

文章编号:1674-2869(2015)10-0045-06

# 一种基于领域情感词典的网络评论倾向分析方法

何成万,王 格

武汉工程大学计算机科学与工程学院,湖北 武汉 430205

**摘 要:**移动互联网的快速发展使得网络上数据量剧增,如何从纷繁复杂的信息中提取出对人们有用的信息就成为一个亟待解决的课题.本文提出一种改进的基于情感词典的倾向分析方法,该方法在情感词典中加入领域情感词,并且通过构建辅助词典集来进行辅助分析.同时给出了一种半自动的词典维护方法来发现新词和更新词典集.通过对手机领域的评论进行文本级的情感倾向分析,正面情感分析的准确率和召回率达到0.713和0.769,负面情感的准确率和召回率达到0.738和0.706,与传统基于情感词典的方法相比准确率和召回率都有较大提高.

**关键词:**情感倾向分析;情感词典;辅助词典集

**中图分类号:**TP311

**文献标识码:**A

**doi:**10.3969/j.issn.1674-2869.2015.10.009

## 0 引 言

随着互联网的普及和互联网技术的发展,越来越多的人选择在网上购物、发表评论.如何从如此多纷繁复杂的信息中提取出对人们有用的信息就成为一个亟待解决的课题.文本情感倾向性分析的目的是对整篇文本所体现出的态度,即文本中的主观信息进行判断<sup>[1]</sup>.

目前进行情感倾向分析主要的方法为基于机器学习的方法和基于情感词典的方法.基于机器学习的方法涉及特定领域,在领域相关的文本倾向性分析中有较好的表现,但是对于领域无关的文本进行分析时,效果不是很好.基于情感词典的方法核心思想是构造情感词典,对待分析的文本进行预处理之后,根据情感词典进行情感词匹配,这种方法通用性强,不需要语料处理工作,但是对于特定领域的文本,倾向分析效果不是很好.

本文以基于情感词典的方法为基础,借鉴机器学习的思想,在情感词典中加入领域情感词,并构造辅助词典集合,提出一种改进的基于情感词典的倾向分析方法.对于动态词,本文给出了一种通过处理评论主体来处理动态表达的方法,同时给出一种有监督的词典维护方法.

## 1 相关研究

观点挖掘按照挖掘粒度来分可以分为文本

级、句子级、词语级.文本级的挖掘又叫做情感倾向分析或文本分类,目的是判断文本是褒义、贬义还是中性.

2002年Turny就提出了基于种子词汇来发现情感词的方法<sup>[2]</sup>.Pang等使用了贝叶斯、最大熵、支持向量机(SVM)等机器学习的方法来构造分类器,并对这几种方法进行了对比<sup>[3]</sup>.Nozomi Kobayashi等构建了一个模式库,收录了8种命中率比较高且比较准的模式用来提取评价主体、评价方面和评价之间的关系<sup>[4]</sup>,随后做了大量的语料工作,研究了主体和属性、属性和属性之间的层次关系,还对句子间的关系抽取做了一定的研究<sup>[5]</sup>.Marrese等考虑到用户对不同的产品发表不同的评论,找出旅游领域的特征,构造出更复杂的扩充和自然语言处理模型用于旅游领域的挖掘<sup>[6]</sup>.Penalver等通过在特征选择阶段使用本体来提高基于特征的观点挖掘,同时给倾向性分析提出了一种基于向量的分析方法<sup>[7]</sup>.

朱嫣岚提出了基于知网的语义倾向计算<sup>[8]</sup>.孙建旺等利用词典和机器学习相结合的方法来进行中文微博的情感分析<sup>[9]</sup>.肖红等提出结合句法分析和情感词典的网络舆情分析的方法<sup>[10]</sup>.王勇等构建一种情感极性词典来对微博情感进行分类<sup>[11]</sup>.谢松县等提出一种用语义关系构建情感词典的方法将中文情感词转换为对应的英文,可以解决同一个词语在不同语境下的倾向值分析<sup>[12]</sup>.徐晓丹等

收稿日期:2015-09-14

基金项目:国家自然科学基金(61272115,60873024)

作者简介:何成万(1970-),男,湖北荆门人,教授,博士.研究方向:软件工程.

