ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 137－147
DOI:10.16337/j.1004-9037.2015.01.013

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

# 基于词亲和度的微博词语语义倾向识别算法

唐浩浩　王　波　周　杰　陈　东　刘绍毓

(解放军信息工程大学信息系统工程学院,郑州,450001)

**摘　要:** 准确识别词语语义倾向并构建高质量的情感词典,从而提高微博文本情感分析的准确率,具有重要意义。传统的基于语料库方法对种子词选取敏感,并且不能有效对低频词语语义倾向进行识别。本文提出了一种基于词亲和度的微博词语语义倾向识别算法。利用词性组合模式提取候选词集,选取微博表情符号作为种子词,并构建词亲和度网络,利用同义词词林对低频词进行扩展,计算候选词与种子词之间语义倾向相似度。根据设定阈值判断词语语义倾向。在200万条微博语料上分别将本文算法与传统算法进行对比,实验结果表明本文算法优于传统算法。

**关键词:** 微博;情感词;情感分析;语义倾向;词亲和度

**中图分类号:** TP391　　**文献标志码:** A

## Semantic Orientation Identification for Terms From Chinese Micro-blogs Based on Word Affinity Measure

Tang Haohao, Wang Bo, Zhou Jie, Chen Dong, Liu Shaoyu

(Institute of Information and System Engineering, PLA Information Engineering University, Zhengzhou, 450001, China)

**Abstract:** How to identify the semantic orientation of terms and build a high-quality sentiment dictionary to improve the accuracy of sentiment analysis on Micro-blogs has significant importance. Traditional algorithms based on corpus are sensitive to the seed words, and cannot effectively identify semantic orientation identification on low-frequency terms. To solve this problem, an algorithm based on word affinity measure is proposed to identify the semantic orientation of terms from Chinese Micro-blogs. Firstly, candidate words are extracted by the part of speech combination patterns. Secondly, Micro-blog emoticons are selected as seed words, and word affinity networks are built. Then, low-frequency words are expanded by a synonyms dictionary during calculating the semantic orientation similarity between candidate words and seed words. Finally, the semantic orientation is determined according to the threshold. Experiments are conducted on a corpus with two million Micro-blogs using the proposed algorithm and traditional algorithms respectively. Experimental results show the advantage of the proposed algorithm.

**Key words:** Micro-blog; opinioned terms; sentiment analysis; semantic orientation; word affinity measure

---

**基金项目:** 国家高技术研究发展计划("八六三"计划)(2011AA7032030D)资助项目;全军军事研究生课题(2011JY002-158)资助项目。

**收稿日期:** 2013-11-25;**修订日期:** 2014-04-09

## 引 言

近年来,微博客(Micro-blog,简称微博)发展日益迅速。CNNIC 统计显示[1],截至 2013 年 6 月底,我国微博用户规模已达 3.31 亿,庞大的用户规模进一步巩固了其舆论传播的中心地位。微博传播速度快、影响大,其短文本及口语化的特点吸引了大批学者对其研究[2]。对于一些群体性事件,可以基于微博中反映这些事件变化的用户情绪,进行准确预测或及时预警[3]。要实现上述目标,需要捕捉事件并对其进行准确的情感分析,而词语语义倾向识别是情感分析的基础,对情感分析起着举足轻重的作用,因此准确有效地对词语语义倾向进行识别,进而构造高质量的情感词典具有重要意义。

现有的词语语义倾向识别方法主要有 3 种:人工标注的方法、基于语料库的方法和基于词典资源的方法[4]。人工标注的方法耗费大量人物力资源且可扩展性差,常用作自动构建词典方法的补充。当前,相关研究主要集中在基于语料库和基于词典资源两类方法。

基于语料库的方法主要是依据语言的约束关系对大规模语料中词语的分布特性进行统计,判别词语倾向性。早期的一些学者发现,由连词连接的两个形容词的极性往往存在一定的关联性,比如"and"连接的形容词(如 lovely and beautiful)极性相同,然而"but"连接的形容词(如 lovely but unnatural)极性相反[5]。为了摆脱词性的限制,Turney 和 Littman[6]提出了点互信息(Point mutual information)的方法判别某个词语是否是情感词语,这种方法适用于各种词性,但是较为依赖褒、贬种子词集。在词语语义倾向识别中为了充分考虑词语的上下文信息,Li 等人[7]利用词激活力模型构建亲和度网络,通过亲和度传播来识别词语倾向,有效提高了情感词识别的性能,但对于低频词语的识别准确率有待提高。Tan 等人[8]将词和文档之间知识进行有机融合,提出了一种基于随机游走模型的领域情感词典构建方法。此外,针对微博中不断涌现的新词,Feng 等人[9]以表情符号作为基准词,使用点互信息的方法计算词语倾向,并通过对微博文本进行情感分类验证其有效性。基于语料库的方法最大的优点在于简单易行,缺点则在于可利用的语料库有限,同时情感词语在大规模语料库中的分布并不容易归纳,识别结果受种子词影响较大。

