性相加来判别文本的情感。将词的情感倾向值 $SO(w_i)$ 使用式(5)进行线性变换,使得 $SO \in [0, 1]$。

$$SO(w_i)=\begin{cases}\dfrac{1+SO_{\text{new}}(w_i)}{2}, & w_i \in W_P \\ \dfrac{1-SO_{\text{new}}(w_i)}{2}, & w_i \in W_N\end{cases} \tag{5}$$

设计SO-A分类器,SO-A是利用 SO 的和作为分类依据,如式(6)所示:

$$c_{\text{SO-A}}=\arg\max_{c_j \in C}\left\{\sum_{i=1}^{n} SO_{c_j}(w_i)\times wt(w_i)\right\} \tag{6}$$

以SOSL为情感词典,采用SO-A对测试语料集M_test_set、P_test_set和ALL_test_set进行微博情感分类,结果如表3所示。表中 PF 表示正向文本的分类微平均,NF 表示负向文本的分类微平均,F 表示整体的分类微平均。

Table 3 Result of text sentiment classification based on SO-A

表3 基于SO-A的情感分类结果

语料库	正向	负向	PF/(%)	NF/(%)	F/(%)
M_train_set	130	1 370	73.74	25.21	59.96
M_test_set	144	3 856			
P_train_set	768	232	65.21	30.14	**68.11**
P_test_set	916	584			

从表3中可以看出,对于产品评论语料,正向文本的分类效果明显比负向文本的分类效果好很多,主要有以下原因:(1)人们更喜欢正向评论,正向评论远超过负向评论;(2)仅考虑情感词,忽略了转折词、否定词等的影响;(3)负向评论文本较少,情感特征词少。对于微博语料,微博的短文本、口语化、转折语等多方面的影响,导致负向语料分类效果较差。对比NB分类方法下的性能,基于SO-A分类效果优于NB分类器。

本文将2.1节中HowNet、NTUSD和褒贬义词典的并集(记为USL,union sentiment lexicon)作为情感词典,对语料进行情感分析,分析结果如表4所示。

Table 4 Result of text sentiment classification based on USL

表4 基于USL的情感分类结果

语料库	正向	负向	$F1$/(%)
M_train_set	130	1 370	48.02
M_test_set	144	3 856	
P_train_set	768	232	63.23
P_test_set	916	584	

对比表3的实验结果,表4的实验结果中微博语料的分类效果明显较差。主要原因是微博的情感分布稀疏,需要更细粒度的情感权值来作为情感分析的标准;而产品评论分类 $F1$=63.23%<68.11%,相对微博语料的分类效果有较大的提升,主要是情感特征比较丰富,而且情感词与评价对象之间的对应关系比较清晰。

3.3.3 混合语料下的情感分类实验(实验3)

利用训练语料集ALL_train_set进行训练,得到NB分类器的基础概率模型。将语料库ALL_test_set定量200条不断输入到分类器SO-A和NB中进行测试,验证在混合语料环境下,随着语料的增加,两种分类方法的微平均 $F1$。

在利用NB进行分类时,每次输入需要先对已经输入的所有语料进行测试,得到测试的评价指标,之后使用最新一组的输入语料修正概率模型。这里采用了两种方式进行修正:一种是使用语料的标注类别;另一种是根据最大期望算法(EM)的思想,模拟对语料标注不明情况下的情感分类,即利用分类的结果作为修正概率模型。分类结果如图2所示。

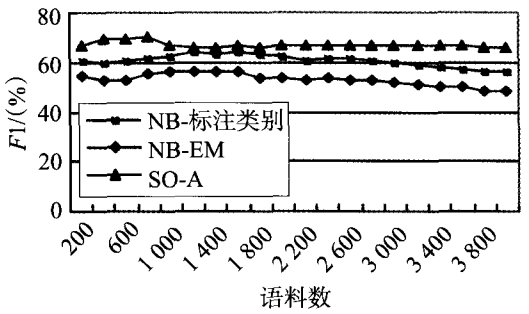

Fig.2 Result of text sentiment classification under mixed corpus environment

图2 混合语料环境下的情感分类结果

随着混合语料的增加，基于情感词典的文本情感分类的微平均先上升后下降，直趋平稳；利用NB分类器的情感分类，在两种修正概率模型下，$F1$值先上升后下降。就总体性能来说，基于情感词典的分类方法优于基于朴素贝叶斯的方法。从图2中可以总结出，基于情感词典的分类方法，选用优秀的情感权值计算方法，可有效用于多个领域的语料情感分类，并且分类性能比较稳定。实验结果也说明了本文的情感权值计算方法是有效的。

4 结束语

本文将互联网作为海量语料试验场，选用40个情感种子词，利用搜索引擎返回的共现数，通过改进的PMI算法计算通用情感词典的情感权值。采集微博语料和产品评论语料，并在混合语料下验证了情感词典的分类性能。在实验1中，将获得的情感词典用于情感特征选择，并利用基于朴素贝叶斯的方法进行情感分类，从分类的性能来看，本文生成的情感词典可用于文本情感分类中的特征选择，说明了情感词的选择是有效的。在实验2中，直接利用情感词典进行情感分类，发现分类效果优于基于朴素贝叶斯的方法（$F1$：68.11%>65.24%）。同时对比了HowNet、NTUSD和褒贬义词典的并集USL的分类性能，发现本文方法可取得更高的$F1$。说明了情感权值的计算结果可直接用于情感分类，并且分类器设计简单，分类速度快。在实验3中，在混合语料环境下对比了随着语料的增加，基于朴素贝叶斯分类和基于情感词典分类的性能。实验结果表明，基于情感词典的分类性能优于基于朴素贝叶斯的分类性能，并且具有分类性能稳定的优势。后续工作中，将考虑网络用词、表情符号等更多的情感特征，优化本文方法，构建更全面、更高质量的情感词典，并发布词典、相关语料和测试报告。

References:

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

[2] Yang Aimin, Zhou Yongmei, Lin Jianghao. A method of Chinese texts sentiment classification based on Bayesian algorithm[J]. Applied Mechanics and Materials, 2013, 263/266: 2185-2190.