利用情感词典及特征加权相结合的方法来进行倾向性判别<sup>[13]</sup>. 周咏梅等针对微博上的网络用语, 使用上下文熵和词频统计阈值在过滤后的微博的语料部分中来滤筛选网络用语, 构建网络情感词典<sup>[14]</sup>.

## 2 词典构造

本文构造的词典主要包括情感词典和辅助词典集, 各词典的关系如图 1 所示.

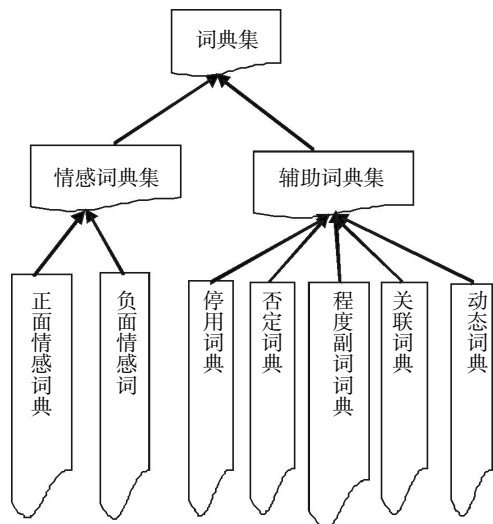


图 1 词典集结构关系

Fig.1 Structural relationship of dictionaries

### 2.1 情感词典构造

将 HowNet<sup>[15]</sup>发布的正面评价词典、正面情感词典、情感词汇本体<sup>[16]</sup>中极性为“1”的词合并去重, 构成正面基础情感词典, 且词语的情感倾向值都为“+1”; 将 HowNet 中的负面评价词典、负面情感词典、情感词汇本体中极性为“2”的词合并去重, 构成负面基础情感词典, 且词语的情感倾向值都为“-1”.

基础情感词典是通用词典, 这种词典的优点通用性强. 但是通用的情感词典在领域相关的文本分析中, 面对含有某些领域相关的情感词的评论的时候并不能准确分析出结果. 为了增加对特定领域的识别能力, 本文借鉴机器学习的思想, 对特定领域进行语料分析和学习, 扩充基础情感词典. 基础情感词典的扩充过程如下:

step1: 对于每一条训练语料, 进行分词和停用词过滤.

step2: 将经过预处理之后的文本与基础情感词典中的词语匹配, 如果匹配成功, 则说明该词语已经收录在基础情感词典中, 不再考虑; 如果未匹配成功, 则该词语属于未收录词语, 将该词语加入未收录候选词表中.

step3: 待所有训练语料处理完毕之后, 对于未收录候选词表中的词, 进行人工判断词语的情感倾向, 并且根据倾向性赋予倾向值(褒义为“+1”, 贬义为“-1”), 将倾向值为正的词, 加入正向基础情感词典中, 将倾向值为负的词, 加入负向基础情感词典中.

本文没有使用 SO-PMI 或 PMI 等方法去计算情感词更为具体的倾向值, 原因有如下几点:

a. 对于整个分析过程来讲, 正向词倾向值定为+1, 负向词倾向值定为-1, 这样可以简化分析过程的计算量, 并且具体的倾向值对倾向分析的影响没有否定词、程度副词、关联词等的影响大. 如本文这样定义情感词的倾向值, 在倾向值计算的时候节省了开销, 配合辅助词典进行分析得到的分析结果已经能够满足要求.

b. 使用 SO-PMI 或 PMI 算法计算词语倾向值, 需要选择种子词, 而种子词的选择方法没有固定标准, 理论上是选择情感倾向强烈的词作为种子词, 但是计算词语倾向性强烈程度本身就是要解决的问题. 另外, 选择的种子词的数量对分析结果也有影响.