基于词典资源的方法主要是利用现有知识库中各个词语的标注信息(如同义词、概念描述等)以及相互联系,识别词语的语义倾向。2004 年,Hu 和 Liu[10]同时考虑 WordNet 中同义和反义关系的词语集合对情感词典进行扩展。朱嫣岚等[11]利用 HowNet 提出基于语义相似度和语义相关场的词语语义倾向计算方法,通过计算目标词语与 HowNet 已标注褒贬词语的相似度,获取目标词语的倾向性。Hassan 等[12]利用 WordNet 计算词之间的相似度,建立词之间的语义图,利用图算法识别词语倾向性。此外,为了利用相对完备的英文资源,Su 等[13]利用机器翻译将语料从源语言翻译为另外一种语言(英文或中文),并选取英文正负面种子词,通过点互信息来构建中文情感词典。针对情感词的领域相关问题,Liu 等[14]使用 HowNet 情感词典,从在线产品评论中抽取<产品属性#情感词>候选对来完成领域相关情感词典的构建。基于词典资源的方法能够获得较高的识别准确率,并且具有较快的识别速度,主要缺点是依赖于语义知识库,可扩展性较差,对于语义知识库中不包含的词语,需要根据规则进行人工标注。针对传统的基于语料库的方法对种子词敏感以及不能有效对低频词语进行识别的问题,本文提出了一种基于词亲和度的微博词语语义倾向识别算法。

## 1 原理描述与算法流程

### 1.1 原理描述

通常,微博中的情感词语是由用户的情感激活的,而用户的情感直接由表情符号来表征,因此可以从微博中倾向性鲜明的表情符号出发来识别情感词及其语义倾向;此外,相似的词具有相似的上下文[15],词亲和度网络充分考虑词语的上下文信息,并且同一条微博表达的情感在上下文中具有一致性。

因此,可以通过词亲和度网络来计算词语的倾向强度。本文方法的基本原理是利用这两个特性建立表情符号与词语间的亲和度网络,再利用词语跟表情符号间亲和度向量的相似度计算词语的语义倾向强度,以此完成微博词语的语义倾向识别。

### 1.2 算法流程

本文算法流程如图 1 所示,主要分为数据预处理和词语语义倾向识别两个部分。

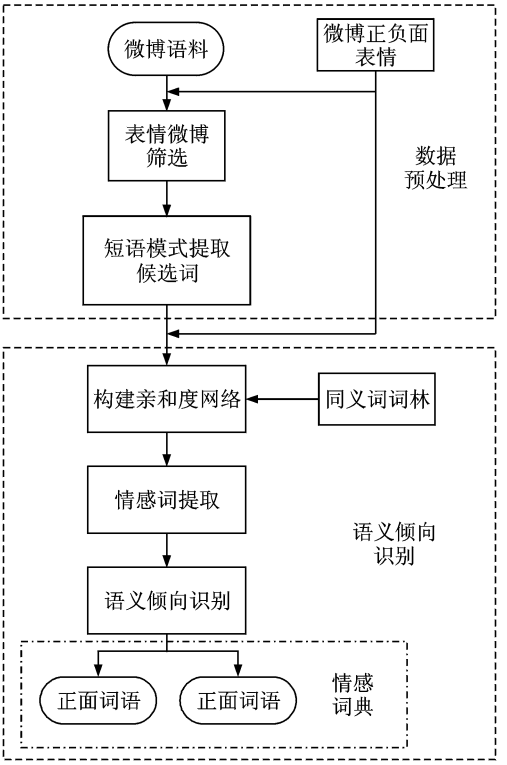


图 1 微博词语语义倾向识别流程

Fig. 1 Process of identifying semantic orientation of terms from chinese micro-blogs

数据预处理:首先,利用微博中倾向明确的正负面表情符号筛选表情微博(包含表情符号的微博);然后通过词性组合模式[16]从已筛选出的表情微博中提取候选词集。

语义倾向识别:首先,利用词激活力模型对数据预处理得到的候选词集构建词亲和度网络;然后,选择新浪微博表情集合中倾向明显且在候选词集中词频较高的 18 组微博表情符号作为种子词,并利用同义词词林对低频词进行扩展,提取情感词;最后,利用种子词以及构建的亲和度网络计算词语的语义倾向强度,完成词语语义倾向识别。

