doi:10.3969/j.issn.1673-9833.2015.05.016

基于平滑SO-PMI算法的微博情感词典构建方法研究

杜 锐，朱艳辉，田海龙，刘 璟，马 进

（湖南工业大学 计算机与通信学院，株洲 湖南 412007）

摘 要：对现有情感词典在微博情感分类中的适用性进行了分析，针对现有情感词典在微博中情感词覆盖度低的问题，整合现有情感词典资源，构建了一个微博基础情感词典，同时提出了一种基于拉普拉斯平滑的SO-PMI算法对微博基础情感词典中没有收录的情感词倾向性进行判断，最后利用微博情感词典与拉普拉斯平滑的SO-PMI算法对微博情感词典进行了构建，并对所构建微博情感词典的分类性能进行了实验。实验结果表明，该方法所构建的情感词典在微博情感分类中能达到较好的分类效果。

关键词：中文微博；情感词典；情感分类；平滑

中图分类号：TP391.1 **文献标志码**：A **文章编号**：1673-9833(2015)05-0077-05

Research on Construction of Microblog Sentiment Lexicon Based on the Smooth SO-PMI Algorithm

Du Rui，Zhu Yanhui，Tian Hailong，Liu Jing，Ma Jin

（School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：Analyzed the applicability of the existing sentiment lexicon in the microblog sentiment classification. In view of low coverage of the existing sentiment lexicon, built a basic microblog sentiment lexicon by integrating the existing sentiment lexicon, and put forward a Laplacian-based smooth SO-PMI algorithm to judge emotional orientation of the words which not included in the basic sentiment lexicon, finally applied the microblog sentiment lexicon and the Laplacian smooth SO-PMI algorithm to construct the microblog sentiment lexicon, and tested the constructed lexicon classification capabilities. Experimental results showed that the constructed microblog sentiment lexicon achieved good effect inmicroblog sentiment classification.

Keywords：chinese microblog；sentiment lexicon；sentiment classification；smoothing

0 引言

随着移动互联网的快速发展，以微博为代表的社交媒体得到了广泛的应用。在微博中，人们可以享受快捷的获取信息的方式，也可以分享自己身边有趣的人或事。在海量的微博文本中包含着大量表达人们情感的主观性文本，这些主观性的微博在文本长度、表达方式、语言风格等方面与传统评论存在着较大的区别，分析主观文本的情感倾向是舆情

收稿日期：2015-07-13

基金项目：国家自然科学基金资助项目（61170102），国家社会科学基金资助项目（12BYY045），湖南省教育厅重点项目基金资助项目（15A049），湖南工业大学研究生创新基金资助项目（CX1313）

作者简介：杜 锐（1987-），男，湖北仙桃人，湖南工业大学硕士生，主要研究方向为文本处理，E-mail：578781015@qq.com

通信作者：朱艳辉（1968-），女，湖南株洲人，湖南工业大学教授，主要从事文本分类和信息检索方面的教学和研究，E-mail：swayhzhu@163.com

监控的重要基础。

在微博情感分析中，微博情感词典的构建具有重要的研究意义和使用价值，其不仅能为情感分析的研究提供参考，而且在情感特征选择及特征降维等方面有着重要的应用。在情感词典的构建过程中，情感词的倾向性计算是重点也是难点。目前，计算情感词倾向性的方法主要有基于语义相似度的计算与基于统计的计算方法。

文献[1]采用HowNet和NTUSD 2种资源对现有情感词典进行扩展，建立了一个具有倾向性程度的情感词典。文献[2]提出了一种自动构建与上下文相关的情感词典的最优化方法。文献[3]利用知网进行同义词扩展，提出一种HowNet和PMI（pointwise mutual information）相融合的词语极性计算方法。