### 2.2 辅助词典集的构造

2.2.1 停用词典 停用词是在表达中为了语义连贯而添加的没有实际意义的词语, 或者对于情感倾向分析来说没有帮助的词. 例如, 在“这部手机太好了”和“这部手机用着真流畅”中, “了”和“着”这两个词在语言表达上都没有实际的意义, 仅仅只是习惯表达或者为了语意连贯而加上的. 本文构建的停用词典共收录停用词 1 118 个.

2.2.2 否定词典 在自然语言中, 否定表达根据否定词数量的不同, 分为否定表达和多重否定. 否定词数量的多少能够直接影响情感词的倾向值. 其中否定和三重否定表示的否定的意思, 而双重否定表达的是着重肯定的意思. 这里引入乘积因子  $\beta$ , 用来处理否定词数量对情感词倾向值的影响. 在匹配的过程中记录否定词的匹配的次数, 匹配过程结束后, 若否定词的数量为偶数, 将  $\beta$  置为 2, 若否定词的数量为奇数, 将  $\beta$  置为-1.

2.2.3 程度副词词典 程度副词既可以改变情感词的倾向性, 又可以改变情感词的倾向值, 从而在分句及整句的倾向值计算中影响倾向分析的结果, 例如“手机好用”是正面评价, “手机很好用”是对“好用”的加强. 程度副词词典收录了 115 个程度副词, 是将 HowNet 提供的程度级别词典进行人工筛选过滤之后, 添加进部分新词构成的. 本文定

义乘积因子  $\alpha$  来综合表征这三类程度副词的影响,在程度副词词典中记录程度副词的权值,在进行程度副词匹配时,若匹配成功,则将情感词的倾向值乘以  $\alpha$ ,并用于后续计算. 程度副词词典结构如表 1 所示.

表 1 部分程度副词及权值

Table 1 Weights of some degree adverbs

程度副词	权值
很	2
非常	2
稍微	0.5
一点	0.5
过于	-1

**2.2.4 关联词词典** 关联词是对倾向性分析影响较大的一个原因,尤其在中文表达中,连词的出现频率非常高,忽视这些连词对倾向性的影响可能会使倾向值的计算不准确,甚至导致倾向分析出现错误,例如“这个手机虽然贵,但是性价比高”,这条评论里重点要表达的是对手机的正面情感,如果不考虑关联词,“贵”的倾向为“-1”,“性价比高”倾向值为“+1”,综合倾向值为“0”,为中性,但是从汉语语法来看,这条评论表达了作者正面情感. 为了处理关联词对于倾向分析的影响,本文构建了关联词词典. 关联词词典中收录的是常用的汉语关联词,根据这些关联词对汉语表达不同分句之间语义的影响,赋予关联词所影响的分句不同的权值. 定义参数  $\theta$  表征重点分句的权值(非重点分句权值为1),部分关联词权重如表 2 所示.

表 2 部分关联词及权重

Table 2 Weights of some conjunctive words

关联词	第一分句权重	第二分句权重
尽管.....可是	1	2
虽然.....但是	1	2
.....却.....		2
不但.....而且	1	2
不仅.....还	1	2

**2.2.5 动态词典** 动态词的处理是近年中文情感倾向分析中的一个难点. 所谓动态词,就是情感倾向不固定,随着语境和评论主体的不同会发生变化的词,如“快”、“高”等情感词,单独从词语的倾向性来看是褒义的,收录在正面情感词典中,但是在“耗电快”、“发热高”等表达中,显然是负面的.

在不涉及领域的通用词典中,对于这类词语是无法识别的,因为通用词典是基于词语匹配来分析的,无法识别语义信息. 无法识别动态情感词将会

大大降低情感分析的准确度.