## 2 微博词语语义倾向识别

### 2.1 数据预处理

由于整个语料中词语集合庞大,且客观词语居多,若建立所有词语间的亲和度网络,不仅会影响识别结果,还会使计算量急剧增加。利用文献[16]中的词性组合模式提取候选词集。表 1 列出了部分词性组合模式及其实例。其中,〈·〉表示词性组合模式中语义倾向词语的候选项;BOS 和 EOS 为句首和

句尾标记;符号"/"表示可任选一项;[·]为可选项。

本节的分词和词性标注选用 NLPIR 汉语分词系统 2013 版。采用二元关系(word,freq)表示模式中的各个词语,得到候选项集合 $W=\{w_1,w_2,\cdots,w_N\}$,$N$ 为候选词总数。

**表 1 部分词性组合模式及其实例**

**Table 1 Examples of some speech combination patterns**

| 词性组合模式 | 模式表示 | 实例 |
|---|---|---|
| 〈形容词〉+连接词+〈形容词〉 | 〈a〉+cc+〈a〉 | 真诚/an 与/cc 豁达/an |
| 〈形容词〉+的+名词 | 〈a〉+udel+n | 优秀/a 的/udel 教练/n |
| 副词+〈形容词〉 | d+〈a〉 | 太/d 狭隘/a |
| BOS/分隔符+〈形容词〉+EOS/分隔符 | BOS/w+〈a〉+EOS/w | 霸道/a ! /wt |
| 〈不及物动词/名动词〉+〈不及物动词/名动词〉 | 〈vi/vn〉+〈vi/vn〉 | 投机/vn 暴富/vn |
| BOS+[副词]+〈动词〉+[部分标点]+EOS | BOS+[d]+〈v〉+[wj/wt/wf]+EOS | [不/d] 支持/v [! /wt] |

## 2.2 亲和度网络

词亲和度是词激活力模型(Word activation forces, WAF)[17]中的一个概念。该模型是一种表现文本中词与词之间关系的统计模型,可以将人脑中复杂的词关系网络映射成计算机可读的词激活力矩阵,能够充分利用词语的上下文信息,更深层地剖析其内在语义。模型涉及两个重要概念,分别是词激活力和词亲和度。

词激活力体现了两个词语在二者全局邻接网络中的有序共现程度,表示一个词对另一个词出现与否的激活权重。对于给定的词语 $i$ 和 $j$,词语 $i$ 对于词语 $j$ 的词激活力定义如下

$$waf_{ij}=(f_{ij}/f_i)\cdot(f_{ij}/f_j)/d_{ij}^2 \tag{1}$$

式中:$f_i$ 和 $f_j$ 分别表示两个词在文档中出现的频次;$f_{ij}$ 表示词语 $i$ 和 $j$ 在设定共现窗距离内出现的频次;$d_{ij}$ 为两个词的平均共现距离。根据定义,$waf_{ij}$ 的数值区间是[0,1],0 表示文档中词语 $j$ 从来没有在词语 $i$ 后的 $d_{ij}$ 个词语内出现,1 表示文档中词语 $j$ 总是毗邻出现在词语 $i$ 之后。

根据 $waf$ 定义,可以将一篇文档(或文档集)表示为词激活力矩阵,即

$$\boldsymbol{WAF}=\begin{bmatrix} waf_{11} & waf_{12} & \cdots & waf_{1N} \\ waf_{21} & waf_{22} & \cdots & waf_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ waf_{N1} & waf_{N2} & \cdots & waf_{NN} \end{bmatrix} \tag{2}$$

由于 $waf_{ij}$ 是有向值,所以词激活力矩阵是一个非对称矩阵,也可称作有向词网。矩阵中元素 $waf_{ij}$ 表示词语 $i$ 以 $waf_{ij}$ 的权重激活词语 $j$。第 $i$ 行表示词语 $i$ 对其他所有词的激活力,即词语 $i$ 的出链;第 $i$ 列表示其他所有词对词语 $i$ 的激活力,即词语 $i$ 的入链。基于该有向词网来计算词语间的词亲和度,其定义如下

$$A_{ij}^{waf}=\left(\frac{1}{|K_{ij}|}\sum_{k\in K_{ij}}\mathrm{OR}(waf_{ki},waf_{kj})\right)^{\frac{1}{+}}\cdot\left(\frac{1}{|L_{ij}|}\sum_{l\in L_{ij}}\mathrm{OR}(waf_{il},waf_{jl})\right)^{\frac{1}{+}} \tag{3}$$

