ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 6,Nov. 2016,pp. 1220－1227
DOI:10.16337/j.1004-9037.2016.06.019
© 2016 by Journal of Data Acquisition and Processing
http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种改进的融合关联词典的微博倾向性分析方法

赵　军[1,2]　王　红[1,2]　朱华方[1,2]

(1.山东师范大学信息科学与工程学院,济南,250014; 2.山东省分布式计算软件新技术重点实验室,济南,250014)

摘　要:大多数研究者对微博倾向性分析过多关注的是情感词、形容词和否定词,忽略了关联词对其情感倾向的影响。为了提高微博情感倾向性分析的准确率,提出了融合关联词的微博倾向性分析方法,考虑微博文本中形容词、程度副词以及关联词之间的组合关系。本文充分考虑了关联词的结构特点并在已有词典的基础上构建专门用于微博倾向性分析的微博词典、否定词词典和关联词词典,同时考虑到网络新词对微博倾向性的影响,还构建了一个全新的网络新词词典。借助支持向量机(Support vector machine,SVM)将微博文本分为负向、正向和中性3类,通过结合情感词典和SVM的方法提高微博文本倾向性分析的准确率。通过对COASE 2014数据实验可以表明,本文方法对微博倾向性分析取得了较好的效果。

关键词:中文微博;倾向分析;支持向量机;关联词
中图分类号:TP391　　**文献标志码**:A

Improved Method for Analyzing Microblog Orientation Based on Association Lexicon

Zhao Jun[1,2], Wang Hong[1,2], Zhu Huafang[1,2]

(1. School of Information Science and Engineering, Shandong Normal University, Jinan, 250014, China; 2. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan, 250014, China)

Abstract: At present, a larger number of researchers focus on Micro-blog orientation on the emotional words, adverb and negative words without considering the impact of connectives. To improve the accuracy of orientation analysis, a method of analyzing Mico-blog orientation is proposed. In the paper, we sufficiently analyze the structure characteristics of associated words and consider the combination laws of negative words , adversative words and conjunctions in Microblog. In addition, a specific dictionary is created based on the existing resources, which contains a turning words lexicon, a connective lexicon and a negative words lexicon. At the same time, we take into account the impact of new network words and phrases of the microblog text, so we also build a new network words dictionary. Therefore, the Microblog texts are classified into three categories including negative, positive and neutral one by support vector machine (SVM). By combining Lexicon-based and SVM machine learning method, better accuracy of classification can be achieved. Experimental results verify that the method achieves higher classification accuracy through experiments using COASE 2014.

Key words: Chinese microblog; orientation analysis; support vector machine (SVM); connectives

---

基金项目:国家自然科学基金(61373149,61472233)资助项目;山东省科技计划(2012GGX10118,2014GGX101026)资助项目;山东省教育科学规划(ZK1437B010)资助项目。
收稿日期:2015-06-05;**修订日期**:2015-06-30