通过对大量语料研究,发现动态表达呈现一定的规律,即情感词之所以能够表现出动态的情感倾向,是因为这些动态词描述的是评论主体或评论主体某一属性的特点,而这些属性往往是领域相关的,且情感词往往具有与平时表达相反的情感倾向,如上文“耗电快”、“发热高”等表达中的“快”、“高”.

在不涉及特定领域的倾向分析中,这些评论主体数量很多,但是具体到特定的领域,这些评论主体的数量就非常有限了,而本文的研究目的正好是在保留通用性的情况下来满足领域适用性,所以经过大量的语料研究,提出了一种解决动态词倾向识别的方法,即构建一个动态词典,其中收录能够改变情感词倾向性的评论主体. 通过上面的分析,笔者认为解决动态词的核心不是怎样去处理动态词本身,而是怎样去识别动态词所描述的对象. 通过语料分析,动态词的数量是不多的,涉及到特定领域的就更少了,但是这部分少量的动态词却对某些评论的倾向分析的结果的准确性起到了决定性的作用.

由于目前对动态词的研究还处于初步阶段,本文给出了处理动态词的一种方法,但是动态词典的构建需要手动完成.

### 2.3 新词发现及词典维护方法

倾向分析中另外一个难点,就是新词识别的问题. 这里的新词指的是情感词典中未收录的词语. 例如“给力”、“坑爹”、“不明觉厉”等,这些词是不符合汉语语法规则的,但是借由网络快速传播,从而让大部分网民了解并广泛使用.

目前国内有部分这方面的研究,都是通过复杂的自动发现算法来识别新词,这样做的优点是完全不需要人工参与,但是就目前的研究现状来看,识别效果并不是很好,而且算法复杂,开销很大. 综合考虑识别效果、算法复杂度、计算开销等因素,本文提出一种有监督的新词发现方法.

横向比较每个网络用语出现的时间,发现如此多的网络用语并不是在短时间内集中出现的,而是在相对较长的时间里零散的出现. 考虑到这种事实,花费少量的人工来换取高效的新词发现效率是可行的,主要过程如下:

**a.** 当发现未匹配的分词片段的时候,不仅仅只记录该片段,同时记录该片段的上一片段和下一片段,这样当提交给人工判断时候方便人工进行正确处理;

**b.** 当未匹配片段达到阈值,交由人工判断,若

判定为未收录新词,将该词加入情感词典或辅助词典集中对应的词典;对于判断为不是新词的组合,添加进非情感词典中,这样可以避免重复判断不是新词的高频组合。

关于阈值  $T$  的设定可以根据分析规模的大小来设定。

这种新词发现方法可以用来收录情感词典中为收录的新词,也可以用来维护构建辅助词典集。

### 3 倾向计算与结果分析

#### 3.1 倾向值计算方法

通过分析可知,进行倾向分析的评论必须是同一主体,若评论中含有多个评论主体,则需进行句子级的挖掘。

设计倾向值计算方法的思路是先计算分句的倾向值,再计算整句的倾向值,最后计算文本的倾向值。算法描述如下:

输入:待分类文本  $d$ ,情感词典  $\text{sendic}$ ,停用词典  $\text{stopdic}$ ,否定词典  $\text{ndic}$ ,程度副词词典  $\text{levdic}$ ,关联词典  $\text{reldic}$ ,动态词典  $\text{dyndic}$ ;