式中:$K_{ij}=\{k\mid waf_{ki}>0 \text{ or } waf_{kj}>0\}$ 表示词语 $i$ 和词语 $j$ 的入链集合;$L_{ij}=\{l\mid waf_{il}>0 \text{ or } waf_{jl}>0\}$ 表示词语 $i$ 和词语 $j$ 的出链集合;$\mathrm{OR}(x,y)=\min(x,y)/\max(x,y)$ 为重叠率计算;词亲和度 $A_{ij}^{waf}$ 是词语 $i$ 和词语 $j$ 在词激活力矩阵中所有入链和出链重叠率的几何平均值,体现了两者在整个文档中的亲密程

度。

同样地,可以用词亲和度 $A_{ij}^{waf}$ 将文档表示为亲和度矩阵,即

$$\boldsymbol{A}^{waf}=\begin{bmatrix} A_{11}^{waf} & A_{12}^{waf} & \cdots & A_{1N}^{waf} \\ A_{21}^{waf} & A_{22}^{waf} & \cdots & A_{2N}^{waf} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1}^{waf} & A_{N2}^{waf} & \cdots & A_{NN}^{waf} \end{bmatrix}=\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix} \tag{4}$$

词亲和度矩阵是一个无向对称矩阵,其中第 $i$ 行表示其他所有词与词语 $i$ 的亲和度。在词语的语义倾向识别过程中,可以将词语作为节点,各节点间的亲和度作为边,构建亲和度网络,亲和度越强则节点语义倾向越相近。设节点集合为 $W=\{w_1,w_2,\cdots,w_N\}$,节点$\langle w_i,w_j\rangle$间的亲和度为 $A_{ij}^{waf}$。图 2 为词亲和度网络示意图。

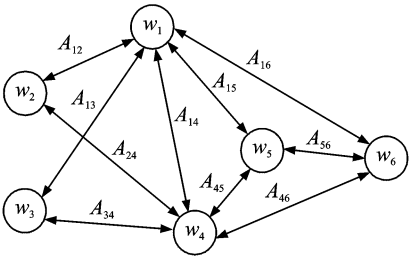


图 2 词亲和度网络示意图

Fig. 2 Diagrammatic sketch of word affinity measure network

## 2.3 情感词提取

情感词提取主要包括两个部分:种子词的选取和情感词提取算法。情感词提取结果的好坏很大程度上取决于种子词选取的质量。种子词的选取通常有两种方法:一种是基于语料统计词频,选出词频最高且倾向明显的一组词作为种子词;另一种是基于词典资源,选出词典中倾向最明显的一组词作为种子词。前者容易造成语义覆盖面不全从而影响准确率,后者很大程度上依赖于主观因素。本文选取微博中倾向明显且在语料中出现频次最高 18 组正负倾向的表情符号作为种子词,语义覆盖面相对全面,且语义倾向客观,避免了主观因素的影响,如表 2 所示。

经过预处理后的语料中低频词占多数,低频词与种子词共现次数较少,且一旦共现二者之间亲和度很大,采用目前情感识别研究中的最优算法亲和度传播(Affinity propagation, AP)[7]会把大量低频的非情感词误识别为情感词,严重影响情感词提取的准确率。为了解决该问题,本文引入同义词词林对 AP 算法进行改进,即在情感词提取时,利用同义词词林对低频词进行扩展,同时考虑低频词及其同义词集与种子词之间的亲和度来完成情感词提取。

**表 2 18 组正负面微博表情**

**Table 2 Eighteen groups of negative Micro-blog emoticons**

| 正面表情 | 负面表情 |
|---|---|
| [给力][威武][礼物] | [弱][怒骂][伤心] |
| [呵呵][嘻嘻][哈哈] | [晕][悲伤][抓狂] |
| [可爱][爱你][偷笑] | [怒][汗][失望] |
| [亲亲][太开心][抱抱] | [衰][委屈][吐] |
| [酷][鼓掌][耶] | [生病][泪][可怜] |
| [good][赞][蛋糕] | [囧][吃惊][鄙视] |

情感词提取算法见算法 1。其中，$S$ 为初始种子词集，$S \in W$；$W$ 为候选词集，总量为 $N$；$\boldsymbol{A}^{waf}$ 为 $N$ 维亲和度矩阵；CiLin 为同义词词林；$f$ 代表低频词阈值，即词频小于该阈值的词为低频词；$\gamma_a$ 为情感词判决阈值；$T$ 为最大迭代次数。经过 $T$ 次迭代，输出情感词集 OPW。

**算法 1** 情感词提取算法

输入：初始种子词集 $S$，候选词集 $W$，亲和度矩阵 $\boldsymbol{A}^{waf} \in \mathbf{R}^{N*N}$，同义词林 CiLin，低频词阈值 $f$，情感词判决阈值 $\gamma_a$，最大迭代次数 $T$。