引　　言

自从进入 Web 2.0 时代以来,社交媒体不断涌现,微博的出现更是加快了实时消息的传递,也极大丰富了人们的日常生活。微博简短、无特定格式、便于发布以及实时交互的特点迎合了当代人们的生活节奏。越来越多的用户喜欢在微博平台分享信息、评论话题以及交流观点和情感等。通过对这些简短的微博信息进行挖掘和分析从而得到微博的情感倾向,它可以广泛用于社会舆情分析、微博营销、影视评论和产品在线跟踪与质量评价等领域。截至 2014 年 12 月,新浪官方发布消息称,新浪微博用户数已经突破 3.9 亿,如此庞大的用户群产生了海量的微博信息,这些海量的微博信息中隐藏着巨大的、可供挖掘的有用信息;所以越来越多的研究者开始对微博的情感倾向性进行探究。微博的情感倾向性分析是指通过挖掘和分析微博文本中的立场、观点、看法和情绪等主观信息,对文本的情感倾向做出正面或者负面的类别判断。目前,国内外研究者已经在文本倾向性分析领域进行了大量的深入研究,并且取得了很多辉煌成果。对文本的倾向性研究主要有以下两种方法:基于特征分类的方法和基于情感知识的方法。基于特征分类的方法主要使用机器学习的方式,把文本情感倾向分析问题转化为传统的分类问题来解决,抽取文本中的特征来做训练,使其训练出一个模型,然后使用这个模型对需要判别情感极性的微博文本进行预判。基于情感知识的方法主要通过建立情感词典或领域情感词库的方式,通过比较文本中的正负情感词个数来判别该文本的情感极性。对文本情感的倾向性研究起源于 20 世纪 90 年代,最初,文献[1]在各种文本数据的基础上构建了语义词典,为以后的基于词典的情感倾向分析提供了便利。文献[2]在不同的文本数据集上同时考虑形容词和连词之间的约束规则,从而对英文文本进行情感倾向性分析。近年来,文献[3]在汽车评论的语料集上通过提取主要特征的方法,实现了对评论文本从特征级别的倾向性分析。文献[4]基于主题目标先对 Twitter 语料进行分类,再对已分类的文本进行倾向性分析,该方法取得了不错的分析效果。相应地,文献[5]针对中文微博提出了一种基于主题相关性分类的微博话题立场预判方法,该方法也取得了不错的预判结果。文献[6]利用已经构建好的情感词库中的词作为种子词,来确定与种子词有一定句法关系的词语的情感极性,然后对文本中所有句子加权求和,从而得到整个文本的情感倾向。这种基于情感知识方法判别文本极性的重点就在于情感词库的构建、情感词组合以及情感极性求和。文献[7]首次使用了基于特征分类(机器学习)的方法实现对文本的情感极性分析,使用的是 N-gram 词语特征与词性特征,同时对比了支持向量机(Support vector machine,SVM)、朴素贝叶斯(Native Bayesian,NB)和最大熵(Maximum entrpy,ME)这 3 类常用分类模型,最后通过实验发现,Unigram 取得了较好的分类效果。文献[8]提出了基于转折句式的文本情感倾向性分析,通过列举各种否定词和关联词重数来分析其对文本极性的影响,实验结果显示该方法对正面极性的准确率达到了 79.6%,对负面极性文本的预判准确率最高为 81.2%。

微博作为一个新兴的社交工具,不仅对人们的日常生活有重大影响,而且对商业甚至政府决策也有不可低估的作用,因此吸引了越来越多的学者对其进行研究。文献[9-11]分别采用机器学习和情感词典的方法对 Twitter 文本极性进行分析。由于对中文微博研究起步较晚,所以相关文献也相对较少,文献[12]通过共用 4 个特征,采用 SVM 的方法对新浪微博数据进行实验研究,该实验结果显示,使用主题无关特征获得的最高准确率为 66.467%,考虑主题相关这一特征获得的准确率为 67.283%。在微博倾向性分析中,基于情感知识的方法主要是通过建立情感词典或领域情感词库的方式,通过比较文本中的正负情感词个数来判别该文本的情感极性。对于微博中没有感情词的文本,可以粗略地认为其倾向性为中性,例如,“明天会下雨”,这个句子没出现任何能表示情感的词语,所以认为其倾向性为中性。对于文本中有感情词的文本,通过计算感情词极性的加权和来判断文本的倾向性,微博文本情感倾向性判别方式为

$$情感=\left\{\begin{array}{ll}正情感词数+正表情数>负情感词数+负表情数 & 正向\\ 正情感词数+正表情数=负情感词数+负表情数 & 中性\\ 正情感词数+正表情数<负情感词数+负表情数 & 负向\end{array}\right\}$$

文献[13]又考虑了否定词对情感词极性的影响，例如词语“喜欢”本来的情感极性为正极性，但是如果和否定词“不”连用就会组合成“不喜欢”，极性转变为负极性，若与两个否定词连用组合成“不是不喜欢”，那么情感极性不变。通过实验可以看出，考虑否定词这一特征对微博倾向性分析具有重大意义。本文也考虑否定词这一重要因素，并且构建专用于微博情感分析的否定词词典，并且对词典中的所有词赋值为-1。让否定词权值和感情词的权值相乘，这样所得到的乘积就为组合词的极性。判断规则也可以是：统计一句话中的否定词个数，若个数为偶数则该句的倾向性不变；若否定词的个数为奇数，那么语句的倾向性发生逆转，可以表达为

$$\mathrm{IsReverse}=\left\{\begin{array}{ll}\mathrm{true} & n\%2!=0\\ \mathrm{false} & n\%2==0\end{array}\right\} \tag{1}$$