输出:文本  $d$  的分析结果

step1:将待分类文本  $d$  使用中科院的分词系统 ICTCLAS<sup>[17]</sup>进行分词;

step2:将分词完成的文本使用停用词典  $\text{stopdic}$  进行过滤;

step3:将文本按照句号划分成整句集合  $S=\{S_1, S_2, \dots, S_n\}$ ,并对每一个整句根据逗号和分号划分成分句  $s_1, s_2, \dots, s_n$ ;

step4:对每一分句使用情感词典  $\text{sendic}$  匹配情感词,若匹配失败,则对下一分句进行匹配,直至所有分句匹配失败;若匹配成功,进行下一步;

step5:从情感词的位置开始反向遍历分句,根据程度副词词典  $\text{levdic}$  匹配程度副词,若匹配到程度副词,则置乘积因子  $\alpha=2$ ,否则置乘积因子为 1;

step6:从情感词的位置反向遍历分句,根据否定词典  $\text{ndic}$  匹配否定词,记录否定词的数量  $n$ ,若  $n$  为奇数,则置乘积因子  $\beta=-1$ ,若为非 0 偶数,则置为 2,若为 0 则置为 1;

step7:从情感词的位置反向遍历分句,根据动态词典  $\text{dyndic}$  匹配动态情感词的评论主体,若匹配成功,则将情感词的倾向性改变,倾向值的绝对值不变;

step8:从情感词的位置反向遍历分句,根据关联词典  $\text{reldic}$  匹配关联词,根据关联词典中定义的分句权值确定乘积因子  $\theta$  的值;

step9:根据公式(1)计算文档  $d$  的倾向值  $\text{Orientation}(d)$ ;

step10:若  $d$  的倾向值  $>0$ ,则文档为褒义;若  $d$  的倾向值  $<0$ ,则文档为贬义。

$$\text{Orientation}(d) = \frac{\sum_{i=1}^n \left( \frac{\sum_{j=1}^m \alpha \beta \theta \text{Ori}(w)}{m} \right)}{n} \quad (1)$$

式(1)中  $m$  为整句中分句的数量,  $n$  为文档中整句的数量。  $\text{Ori}(w)$  为分句中情感词结合动态词典  $\text{dyndic}$  之后的倾向值,  $\alpha$  表示程度副词对倾向性分析的影响;  $\beta$  表示否定词对倾向性分析的影响;  $\theta$  表示关联词对情感分析的影响。

#### 3.2 试验和结果分析

以手机领域的评论为例,进行情感词典的扩充和动态词典的构造。实验数据为数据堂下载的真实手机评论。从获取的手机评论中随机选择 2 000 条作为训练语料,用于情感词典的构建。从中选择 1 000 条作为测试语料用于实验结果分析。

通过对 2 000 条训练语料进行以上处理,新增了 71 个领域情感词。其中褒义 23 个,贬义 48 个。将褒义词添加进正向基础情感词典,构成正面情感词典,将贬义词添加进负向基础情感词典,构成负面情感词典。构成的正面情感词典和负面情感词典部分内容如表 3 所示。

表 3 部分正面情感词  
Table 3 Some positive emotional words

倾向	举例
褒义情感词	赞、持久、经摔、耐用、牛掰、给力
贬义情感词	坑爹、黑屏、死机、翻新机、垃圾、山寨机、失灵、雪花屏

然后人工在手机评论的语料中搜集动态情感词的评论主体和评论主体的属性,构成动态词典,部分动态词典如图 2 所示。

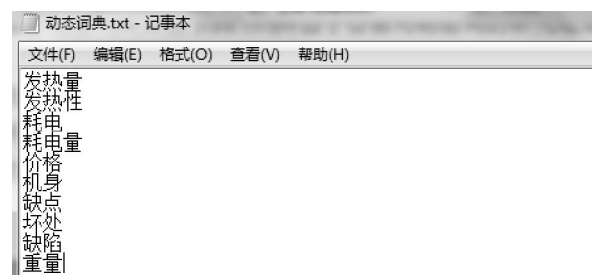


图 2 部分动态词

Fig.2 Some dynamic terms

邀请 3 位标注者分别对 1 000 条测试语料进行人工标注,然后对标注结果进行汇总,若有两名标注者或所有标注者对语料标注为正向,则语料倾向性为正,若有两名标注者或所有标注者对语料标注为负向,则语料倾向性为负。