输出：情感词集 OPW。

初始化：情感亲和度 $\mathrm{OPI}_j = 0.0$，for all $w_j \in W$

01：OPW = [], iternum = 0

02：iternum += 1

03：for $w_j \in W$

04：　if $w_j \notin S$ and $w_j \notin$ OPW then

05：　for $s_t \in S$

06：　if $tf_{w_j} > f$

07：　$\mathrm{sum}A_j = \mathrm{sum}A_j + A_{ij}^{waf}$

08：　else

09：　$\mathrm{sum}A_j = \mathrm{sum}A_j + \frac{1}{m}\sum_{t}^{m} A_{tj}^{waf}$（$w_t$ 为低频词 $w_j$ 的同义词集）

10：　end for

11：　$\mathrm{OPI}_j = \mathrm{sum}\mathrm{A}_j / \mathrm{S} \cdot \mathrm{size}$

12：　if $\mathrm{OPI}_j > \gamma_a$

13：　将词语 $w_j$ 添加到 OPW

14：end for

15：if OPW 非空且 iternum < $T$

16：更新情感词集 OPW 到种子词集 $S$

17：重复以上步骤 2～10 次，输出最终情感词集 OPW

## 2.4 语义倾向识别

通过计算词语与正负面种子词间亲和度向量的相似度来衡量词语的正负面倾向强度，进而得出词语的语义倾向强度。

设情感词集 OPW 总量为 $N'$，那么词语 $c_j$（$c_j \in \mathrm{OPW}$，$j \in [1, 2, \cdots, N']$）语义倾向强度可以表示为

$$SO_j = SO_j^+ - \beta * SO_j^- \tag{5}$$

其中，$SO_j^+$ 和 $SO_j^-$ 分别表示词语 $c_j$ 与正负面种子词集的语义倾向相似度，$\beta = \sum_j SO_j^+ / \sum_j SO_j^-$ 为语料中正负面倾向强度比。

$SO_j^+$ 可由下式计算

$$SO_j^+ = \frac{1}{|P|}\sum_{p_i \in P} \mathrm{Sim}(c_j, p_i) = \frac{1}{|P|}\sum_{p_i \in P} \cos(\boldsymbol{v}_{c_j}, \boldsymbol{v}_{t \leftarrow p_i}) \tag{6}$$

式中：$\boldsymbol{v}_{c_j}$ 为词语 $c_j$ 的亲和力向量，$\boldsymbol{v}_{t \leftarrow p_i}$ 正面种子词 $p_i$ 在亲和度矩阵 $\boldsymbol{A}^{waf}$ 中对应的行向量，$t$ 为 $p_i$ 在 $\boldsymbol{A}^{waf}$ 中对应的行。对于 $\boldsymbol{v}_{c_j}$，有

$$\boldsymbol{v}_{c_j} = \begin{cases} \boldsymbol{v}_{t \leftarrow c_j} & tf_{c_j} \geqslant f \\ \dfrac{1}{|CL(c_j)|}\displaystyle\sum_{w \in CL(c_j)} \boldsymbol{v}_{t \leftarrow w} & tf_{c_j} < f \end{cases} \tag{7}$$

式中:$CL(c_j)$为词语 $c_j$ 在语料中的同义词集,$\boldsymbol{v}_{t\leftarrow c_j}$ 表示词语 $w(w\in CL(c_j))$在亲和度矩阵 $\boldsymbol{A}^{waf}$ 中对应的行向量。同理,可对 $SO_j^-$ 进行类似计算。

由 $SO_j^+$ 和 $SO_j^-$ 代入式(5),得到 $SO_j$,有

$$\begin{cases} SO_j > \gamma_p & \text{判为正面词语} \\ SO_j < \gamma_n & \text{判为负面词语} \\ \text{其他} & \text{判为中性词语} \end{cases} \tag{8}$$

## 3 实验结果与性能分析

### 3.1 实验数据分析

本文选取文献[9]采集的从 2011 年 10 月到 2011 年 12 月的 3 亿新浪微博语料作为实验语料。考虑到情感词的扩展可能受转折以及否定的影响,为了保证情感词识别的准确率,本文首先从 3 亿微博语料中删除同时包含正负面倾向的微博以及包含否定、转折的微博,然后利用选定的正负面表情符号筛选出 2 203 600 条表情微博,其中包含正面表情的微博 1 456 095 条,包含负面表情的微博 747 505 条。

微博文本中,虽然包含大量情感词语,但中性词语仍占绝大多数。采用词性组合模式选取候选词集,并依据候选词在语料中出现的频数进行排序,得到词语频数分布图如图 3 所示。