基于HowNet的语义相似度计算方法[4]以及基于SO-PMI（semantic orientation-pointwise mutual information）的情感词倾向性计算方法[5]，这2种方法的共同点是：需要选取一定数量的正面种子词和负面种子词。不同点是：前者通过计算未知词与正、负面种子词相似度的方法判断未知词的倾向性，其中词语的相似度采用计算2个词语义原相似度的最大值而得到[6]；而后者通过互信息计算未知词与正、负面种子词关联度的方法对未知词的倾向性进行判断。上述2种方法在情感词的倾向性判断中取得了一定的效果，但是在中文微博中，由于网络新词较多，部分词如“给力”“正能量”“坑爹”等在HowNet中找不到义原，进而也就无法根据2个词义原的相似度计算词语的相似度。因此，基于HowNet的语义相似度计算方法对微博中部分词的倾向性判断并不适用。

基于SO-PMI的方法需要计算候选情感词与种子词的互信息，种子词通常选词频较高的情感词。而在微博中，若选取词频较高的情感词作为种子词则会带来如下问题：由于同一情感词可能在一条微博中出现多次，而在其他微博中出现的次数较少或并不出现，若将该情感词选为种子词则候选情感词与该种子词在整个语料中同现的次数可能为0，候选情感词与种子词的互信息无法计算，进而无法判断候选情感词的情感倾向性。因此基于SO-PMI的方法在判断微博中情感词的倾向性时也存在局限性，本文在已有情感词典资源的基础上，提出了一种基于改进SO-PMI的微博情感词典构造方法。

1 现有情感词典适用性分析

情感词典是指由一系列情感词及其相应的倾向性值构成的词集合。在含有情感词的微博情感句中，情感词是进行倾向性判断的重要特征。虽然，已有一些研究机构发布了一系列情感词典如《知网》情感词词典[7]、《大连理工大学情感本体》[8]等，这些情感词典为情感分类的研究提供了重要的参考。但由于微博中网络新词和网络用语层出不穷，现有情感词典对微博中所有情感词的覆盖程度难以确定。为此，本文对现有情感词典在微博中情感词的覆盖程度进行了分析。

课题组首先从COAE2013（2013年中文倾向性评测）任务三发布的微博评测标注语料中随机选取正、负面微博各100条，采用ICTCLAS[9]分词后，人工挑选微博中所有的正、负情感词及情感短语，其中正面情感词及短语119个、负面情感词及短语99个。然后，对现有情感词典资源《知网》《大连理工大学情感本体》分别进行了整理，情感词典整理结果如表1所示。

表1 微博基础情感词典

Table 1 The basic sentiment lexicon of microblog

项 目	HowNet	Dalian	Co
正面情感词/个	4 528	11 229	13 476
负面情感词/个	4 320	10 862	13 279

表1中，Dalian为大连理工大学情感本体，Co为合并HowNet与Dalian情感词典并去掉重复的词构成的情感词典，即微博基础情感词典。将上述整理后的情感词典分别与人工挑选的微博正、负情感词进行对比，并计算整理后的情感词典对微博中情感词的覆盖程度，覆盖度计算为：

$$q=\frac{m}{n}, \tag{1}$$

式中：m为整理后与基础情感词典完全匹配的情感词个数；n为正、负微博情感词个数。

利用上述公式分别计算HowNet正、负情感词典及Dalian正、负情感词典对微博正、负面情感词的覆盖度，其覆盖度结果如图1所示。

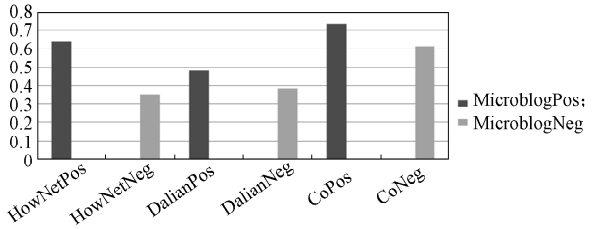

图1 正、负情感词覆盖度

Fig. 1 The coverage of positive and negative sentiment words

由图1可以看出，HowNet正面情感词典对微博正面情感词的覆盖程度较好，Dalian负面情感词

对微博负面情感词典的覆盖程度较好，微博基础情感词典对微博的正、负面情感词覆盖程度有显著提升，因此，整合现有情感词典在一定程度上能提高微博中情感词的覆盖度。但是，整合后的情感词典离完全覆盖微博中的正、负情感词还有一定的差距。