将本文提出的方法和传统的使用HowNet 计算的方法进行比较,为了更具体的对比分析两种方法,对正面情感和负面情感分开计算,结果如表 4 所示。

表 4 实验结果

Table 4 Experimental results

方法	评估指标	正面情感	负面情感
本文提出的方法	文本数量	698	302
	准确率 $p$	0.713	0.738
	召回率 $r$	0.769	0.706
	$F$ 值	0.739	0.721
基于 HowNet 的方法	文本数量	788	212
	准确率 $p$	0.503	0.462
	召回率 $r$	0.611	0.310
	$F$ 值	0.672	0.371

通过对上表实验数据的分析可以得到以下结论:

在进行领域相关文本的倾向性分析时,本文提出的方法在正面情感的判断和负面情感的判断上,无论准确率和召回率都要比基于 HowNet 的传统情感词典方法要高。首先是因为对 HowNet 进行了筛选过滤,去除了不符合收录规则的词;其次是进行了领域相关的训练语料的学习工作,引入了领域情感词,并且扩充了网络流行用语,对领域相关的评论进行分析可以提高分析的准确度;第三是辅助词典集全面的考虑了语义和句子间的结构关系,能够处理否定词、程度副词、关联词,能够提高倾向值计算的准确性,对于动态词的识别能够增强分析的准确度。

## 4 结 语

以上提出的一种基于领域情感词典的网络评论倾向分析方法,给出了情感词典的构建和扩充方法,并通过构建辅助词典集以及有监督的维护方法来进行辅助分析。

目前的方法还存在以下不足:不能识别网络表达中表情符号和不符合语法规则的表达;需要研究出自动的新词发现方法;对于特殊句式如反问句等还不能识别。将在今后的工作中对这些问题做进一步的研究。

## 参考文献:

- [1] 黄萱菁,张奇,吴苑斌.文本情感倾向分析[J].中文信息学报,2011,10(9):118-126.  
HUANG Xuan jing,Zhang Qi, WU Yuan bin. A Survey on Sentiment Analysis [J]. Journal of Chinese information,2011,10(9) 118-126.(in Chinese)
- [2] Turney PD. Thumbs up or thumbs down semantic orientation applied to unsupervised classification of views [C]//Proceedings of the 40th annual meeting on association for computational linguistics. Philadelphia: Association for Computational Linguistics,2002:417-424.
- [3] Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales [C]//Proceedings of the 43rd annual meeting on association for computational linguistics. New York: NJ,ACM,2005:115-124.
- [4] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto. Collecting Evaluation Expressions for Opinion Extraction[C]//Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP). 2004:584 - 589.
- [5] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto. Extracting Aspect-Evaluating and Aspect-of Relations in Opinion mining [C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007:1065-1074.
- [6] Marrese-Taylor E, Velasquez J D, Bravo-Marquez F. A novel deterministic approach for aspect-based opinion mining in tourism products reviews[J]. Expert Systems with Applications,2014,41(17):7764-7775.
- [7] Penalver-Martinez I, Garcia-Sanchez F, Valencia-Garcia R, et al. Feature-based opinion mining through ontologies [J]. Expert System with Applications,2014, 41(13):5995-6008.
- [8] 朱嫣岚,闵锦,周雅倩,等.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 Yan-lan,MIN Jin,ZHOU Ya-qian,etc. Semantic orientation computing based on hownet[J]. Journal of Chinese Information,2006,20(1):14-20.(in Chinese)
- [9] 孙建旺,吕学强,张镭翰.基于词典与机器学习的中文微博情感分析研究[J].计算机应用与软件,2014,31(7):177-181.  
SUN Jianwang LV Xueqiang Zhang Leihan. On sentiment analysis of chinese microblogging based on lexicon and machine learning[J]. Journal of Computer Applications and Software,2014,31(7):177-181.(in Chinese)