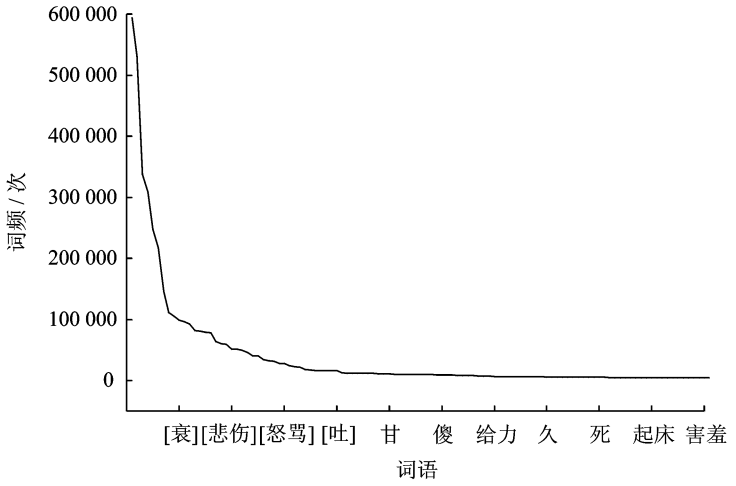


图 3 词语频数分布图

Fig. 3 Distribution of word frequency

经过组合模式筛选出的候选词总量为 33 440。由图 3 可以看到,语料中表情符号的频数远远高于其他候选词语,这也从一定程度上说明了本文选取表情符号作种子词的可行性;此外,低频词语在语料中的比例不容忽视,对于该部分词语的处理为本文改进文献[7]的方法提供了契机。

### 3.2 实验结果比较

本文分别选取文献[6]和文献[7]两个算法进行比较。文献[6]中 Turney 提出 SO-PMI 算法识别词语的语义倾向,该算法中种子词的选择对实验结果影响较大,一般要求种子词为具有较强正负面倾向的高频词语,且正负面词语频数相近。首先使用本文选取的表情符号作为种子词进行实验,结果记为 SO-PMI-E;然后使用语料中频数最高且倾向明显的词语为种子词实验,结果记为 SO-PMI-S。文献[7]中 Li 提出的 AP 算法对种子词选择相对不敏感,但对低频词识别效果欠佳。首先使用本文选取的表情符号

作为种子词实验，结果记为 AP-E；然后选取语料中词频最高且倾向明显的词语作为种子词实验，结果记为 AP-S。

针对文献[7]不能很好地对低频词进行识别的问题，本文利用同义词林进行扩展来解决低频词带来的稀疏问题；此外，结合微博文本的特点，本文使用表情符号作为种子词来进行词语语义倾向识别。为了验证其有效性，设置两组实验，首先使用表情符号作为种子词进行实验，结果记为本文方法 1；然后选取语料中词频最高且倾向明显的词语作为种子词进行实验，结果记为本文方法 2。

由于本文提出的方法是一种基于语料的方法，不需要对原始语料进行标注，因此只采用识别准确率作为评价指标。在准确率的计算中，本文合并 HowNet 情感词典（共计情感词语 2 090 个）与台湾大学 NTUSD 情感词典（共计 11 084 个），并去除其中标注不一致词语，然后对识别结果进行标注并计算识别准确率，准确率计算公式如式(9)所示。本文阈值参数均取经验最优值，其中低频词阈值 $f=20$，情感词判决阈值 $\gamma_a=0.002$，正面情感词判决阈值 $\gamma_p=0.009$，负面情感词判决阈值 $\gamma_n=-1E-4$，实验结果如图 4 所示。