当微博条数增加时，需要判断整合后的情感词典是否具有稳定性，即随着微博条数的增加，情感词典对微博中情感词的覆盖度是否保持不变。为了对情感词典的稳定性进行分析，本文共进行了8组实验，每组分别选取50, 100, 150, 200, 250, 300, 350, 400条微博，人工挑选出每组中的所有微博情感词，并利用微博基础情感词典计算其对微博中情感词的覆盖度，其计算结果如图2所示。

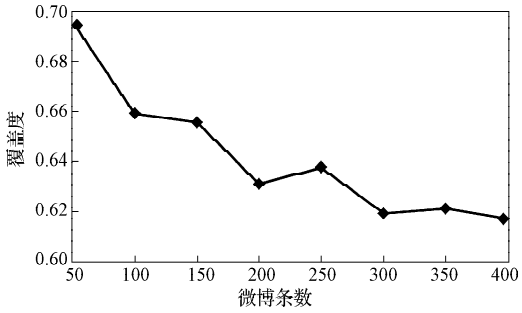

图2　情感词覆盖度随微博条数的变化趋势

Fig. 2　The sentiment words coverage changed with the numbers of microblog

由图2可知，当微博条数增加时，微博基础情感词典对微博中的情感词覆盖度降低。分析其原因，随着微博条数增加时，微博中的情感词也随之增加，而部分情感词如网络新词、情感短语等并没有在整合后的微博基础情感词典中收录。因此，整合后的基础情感词典在微博中并不具有稳定性。

2　基于改进的SO-PMI算法的微博情感词典构建

2.1　候选微博情感词的提取

候选微博情感词是指微博中可能是情感词的词或短语，其主要以名词、动词、形容词、副词存在。因此，微博中候选情感词的提取可以通过分词后词语的词性而得到，但仅仅以词性作为候选情感词的提取方式则会产生过多的候选情感词，为了减少候选情感词的粗糙程度，本文采用如下方式提取微博中的候选情感词。

首先，采用ICTCLAS对微博进行分词，提取词性为/a, /v, /n, /vn, /ag, /vi的词作为待入选候选微博情感词；然后，待入选候选微博情感词分别与微博基础情感词典中的正、负情感词匹配，正面匹配相同的词存入sp，记$sp=\{a_1, a_2, \cdots, a_n\}$，负面匹配相同的词存入$sn$，记$sn=\{b_1, b_2, \cdots, b_m\}$；则未匹配的词即为候选微博情感词，将该类词存入$sd$，记为$sd=\{c_1, c_2, \cdots, c_p\}$。

2.2　判断候选微博情感词倾向性

在利用SO-PMI算法对候选微博情感词的倾向性进行判断时需选取情感种子词，通常种子词采用词频统计的方式选取词频较高的情感词作为种子词。但由于微博文本长度较短，词频较高的种子词其文档频率并不一定高，若情感种子词在较少的微博中出现，则微博候选情感词与情感种子词在训练语料中同现的次数较少或不同现而无法计算其互信息。因此，为了避免上述问题的出现，认为，种子词的选取应选取文档频次较高的情感词而并非词频较高的情感词。若候选情感词与情感种子词在整个语料中同现的次数为0，则候选情感词与情感种子词的互信息无法计算，进而无法判断候选情感词的情感倾向性。为了避免解决上述问题，本文对SO-PMI算法进行了如下改进。