- [10] 肖红, 许少华. 基于句法分析和情感词典的网络舆情倾向性分析研究[J]. 小型微型计算机系统, 2014, 35(4): 811–813.  
XIAO Hong, XU Shao-hua. Analysis on web public opinion orientation based on syntactic parsing and emotional dictionary [J]. Small Microcomputer system, 2014, 35(4): 811–813. (in Chinese)
- [11] 王勇, 吕学强, 姬连春, 等. 基于极性词典的中文微博客情感分类 [J]. 计算机应用与软件, 2014, 31(1): 35–37, 126.  
WANG Yong, LV Xueqiang, JI Lianchun, et al. Sentiment classification for chinese microblogging based on polarity lexicons [J]. Computer Applications and Software, 2014, 31(1): 35–37, 126. (in Chinese)
- [12] 徐晓丹, 段正杰, 陈中育. 基于扩展情感词典及特征加权的情感挖掘方法[J]. 山东大学学报(工学版), 2014, 44(6): 15–18, 69.  
XU Xiaodan, DUAN Zhengjie, CHEN Zhongyu. The sentiment mining method based on extended sentiment dictionary and integrated features [J]. Journal of Shandong University (engineering science), 2014, 44(6): 15–18, 69. (in Chinese)
- [13] 谢松县, 刘博, 王挺. 应用语义关系自动构建情感词典[J]. 国防科技大学学报, 2014, 36(3): 111–115.  
XIE Songxian, LIU Bo, WANG Ting. Applying semantic relations to construct sentiment lexicon automatically [J]. Journal of National University of Defense Technology, 2014, 36(3): 111–115. (in Chinese)
- [14] 周咏梅, 阳爱民, 林江豪. 中文微博情感词典构建方法[J]. 山东大学学报(工学版), 2014, 4(3): 36–40.  
ZHOU Yongmei, YANG Aimin, LIN Jianghao. A method of building Chinese microblog sentiment lexicon [J]. Journal of Shandong University (engineering science), 2014, 4(3): 36–40. (in Chinese)
- [15] 周强, 冯松岩. 构建知网关系的网状表示[J]. 中文信息学报, 2000, 14(6): 21–27.  
ZHOU Qiang, FENG Song-yan. Build a relation network representation for How-net. Journal of Chinese information, 2000, 14(6): 21–27. (in Chinese)
- [16] 郭凯. 基于评论情感的微博谣言检测研究 [D]. 大连: 大连理工大学, 2014.  
GUO Kai. The research of Microblog Rumors Detection Based on Comments Sentiment [D]. Dalian: dalian university of technology, 2014. (in Chinese)
- [17] 蔡小艳, 寇应展, 沈巍, 等. 汉语词法分析系统 ICTCLAS 在 Nutch-0.9 中的应用与实现 [J]. 军械工程学院学报, 2008, 20(5): 63–70.  
CAI Xiao yan, KOU Ying zhan, Shen Wei, et al. The Application and Realization of ICTCLAS on Nutch-0.9 [J]. Journal of ordnance engineering college, 2008, 20(5): 63–70. (in Chinese)

## Method of semantic orientation analysis based on domain specific sentiment dictionary

HE Cheng-wan, WANG Ge

School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China

**Abstract:** The volume of data on the network increases remarkably with the rapid development of mobile network. How to extract useful information from the complicated information becomes an urgent problem to be solved. An improved semantic orientation analysis was proposed, in which the sentiment dictionary was expanded with domain emotional words, and an auxiliary dictionary set was constructed for assistant analysis. To maintain all the dictionaries, a semi-automatic method was also presented. The results in the domain of mobile phone show that the rates of accuracy and recall of positive emotion analysis are 0.713 and 0.769, and the rates of accuracy and those of negative emotion analysis are 0.738 and 0.706, which are improved compared with using the traditional method.

**Keywords:** semantic orientation analysis; sentiment dictionary; assistant dictionary set

本文编辑: 陈小平