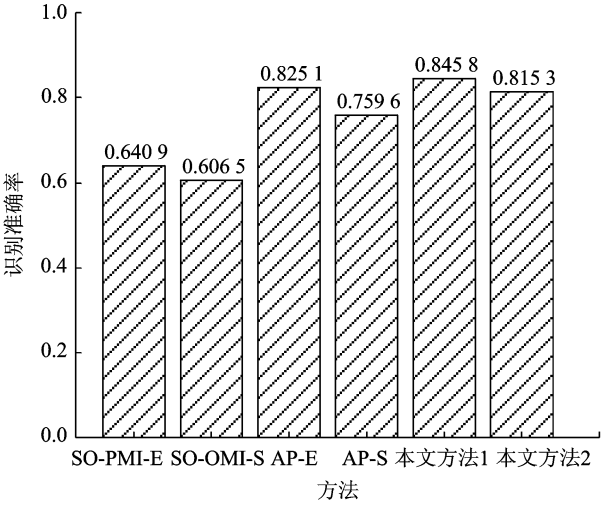


图 4 本文方法与其他方法对比结果

Fig. 4 The results compared with other methods

$$P=\frac{\text{识别正确的个数}}{\text{识别出的总个数}}\times 100\% \tag{9}$$

由实验结果可以看出，选取表情符号作为种子词其识别准确率普遍高于非表情符号作为种子词的方法。由于缺乏必要的语义约束，SO-PMI 算法整体性能较差。当选用合适的基准词时，使用 AP 算法能够取得较好的性能，但对低频词语的识别效果较差。本文算法采用同义词词林对低频词语进行扩展，所以其识别准确率比 AP 算法有明显提高。此外，限制本文算法识别准确率进一步提高的主要原因是，语料中还存在相当一部分词频较低且未在同义词林中出现的网络新词以及变形词等，该问题可以通过进一步加大语料规模以提高其词频来解决。

实验结果中，词语语义倾向强度 $SO_j$ 偏小，通过式(10)进行线性调整，并限定 $SO_j$ 的取值范围为 $[-1,1]$。

$$SO_j=\begin{cases}SO_j+0.5 & c_j \text{ 为正面词语}\\ SO_j-0.5 & c_j \text{ 为负面词语}\end{cases} \tag{10}$$

调整后,摘取部分情感词典如表 3 所示。

**表 3 部分情感词典倾向强度**

**Table 3 Sentiment strength of some words**

| 正面词语 | 调整前 | 调整后 | 负面词语 | 调整前 | 调整后 |
|---|---|---|---|---|---|
| 幸福 | 0.43 | 0.93 | 难受 | −0.52 | −1.0 |
| 快乐 | 0.42 | 0.92 | 郁闷 | −0.27 | −0.77 |
| 开心 | 0.31 | 0.81 | 伤不起 | −0.21 | −0.71 |
| 给力 | 0.20 | 0.60 | 坑爹 | −0.05 | −0.55 |
| 有爱 | 0.05 | 0.55 | 尼玛 | −0.04 | −0.54 |

### 3.3 词典质量评估

这里使用文献[9]的算法对本文扩展的词典质量进行评估,具体算法见算法 2。

**算法 2** 情感词典评估算法

输入:标注微博语料 $WB$,情感词典 $SD$,否定词典 $NG$

输出:每条微博 $WB_i$ 的情感强度 $\text{sentiScore}_i$,$i\in[1,2000]$

初始化:$\text{sentiScore}_i=0.0$

01:split $WB_i$ into sentences

02:for each sentence $s$ in $WB_i$ do

03: segment $s$ into words

04: for each word $w$ of $s$

05: if $w\in NG$

06: $ng++$

07: if $w\in SD$

08: $\text{sentiScore}_i=\text{sentiScore}_i+SO(w)$

09:if $ng$ is odd

10: $\text{sentiScore}_i=-\text{sentiScore}_i/2$

11:return $\sum\text{sentiScore}_i$

实验从 2.1 节提到的 3 亿微博中随机选取 2 000 条微博进行人工标注,然后分别使用 HowNet 情感词典、台湾大学 NTUSD 情感词典以及本文扩展的情感词典对这 2 000 条标注的微博进行情感分析。评价指标采用准确率、召回率和 $F1$ 值。由于 HowNet 情感词典和 NTUSD 情感词典无倾向强度标注,这里分别对其正负面情感词赋予倾向强度值 1 和−1。不同情感词典对 2 000 条微博进行情感分析的结果如表 4 所示。其中,PP 表示正面微博的准确率,RP 表示正面微博的召回率,FP 表示正面微博的 $F1$ 值,PN 表示负面微博的准确率,RN 表示负面微博的召回率,FN 表示负面微博的 $F1$ 值。

**表 4 不同情感词典实验结果对比**

**Table 4 The experimental results compared between different sentiment dictionaries**

| 情感词典 | PP | RP | FP | PN | RN | FN |
|---|---|---|---|---|---|---|
| HowNet | 0.557 | 0.659 | 0.604 | 0.298 | 0.533 | 0.383 |
| NTUSD | 0.501 | 0.749 | 0.60 | 0.644 | 0.611 | 0.627 |
| 本文词典 | 0.541 | 0.695 | 0.626 | 0.627 | 0.509 | 0.562 |

从表 4 可以看出,本文扩展的词典在对微博文本进行倾向性分析时,效果优于人工标注的 HowNet

词典,其原因是微博中的表情符号、网络新词等未在 HowNet 词典中收录;相比 NTUSD 词典,本文负面微博 F1 值较低,其原因是 NTSUD 词典中含有褒义词 2 812 个,贬义词 8 276 个,贬义词数量远大于褒义词数量,因此对负面微博进行识别时有着相对较高的 $F1$ 值。