式中：n为每个句子中的否定词个数。文献[14]考虑了程度副词对微博文本的倾向性影响，提出了程度副词的4个量级：极量、高量、中量和低量，其对应的权重值依次递减，考虑程度副词这一因素也提高了微博情感的分类精度。综合分析现有的研究成果，发现大多数研究者并没有考虑关联词对微博倾向性分析的重大影响。本文在前人的研究基础上主要做了以下4方面工作：(1)探测网络新词并将其添加到网络新词词典中。(2)考虑否定词和副词对文本极性的影响。(3)考虑关联词对微博倾向性的影响。(4)结合支持向量机模型对微博倾向性进行分析。实验表明本文方法取得了不错的结果。本文的主要工作是检测微博文本的正负极性问题，可将其简化为经典的二分类问题处理，文献[15]指出支持向量机是解决二分类问题的首选方法，所以本文最后结合支持向量机模型对微博倾向性进行分析，实验表明本方法取得了较好的结果。

1 基于关联词的微博倾向性分析方法

1.1 微博新词发现

随着Internet的快速发展，网络新词也如雨后春笋般衍生出来，这些网络新词和传统的词语有着很大区别，这些词往往不规则但又具有强烈的感情色彩，比如：“宝马女、毛线、颜值高、稀饭(喜欢)……”。微博无固定书写格式，人们更喜欢口语化的表达，大量的网络新词充斥其间，这些网络新词往往带有浓厚的感情色彩，因此发现这些新词对微博倾向性分析至关重要。本文借助HowNet基础情感词典，通过人工剔除那些不常用和情感倾向不明确的情感词，最终得到6 124个常用情感词作为基础情感词库，其中包含3 219个褒义词和2 905个贬义词。在此基础上，本文采用分词和逐点互信息(Semantic orientation-pointwise mutual information，SO-PMI)来计算微博新词的极性。首先，计算两个词语的点互信值为

$$\mathrm{PMI}(W_i,W_{i+1})=\log\left[\frac{p(W_i\&W_{i+1})}{p(W_i)p(W_{i+1})}\right] \tag{2}$$

式中：$P(W_i\&W_{i+1})$为两个词语W_i和W_{i+1}共同出现的概率；而$P(W_i)$与$P(W_{i+1})$则为这两个词语单独出现的概率。两个词语在数据集的某个小范围内的贡献频率越大，则表明这两个词的关联程度越高，反之，关联度越小。其中，$p(W_i)=\dfrac{df(W_i)}{N}$，$df(W_i)$为文档集中的词$W_i$出现的频数，$N$为总文档数。同理，$p(W_{i+1})=\dfrac{df(W_{i+1})}{N}$，$df(W_{i+1})$为文档集中词$W_{i+1}$出现的频数，$N$为文档总数。同理，$p(W_i\&W_{i+1})=\dfrac{df(W_i\&W_{i+1})}{N}$，$df(W_i\&W_{i+1})$为词$W_i$和$W_{i+1}$同时在文档集中的出现次数，$N$仍为文档集中的文档总数；则式(2)可以表示为

$$\mathrm{PMI}(W_i,W_{i+1})=\log_2\frac{N\times df(W_i\&W_{i+1})}{df(W_i)df(W_{i+1})} \tag{3}$$

网络新词倾向性判别步骤如下:

(1)对微博语料进行预处理。本文采用中科院的 ICTCLAS 分词工具对微博文本进行分词,过滤掉其中的停用词和构词能力较差的词语,生成网络新词候选集,记为 CD=$\{cd_1,cd_2,\cdots,cd_n\}$。