设有n个正面情感种子词：$P=\{p_1, p_2, \cdots, p_n\}$，$m$个负面情感种子词：$N=\{r_1, r_2, \cdots, r_m\}$，则对候选微博情感词中的每个词$c_i$（$i$=1, 2, …, p），其与正面情感种子词p_j（j=1, 2, …, n）的互信息为

$$PMI(c_i, p_j) = \log_2 \frac{p(c_i, p_j)}{p(c_i)p(p_j)}, \quad (2)$$

式中：$p(c_i, p_j)$为词c_i与正面情感种子词p_j在训练语料中同现的概率；

$p(c_i), p(p_j)$为词c_i, p_j在训练语料中出现的概率。

在实际计算过程中，上述概率值可用频率进行估计，因此有以下公式，即

$$p(c_i, p_j) = \frac{count(c_i, p_j)}{q}, \quad (3)$$

$$p(c_i) = \frac{count(c_i)}{q}, \quad (4)$$

$$p(p_j) = \frac{count(p_j)}{q}。 \quad (5)$$

式（3）~（5）中：$count(c_i, p_j)$为表示词c_i与p_j在训练语料中同现的微博条数；

$count(c_i)$为包含词c_i的微博条数；

$count(p_j)$为包含词p_j的微博条数；

q为训练集中总的微博条数。

将式（3）~（5）带入式（2）后得到式（6）：

$$PMI(c_i, p_j) = \log_2 \frac{q \times count(c_i, p_j)}{count(c_i) \times count(p_j)}。 \quad (6)$$

由于在实际计算过程中$count(c_i, p_j)$的值可能为

0，此时计算PMI值将无意义，因此，本文对式（3）引入拉普拉斯平滑技术：

$$p(c_i,p_j)=\frac{count(c_i,p_j)+1}{q+2}。\quad(7)$$

则式（6）可改进为：

$$PMI(c_i,p_j)=\log_2\frac{q^2\times[count(c_i,p_j)+1]}{(q+2)\times count(c_i)\times count(p_j)}。\quad(8)$$

同理，词c_i(i=1, 2, …, p)与负面情感种子词r_j(j=1, 2, …, m)的互信息可进行相应的改进，则词c_i的SO-PMI值可用如下公式计算：

$$SO\_PMI(c_i)=\sum_{j=1}^{m}\log_2\frac{q^2\times[count(c_i,p_j)+1]}{(q+2)\times count(c_i)\times count(p_j)}-\sum_{j=1}^{m}\log_2\frac{q^2\times[count(c_i,r_j)+1]}{(q+2)\times count(c_i)\times count(r_j)}。\quad(9)$$

将式（9）化简后可变为：

$$SO\_PMI(c_i)=\log_2\prod_{j=1}^{m}\frac{[count(c_i,p_j)+1]\times\alpha_j}{count(c_i,r_j)+1},\quad(10)$$

式中$\alpha_j=\dfrac{count(r_j)}{count(p_j)}$。（11）

在封闭的训练语料中，出现正、负面种子情感词的微博条数是固定的，因此α_j可看做一个常数，其取值范围为(0, + ∞)。在训练语料中，如果

$$count(c_i,p_j)=count(c_i,r_j),\quad(12)$$

即词c_i与p_j, r_j在训练语料中同现的微博条数相等，则c_i可视为中性词，即

$$SO\_PMI(c_i)=0。\quad(13)$$

而根据式（10）计算后有

$$SO\_PMI(c_i)=\log_2\prod_{j=1}^{n}\alpha_j,\quad(14)$$

若$\prod_{j=1}^{n}\alpha_j<1$，则$c_i$误判为正面情感词，若$\prod_{j=1}^{n}\alpha_j<1$，则$c_i$误判为负面情感词，为了避免上述$\alpha_j$给情感词倾向性判断带来的影响，本文将$\alpha_j$赋值为1，即

$$\prod_{j=1}^{n}\alpha_j=1。\quad(15)$$

因此，改进后的SO-PMI值的计算公式为：

$$SO\_PMI(c_i)=\log_2\prod_{j=1}^{n}\frac{count(c_i,p_j)+1}{count(c_i,r_j)+1}。\quad(16)$$

最终，候选微博情感词的情感倾向性可通过式（16）进行判断：若式（16）大于0则词c_i被判定为正面情感词；若式（16）小于0则词c_i被判定为负面情感词；若式（16）等于0则c_i被判定为中性词。将被判定为正面的情感词加入到*sp*中，被判定为负面的情感词加入到*sn*中，最后将加入了正、负情感词的*sp*与*sn*合并，组成微博领域情感词典。