## 4 结束语

本文结合微博文本特点,针对传统词语语义倾向性识别方法的不足,提出一种基于词亲和度的微博词语语义倾向识别算法。通过选取微博表情符号作为种子词,并构建词亲和度网络;然后,利用同义词词林对低频词进行扩展,计算词语的语义倾向强度,完成词语语义倾向识别。实验结果表明,本文算法能有效识别微博词语语义倾向,性能优于 AP 算法。需要指出的是,本文利用同义词词林进行扩展以降低数据稀疏的影响,然而微博文本用语随意,网络新词层出不断,如何提高未出现在同义词词林的新词以及网络新词的识别准确率,是本文下一步的研究工作;本文构建的词典属于微博通用情感词典,针对领域情感词,如何自动为其标注不同领域,也值得进一步研究。此外,算法的处理速度还有待提升。

**参考文献:**


[1] CNNIC. 中国互联网络发展状况统计报告[R]. 北京:中国互联网信息中心,2013.
CNNIC. Report of China Internet development statistics[R]. Beijing: China Internet Network Information Center, 2013.

[2] 张波,向阳,黄震华,等. 一种社交网络中的个体间推荐信任度计算方法[J]. 南京航空航天大学学报,2013,45(4):563-569.
Zhang Bo, Xiang Yang, Huang Zhenhua, et al. Recommended trust computation method between individuals in social network site[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2013,45(4):563-569.

[3] 周耀明,李弼程. 一种自适应网络舆情演化建模方法[J]. 数据采集与处理,2013,28(1):69-76.
Zhou Yaoming, Li Bicheng. Adaptive evolution modeling method of Internet public opinions[J]. Journal of Data Acquisition and Processing, 2013,28(1):69-76.

[4] Liu Bing, Zhang Lei. A survey of opinion mining and sentiment analysis[M]. New York: Springer US, 2012:415-463.

[5] 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 Eighth Conference of the European Chapter of the Association for Computational Linguistics. Morristown: ACL, 1997:174-181.

[6] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS), 2003,21(4):315-346.

[7] Li Yan, Li Si, Xu Weiran, et al. Analyzing semantic orientation of terms using affinity propagation[C]//8th International Symposium on Chinese Spoken Language Processing (ISCSLP). Hong Kong: The Chinese University of Hong Kong, 2012: 30-34.

[8] Tan Songbo, Wu Qiong. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon[J]. Expert Systems with Applications, 2011,38(10):12094-12100.

[9] Feng Shi, Wang Lin, Xu Weili, et al. Unsupervised learning Chinese sentiment lexicon from massive microblog data[M]. Berlin: Springer Berlin Heidelberg, 2012:27-38.

[10] Hu Minqing, Liu Bing. Mining and summarizing customer reviews[C]//Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle: ACM, 2004:168-177.

[11] 朱嫣岚,闵锦,周雅倩,等. 基于 HowNet 的词语语义倾向计算[J]. 中文信息学报,2006,20(1):14-20.
Zhu Yanlan, Min Jin, Zhou Yaqian, et al. Semantic orientation computing based on How Net[J]. Journal of Chinese Information Processing, 2006,20(1):14-20.

[12] Hassan A, Radev D. Identifying text polarity using random walks[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala: ACL, 2010:395-403.

[13] Su Yan, Li Shoushan. Constructing Chinese sentiment lexicon using bilingual information[M]. Berlin: Springer Berlin Heidelberg, 2013:322-331.

[14] Liu Lizhen, Lei Mengyun, Wang Hanshi. Combining domain-specific sentiment lexicon with HowNet for Chinese sentiment analysis[J]. Journal of Computers, 2013,8(4):878-883.

[15] Bollegala D, Weir D J, Carroll J. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland: ACL, 2011:132-141.

[16] 周杰.网络舆情话题情感倾向性分析技术研究[D].郑州:解放军信息工程大学,2010.
Zhou Jie. Research on sentiment orientation analysis of online public opinion topic[D]. Zhengzhou: PLA Information Engineering University, 2010.

[17] Guo Jun, Guo Hanliang, Wang Zhanyi. An activation force-based affinity measure for analyzing complex networks[J]. Scientific Reports, 2011,1:1-9.

作者简介:**唐浩浩**(1990-),男,硕士研究生,研究方向:文本倾向性分析,E-mail:tanghao_nlp@126.com;**王波**(1970-),男,副教授,研究方向:网络协议分析、智能信息处理;**周杰**(1984-),男,博士生,研究方向:命名实体识别与消歧;**陈东**(1989-),男,硕士研究生,研究方向:复杂网络分析与图检索;**刘绍毓**(1987-),男,硕士研究生,研究方向:中文实体关系抽取。