(2)进行二元语法模型的统计。统计的对象为 CD 集合中两两之间的词,统计的元素为 W_i,这样就可以得到二元词集合。

(3)将步骤(2)中所得的二元词集合与知网 HowNet 中的词进行比对,从二元词集中删除那些已经包含在 HowNet 中的词项。

(4)使用 SO-PMI 算法计算 W 的情感极性。其基本思想是:首先从 HowNet 词典中选择一组常用褒义词 P_w 和贬义词 N_w 作为基准词,用词语 W_i 与 P_{wi} 的点互信息减去 W_i 和 N_{wi} 的点互信息,根据计算的差值正负来判断词语 W_i 的情感极性,判断公式为

$$\text{SO-PMI}(W_i)=\sum_{i=1}^{n}\mathrm{PMI}(W_i,P_{w_i})+\sum_{i=1}^{n}\mathrm{PMI}(W_i,N_{w_i}) \tag{4}$$

$$\text{SO-PMI}(W)\begin{cases}>0 & \text{将 } W \text{ 添加到正面词库中}\\ =0 & \text{将 } W \text{ 从候选词集中删除}\\ <0 & \text{将 } W \text{ 添加到负面词库中}\end{cases} \tag{5}$$

通过 SO-PMI 算法,可以获得网络流行度较高的新词情感倾向,根据互信息值分别分类到正面词典和负面词典中,其算法流程如下:设 CD=$\{cd_1,cd_2,\cdots,cd_n\}$ 为网络新词候选集,正向情感词典=posColl,负向情感词典=negColl,设定正向种子词=PosSeg,负向种子词=NegSeg。

算法 1　构建网络新词词典

输入:网络新词候选集 CD

输出:posColl 和 negColl

Set posColl=$\{\varnothing\}$; negColl=$\{\varnothing\}$; $\varphi=\{1,2,3,\cdots,i\}$;

posColl ←posColl ∪ {PosSeg}; negColl ←negColl ∪ {NegSeg};

(1)　FOR all $i\in\varphi$ do

(2)　　IF $cd_i\notin$ posColl AND $cd_i\notin$ negColl THEN

(3)　　SO-PMI=PMI(W_i, PosSeg)-PMI(W_i, NegSeg) ;//计算一个候选词语的极性差值。

(4)　　IF SO-PMI>0 THEN

(5)　　posColl←posColl∪{cd_i} //若极性>0,把这个词语添加到正极性词典。

(6)　　ELSE IF SO-PMI<0 THEN

(7)　　negColl←negSeg∪{cd_i}//若极性<0,把这个候选词添加到负极性词典。

(8)　　ELSE

(9)　　CD←CD−W_i//若候选词没有感情极性,则从候选词集合将其删除。

(10)　　$\varphi=\varphi-i$//候选词集合元素依次缩小。

(11)　　END IF

(12)　　END ELSE IF

(13)　　END IF

(14)END FOR

1.2　关联词简介

把两个或两个以上在意义上有密切联系的句子组合在一起的词语统称为关联词语,在汉语中主要

包含转折、假设、并列、递进、选择、因果、承接和条件的 8 类关联词，其中，转折、递进和并列关联词对微博倾向分析的影响更突出，暂且只考虑这 3 类词，如表 1 所示。

表 1 关联词简介表

Tab. 1 Profile of related words

类型	常用词	示例
并列关系	既……，又……；那么……，那么 一方面……，另一方面……	她既美丽又漂亮
递进关系	不仅……，而且…… 不但……，还……	他不仅成绩好而且人缘更好
转折关系	虽然……，但是…… 本该……，却……	他本该是冠军的，却失误了

微博文本倾向性分析算法流程如下：微博情感词典 = MicDictionary，HowNet 词典 = HowDictionary，否定词典 = NegDictionary，副词词典 = AdjDictionary，关联词典 = ConDictionary，网络新词词典 = NetDictionary，表情词典 = EmoDictionary，词语的倾向权值 = WO，文本倾向权值 = SO。

算法 2 Microblog Orientation Analysis

输入：A Microblog Text

输出：Microblog Orientation

MicDictionary←HowDictionary ∪ NetDictionary ∪ EDictionary

str←{Microblog Text}；

(1) While(str. read())

(2) For all $i \in$ str. Length DO

(3) IF (str. words$_i \notin$ MicDictionary) THEN

(4) SO←0；EXIT；//不含感情词，算法结束

(5) ELSE：

(6) IF (str. words$_i \in$ NegDictionary) THEN

(7) IF(count % 2 == 0) THEN

(8) WO$_i$←WO$_i$；//偶重否定，极性不变

(9) ELSE WO$_i$←−WO$_i$/奇重否定，极性转变

(10) IF (str. words$_i \in$ AdjDictionary) THEN

(11) WO$_i$←WO$_i$×Wad$_j$；//副词改变原词极性强度

(12) ELSE WO$_i$←WO$_i$

(13) IF(str. words$_i \in$ ConDictionary) THEN

(14) SWITCH(char ConnetivesType)