3 实验结果与分析

3.1 实验数据选择

实验采用COAE2014任务四的评测语料，其中共有40 000条微博（含干扰数据）。首先，对评测语料进行分词、去除非法字符、数据格式规范化处理；然后，采用2.1~2.2节中的方式提取候选微博情感词、计算候选情感词的权值，并构造微博情感词典；最后利用构造的微博情感词典，对40 000条微博数据进行情感倾向性判断，以判别微博情感词典构建的质量。

3.2 种子词个数对微博情感词典的影响

为了考察种子词的选取对构建微博情感词典的影响，实验分别选取了5, 10, 15, 20, 25个TF-IDF值较高的正、负面种子情感词，并用所选取的情感词利用式（16）构建微博情感词典，采用准确率、召回率、*F-measure*对微博情感词典构建的效果进行量化分析，其中准确率（*presion*）、召回率（*recall*）、*F-measure*的计算方式如下：

$$presion=\frac{\text{正确识别的情感词个数}}{\text{识别出的情感词个数}},\quad(17)$$

$$recall=\frac{\text{正确识别的情感词个数}}{\text{识别出的情感词个数}},\quad(18)$$

$$F\text{-}measure=2*presion*recall/(presion+recall)。\quad(19)$$

实验结果如图3所示。

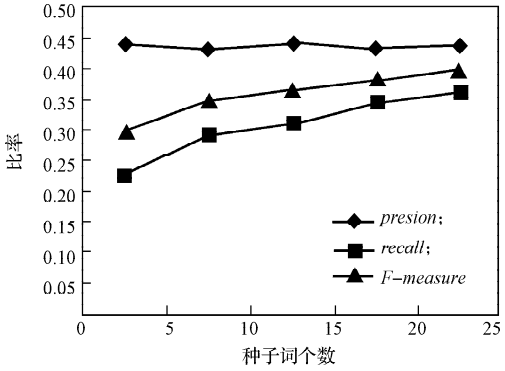

图3 情感词典准确率、召回率、*F-measure*与种子词个数的变化趋势

Fig. 3 The emotional dictionary *presion*, *recall* and *F-measure* changed with seed word numbers

从图3可知，随着种子词个数的增加，准确率、召回率、*F-measure*随之增加，表明情感词典构建越准确。分析其原因，主要由于种子情感词个数越多，其参与有效计算的情感词随之增加，从而减少个别情感词对候选情感词SO-PMI值的影响，因此，计算出的候选情感词的倾向性值的可信度较大，从而使识别出的有效情感词个数增加，准确率、召回率、*F-measure*随之增加。

3.3　微博情感词典的应用验证

为了验证微博情感词典在微博情感分析中的应用适用性，实验应用情感词典并采用基于规则的方法对40 000条微博进行情感倾向性判断，其判断结果如表2所示。

表2　微博倾向性分析结果

Table 2　Result of the microblog tendency analysis

方　法	*PosP*	*PosR*	*PosF*	*NegP*	*NegR*	*NegF*
本文方法	0.965	0.291	0.447	0.963	0.280	0.434
Hit_run3	0.962	0.262	0.412	0.962	0.175	0.296
Medians	0.891	0.299	0.445	0.850	0.281	0.428

表2中*PosP*, *PosR*, *PosF*分别为正面准确率、召回率和*F*值，*NegP*, *NegR*, *NegF*分别为负面准确率、召回率和*F*值，Hit_run3为采用基础情感词典[10]判断微博情感倾向性的结果，Medians为COAE2014评测中的平均值。由表2可知，利用本文方法构建的微博情感词典在情感分析中较Hit_run3效果要好，且高于评测中的平均值。分析其原因：文献[10]所构建的情感词典对网络情感词的覆盖度较低，而本文方法在构建的基础情感词典的基础上，采用改进的SO-PMI算法有效发现候选情感词中的网络情感词，因而识别效果较好，有效验证了本文方法构建的微博情感词典在微博情感分析中的有效性。