[3] Lin Jianghao, Yang Aimin, Zhou Yongmei, et al. Classification of microblog sentiment based on naïve Bayesian[J]. Computer Engineering and Science, 2012, 34(9): 86-90.

[4] Ren Yong, Kaji N, Yoshinaga N, et al. Sentiment classification in resource-scarce languages by using label propagation[C]// Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25), Singapore, Dec 16-18, 2011: 420-429.

[5] Escalante H J, Montes-Y-Gómez M, Solorio T. A weighted profile intersection measure for profile-based authorship attribution[C]//Proceedings of the 10th Mexican International Conference on Artificial Intelligence (MICAI '11). Berlin, Heidelberg: Springer-Verlag, 2011: 232-243.

[6] Jung J J. Maximum entropy-based named entity recognition method for multiple social networking services[J]. Journal of Internet Technology, 2012, 13(6): 931-937.

[7] Xu Ge, Meng Xinfan, Wang Houfeng. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources[C]//Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 1209-1217.

[8] Dai Liuling, Liu Bin, Xia Yuning, et al. Measuring semantic similarity between words using HowNet[C]//Proceedings of the 2008 International Conference on Computer Science and Information Technology (ICCSIT '08). Washington, DC, USA: IEEE Computer Society, 2008: 601-605.

[9] Liu Weiping, Zhu Yanhui, Li Chunliang, et al. Research on building Chinese basic semantic lexicon[J]. Journal of Computer Applications, 2009, 29(11): 2882-2884.

[10] Chang Xiaolong, Zhang Hui. Construction of Chinese polarity lexicon by integration of morpheme features[J]. Journal of Computer Applications, 2012, 32(7): 2033-2037.

[11] Zhu Yanhui, Li Chunliang, Xu Yeqiang, et al. A method of emotional feature extraction in Chinese text based on multiple lexicons[J]. Journal of Hunan University of Technology, 2011, 25(2): 42-46.

[12] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10), Valletta, Malta, 2010: 2200-2204.

[13] Hamouda A, Marei M, Rohaim M. Building machine learning based senti-word lexicon for sentiment analysis[J]. Journal of Advances in Information Technology, 2011, 2(4): 199-203.

[14] Maks I, Vossen P. A lexicon model for deep sentiment analysis and opinion mining applications[J]. Decision Support Systems, 2012, 53(4): 680-688.

[15] Wang Guangwei, Araki K. Modifying SO-PMI for Japanese Weblog opinion mining by using a balancing factor and detecting neutral expressions[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL '07), Rochester, USA, Apr 22-27, 2007. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007: 189-192.

[16] Yang Aimin, Lin Jianghao, Zhou Yongmei, et al. Research on building a Chinese sentiment lexicon based on SO-PMI[J]. Applied Mechanics and Materials, 2013, 263/266: 1688-1693.

附中文参考文献：

[3] 林江豪, 阳爱民, 周咏梅, 等. 一种基于朴素贝叶斯的微博情感分类[J]. 计算机工程与科学, 2012, 34(9): 86-90.

[9] 柳位平, 朱艳辉, 栗春亮, 等. 中文基础情感词词典构建方法研究[J]. 计算机应用, 2009, 29(11): 2882-2884.

[10] 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.

[11] 朱艳辉, 栗春亮, 徐叶强, 等. 一种基于多重词典的中文文本情感特征抽取方法[J]. 湖南工业大学学报, 2011, 25(2): 42-46.

YANG Aimin was born in 1970. He received his Ph.D. degree in computer software and theory from Fudan University in 2005. Now he is a professor at Guangdong University of Foreign Studies. His research interests include text sentiment analysis, machine learning and pattern classification, etc.

阳爱民(1970—),男,湖南永州人,2005年于复旦大学计算机软件与理论专业获得博士学位,现为广东外语外贸大学教授,主要研究领域为文本情感分析,机器学习,模式分类等。发表学术论文50多篇,其中被SCI、EI和ISTP检索25篇,出版著作2部,主持广东省自然科学基金项目、教育部人文社会科学研究青年项目、教育部新世纪优秀人才支持计划项目、广东省社科规划项目等。

LIN Jianghao was born in 1985. He is a master candidate at Guangdong University of Foreign Studies. Her research interests include natural language processing, machine learning and text sentiment classification, etc.

林江豪(1985—),男,广东揭阳人,广东外语外贸大学硕士研究生,主要研究领域为自然语言处理,机器学习,文本情感分类等。发表学术论文7篇,其中4篇为EI检索。

ZHOU Yongmei was born in 1971. She received her M.S. degree in software engineering from Central South University in 2006. Now she is a professor at Guangdong University of Foreign Studies. Her research interests include text sentiment analysis, microblog sentiment analysis and machine learning, etc.

周咏梅(1971—),女,湖南永州人,2006年于中南大学软件工程专业获得硕士学位,现为广东外语外贸大学教授,主要研究领域为文本情感分析,微博情感分析,机器学习等。主持国家社科基金项目、教育部人文社会科学研究青年项目、广东省科技计划项目等。