(15) CASE 'Coordinative'：//并列关系，极性不变

(16) WO$_i$←WO$_i$ BREAK；

(17) CASE 'Progressive'：//递进关系，极性增强

(18) WO$_i$←WO$_i$×1.5；BREAK；

(19) CASE 'Adversative'：//转折关系，极性逆转

(20) WO$_i$←−WO$_i$；BREAK；

(21) DEFAULT ：EXIT SWITCH ；

(22)　　ELSE $WO_i \leftarrow WO_i$;

(23)　　END ELSE

(24)　　$SO=\sum_{i=1}^{N} WO_i$//计算整个文本极性,$N=$str. Length

(25)　END FOR

(26)　END WHILE

2　评价与分析

2.1　数据集和评价方法

本文采用 COASE 2014 任务 4 中的 10 000 条微博文本作为实验数据,有 6 名研究生对其情感倾向作出人工标注,最后汇总到一起,对倾向性标注不一致的文本采取多人讨论的形式进行再次标注,仍有异议的文本舍弃不用,最后得到 8 767条具有明显感情倾向的标注数据,标注数据的分布情况如表 2 所示。

表 2　测试数据分布表

Tab. 2　Test Data Set

数据类型	条数
数据总数	8 767
正向情感	1 273
中性情感	6 284
负向情感	1 210

本文借用 COASE 2014 提供的评价方式,以准确率(Precision)、召回率(Recall)和 F_1 值作为评价标准,则

$$\text{Precision}=\frac{\text{判断正确的类别数目}}{\text{判断为该类别的数目}} \tag{6}$$

$$\text{Recall}=\frac{\text{判断正确的类别数目}}{\text{应该判断为该类别的数目}} \tag{7}$$

$$F_1=\frac{2\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{8}$$

2.2　实验结果与分析

本文采用对比实验的方法,由于测试样本有限,采用五折交叉验证的方式分别与比较成熟的基于 HowNet[17] 方法、基于支持向量机的方法[18] 和基于转折句式的分析方法[8] 作比较,从而分析本文在考虑关联词规则的前提下再使用 SVM 的优势和不足;3 种评价方式的效果对比见图 1～3。从 3 幅对比图中可以看出,与已有的文本情感倾向性分析方法相比,本文融合关联词典的倾向性分析方法取得了较好的效果。基于 HowNet 的微博倾向性分析方法取得的效果不是特别理想,造成这种结果的影响主要有,首先该方法太多依赖微博语料中的情感词,通过统计文本中正负极性的情感词个数来分析文本倾向,没有考虑微博文本中的网络新词。如"谁规定白富美必须要嫁给高富帅,这是什么毛线道理啊",网络热词"白富美"和"高富帅"代表的是形象和家庭都很优越的年轻男女,但是我们使用中科院的 ICTCLAS 分词工具对这条文本进行切词会发现,"白富美"被切成了"白富""美要",而"高富帅"则被分成了"高富"和"帅",对原来词语的意向表达发生了严重偏离,并且"毛线"这个词语在网络上的意思蕴含着一种鄙视的情感。如果不能及时地发现这些网络新词,就会对这些词语造成语义的偏离,从而导致分析结果不够精确。其次,基于 HowNet 的微博倾向分析方法,没有考虑形容词、程度副词对微博倾向的影响,而这两类词语对微博倾向情感强度的影响至关重要,如果不慎重考虑也会降低分析结果的准确度。基于 SVM 的倾向性分析方法相比于 HowNet 方法,在准确率和召回率上都有较大的提高,但是这种方法高度依赖训练集的质量;如果标注的训练集质量太差,使用该方法训练出来的分类模型质量也很差。尽管 SVM 是一种非常理想的分类方法,但是由于中文微博表达方式多种多样,句子结构随意,如果不考虑句子的结构和关联关系,那么就不能很好地考虑句子的语法成分和上下文信息,从而导致分析结果存在一些误差。