4　结语

本文针对SO-PMI算法在判断微博中候选情感词的倾向性时，对情感词倾向性判断不准的问题，在SO-PMI算法的基础上，引入拉普拉斯平滑技术对SO-PMI算法进行了改进。采用改进后的SO-PMI算法在COAE2014评测语料的基础上构建了微博情感词典，利用构建的微博情感词典对微博进行情感倾向性分析。实验结果表明，本文方法构建的情感词典在微博情感分析中具有较好的识别效果。

由于在分词过程中，存在候选情感词分词不准的问题，同时在构建词典中没有考虑微博中表情符号的情感倾向性，因此在应用情感词典进行微博情感倾向性判断时存在召回率不高的问题，这表明采用规则的方法进行倾向性判断存在一定的局限性，因此在规则的基础上融合机器学习的方法对微博进行倾向性判断将是下一步研究工作的重点。

参考文献：

[1] 杨　超，冯　时，王大玲，等. 基于情感词典扩展技术的网络舆情倾向性分析[J]. 小型微型计算机系统，2010，31(4)：691-695.
Yang Chao，Feng Shi，Wang Daling，et al. Analysis on Web Public Opinion Orientation Based on Extending Sentiment Lexicon[J]. Journal of Chinese Computer Systems，2010，31(4)：691-695.

[2] Lu Y，Castellanos M，Dayal U，et al. Automatic Construction of a Context-Aware Sentiment Lexicon：An Optimization Approach[C]//World Wide Web Conference Series. Newyork：ACM，2011，347-356.

[3] 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)：187-189.
Wang Zhenyu，Wu Zeheng，Hu Fangtao. Words Sentiment Polarity Calculation Based on HowNet and PMI[J]. Computer Engineering，2012，38(15)：187-189.

[4] 朱嫣岚，闵　锦，周雅倩，等. 基于HowNet的词汇语义倾向性计算[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 HowNet[J]. Journal of Chinese Information Processing，2006，20(1)：14-20.

[5] Turney P D，Littman M L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transactions on Information Systems，2003，21(4)：315-346.

[6] 知网.《知网》情感分析用词语集[EB/OL]. [2015-04-15]. http://www.keenage.com/html/c_index.html.
HowNet. Words Set for Sentiment Analysis of HowNet [EB/OL]. [2015-04-15]. http://www.keenage.com/html/c_index.html.

[7] 大连理工大学信息检索研究室. 大连理工大学情感词汇本体库[EB/OL]. [2015-04-15]. http://ir.dlut.edu.cn/EmotionOntologyDownload.asp? xutm_source= weibolife.
Information Retrieval Laboratory of Dalian University of Technology. Emotional Vocabulary Ontology Library of Dalian University of Technology[EB/OL]. [2015-04-15]. http://ir.dlut.edu.cn/EmotionOntologyDownload.aspx?utm_source= weibolife.

[8] ICTCLAS分词系统. ICTCLAS下载[EB/OL]. [2015-06-11]. http://ictclas/org/ictclas_download.aspx.
ICTCLAS Word Segmentation System. ICTCLAS Download [EB/OL]. [2015-06-11]. http://ictclas/org/ ictclas_download.aspx.

[9] 刘　群，李素建. 基于《知网》的词汇语义相似度的计算[EB/OL]. [2015-06-15]. http://www.docin.com/ p-655858216.html.
Liu Qun，Li Sujian. Lexical Semantic Similarity Computing Based on HowNet[EB/OL]. [2015-06-15]. http://www.docin.com/ p-655858216.html.

[10] 柳位平，朱艳辉，栗春亮. 中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29(10)：2875-2877.
Liu Weiping，Zhu Yanhui，Li Chunliang. Research on Building Chinese Basic Semantic Lexicon[J]. Journal of Computer Applications，2009，29(10)：2875-2877.

（责任编辑：申　剑）