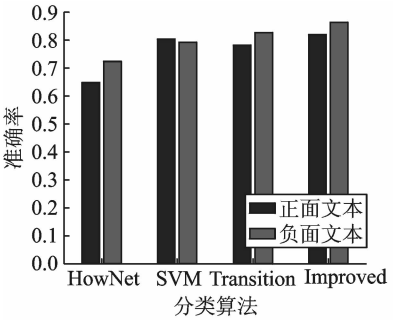

图 1 4 种算法的准确率比较图

Fig. 1 Accuracy comparison chart

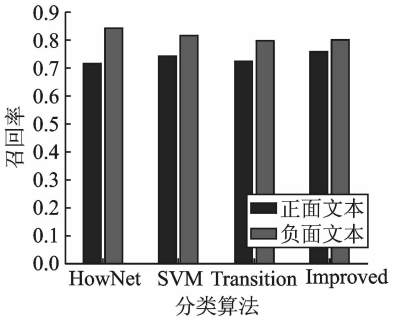

图 2 4 种算法的召回率对比图

Fig. 2 Recall comparison chart

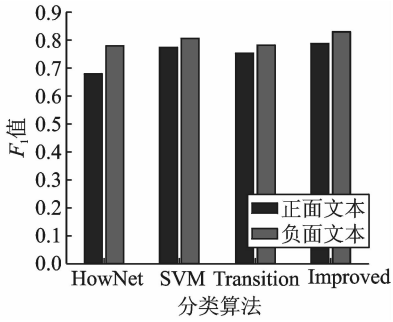

图 3 4 种算法的 F_1 值对比图

Fig. 3 F_1-measure comparison chart

文献[7]提出了基于转折句式的文本情感倾向性分析方法，该方法比前两种分析方法取得了更好的分析结果，专门针对转折句式的特殊结构提出了一种启发式规则，因此能够有效地分析含有转折句式的微博文本，但是其召回率比前两种方法有所下降，可能因为并不是所有的微博文本中都含有大量的转折句式。从图 1 可以看出，本文方法比基于转折句式的分析方法[7]准确率略高，主要因为本文采用了 COASE 2014 任务 4 的数据集，该数据集是句子级别而非篇章级别，着重考虑了带有转折句式的句子，本文的方法主要从关联词出发，由于考虑的粒度相对较小，所以本文方法在该数据集上表现出了更高的准确率。本文方法在 HowNet 情感词典的基础上，充分考虑关联词对文本倾向性的影响，构建了基于关联词的学习规则，并将该学习规则与 SVM 方法相结合对微博进行倾向性分析；通过实验对比，证明了微博新词和关联词提高了倾向性分析的准确率。从图 2 可以看出，本文方法的召回率与其他方法相比并没有明显优势，产生这种结果的原因可能是因为实验数据中，并没有广泛存在关联词，所以提出的基于关联词的方法对这样的数据集表现出了较低的召回率。

3 结束语

微博在当今社交媒体中扮演了越来越重要的角色，对其进行倾向性分析对政府、电商和个性化推荐等领域具有重要意义。本文虽然在前人研究的基础上将关联词考虑进去，对微博的倾向性分析取得了良好的结果，但是也存在以下 3 个难题：(1)由于并不是所有的文本句子都包含关联词，所以该方法的召回率会有所下降。(2)由于汉语博大精深，语法结构复杂，单单考虑关联词还不够精确。(3)没有考虑程度副词和形容词的词序对感情强度的影响，如“不特别漂亮”和“特别不漂亮”就是因为词序不同而表达了不同的情感倾向。下一步的工作应该考虑以上 3 点不足，再结合机器学习、语义规则等方法，集成众多方法的优点，还要从中选择最具代表性的特征对其进行分析，从而提高倾向性的准确率。

参考文献：

[1] Riloffe S J. A corpus-based approach for building semantic lexicons[J]. Association for Computational Linguistics, 1997, 1: 117-124.

[2] Hatzivassiloglou V, Mckeown K R. Predicting the semantic orientation of adjectives[C]// Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Linguistics. Somerset: ACL, 1997: 174-181.

[3] Miao Qingliang, Li Qiudan, Dai Ruwei. Amazing: A sentiment mining and retrieval system[J]. Expert System with Applications: An International Journal, 2009, 36(3): 7192-7198.

[4] Jiang Long, Yu Mo, Zhou Ming. Target-dependent Twitter sentiment classification[C]// Proceeding of the 49th Annual

Meeting of the Association for Compututational Linguistic: Human Language Technology. Oregon, USA: [s. n.], 2011: 151-160.

[5] 王明元,贾焰. 一种基于主题相关性分类的微博话题立场研判方法[J]. 信息网络安全, 2014,9:17-20.
Wang Mingyuan , Jia Yan. A method of discriminating Microblog topic position based on the text classification with correlation of subject[J]. Netinfo Security, 2014,9:17-20.

[6] Tang D, Qin B, Liu T, et al. Learning sentence representation for emotion classification on Microblogs[C]// Natural Language Processing and Chinese Computing. Berlin, Heidelberg: Springer, 2013:212-223.

[7] Kumar S, Morstatter F, Liu Huan. Twitter data analytics[M]. New York, USA: Springer, 2014:134-141.

[8] 邸鹏,李爱萍,段利国. 基于转折句式的文本情感倾向性分析[J]. 计算机工程与设计, 2014,35(12):4290-4295.
Di Peng, Li Aiping, Duan Liguo. Text sentiment polarity analysis based on transition sentence[J]. Computer Engineering and Design, 2014,35(12):4290-4295.

[9] Luciano Barbosa, Feng Junlan. Robust sentiment dection on Twitter from biased and noisy data[C]// Coling 2010, International Conference on Computational Linguistics. Beijing, China: [s. n.], 2010:36-44.

[10] Tsolmon B, Kwon A R, Lee K S. Natural language processing and infornation system[M]. Berlin, Heidelberg: Springer, 2012:265-270.

[11] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining[C]// Proceedings of the Seventh Conference on International Language Resources & Evalution. Velletta, Malta: [s. n.], 2010:1320-1326.

[12] 谢丽星,周明,孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报,2012,26(1):74-87.
Xie Lixing, Zhou Ming, Sun Maosong. Hierarchical structure based hybrid approach to sentiment analysis of Chinese Micro blog and its feature extraction[J]. Journal of Chinese Information Processing, 2012,26(1):74-87.

[13] 唐波,陈光,王星雅, 等. 微博新词发现及情感倾向性判断分析[J]. 山东大学学报:自然科学版,2015,50(1):21-25.
Tang Bo, Chen Guang, Wang Xingya, et al. Analysis on new word detection and sentiment orientation in Micro-blog[J]. Journal of Shandong University: Nature Science, 2015,50(1):21-25.

[14] 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) :2131-2137.
He Fengying. Orientation analysis for Chinese blog text based on semantic comprehension[J]. Journal of Computer Application, 2011,31(8):2131-2137.

[15] 王振宇,吴泽衡,胡方涛. 基于 HowNet 和 PMI 的词语情感极性计算[J]. 计算机工程,2012,38(15):188-193.
Wang Zhenyu, Wu Zeheng, Hu Fangtao. Words sentiment polarity calculaion based on HowNet and PMI[J]. Computer Engineering, 2012,38(15):188-193.

[16] 苟博,黄贤武. 支持向量机多分类方法[J]. 数据采集与处理,2006, 21(3):334-339.
Xun Bo, Huang Xianwu. SVM multi-class classification[J]. Journal of Data Acquisition and Processing, 2006, 21(3): 334-339.

[17] 朱嫣岚,闵锦,周雅倩. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报,2006,20(1):14-20.
Zhu Yanlan, Min Jin, Zhou Yaqian. Semantic orientation computing based on HowNet[J]. Journal of Chinese Iinformation Processing, 2006,20(1):14-20.

[18] 夏火松,陶敏,王一,等. 停用词表对基于 SVM 的中文文本情感分类的影响[J]. 情报学报,2011,30(4):347-352.
Xia Huosong, Tao Min, Wang Yi, et al. The information of stop word removal on the Chinese text sentiment classification based on SVM technology[J]. Journal of the China Society for Scientific and Techical Information, 2011,30(4):347-352.

作者简介:

赵军(1989-),男,硕士研究生,研究方向:大数据、云计算和数据挖掘等,E-mail:1170008793@qq.com。

王红(1966-),女,教授,博师生导师,研究方向:大数据、复杂网络和数据挖掘等。

朱华方(1992-),男,硕士研究生,研究方向:大数据、云计算和数据挖掘等。