文章编号:1001 - 9081(2009)10 - 2875 - 03

中文基础情感词词典构建方法研究

柳位平,朱艳辉,栗春亮,向华政,文志强
(湖南工业大学 计算机与通信学院,湖南 株洲 412008)
(sallow08@gmail.com)

摘 要:词语的情感倾向判别是文章语义情感倾向研究的基础工作。利用中文情感词建立一个基础情感词典,为专一领域情感词识别提供一个核心子集,能够有效地在语料库中识别及扩展情感词集,并提高分类效果。在中文词语相似度计算方法的基础上,提出了一种中文情感词语的情感权值的计算方法,并以 HOWNET 情感词语集为基准,构建了中文基础情感词典。利用该词典结合 TF-IDF 特征权值计算方法,对中文文本情感倾向进行判别,实验结果表明,该方法取得了不错的分类效果。

关键词:基础情感词词典;倾向性分析;情感权值;种子词
中图分类号: TP18 **文献标志码**:A

Research on building Chinese basic semantic lexicon

LIU Wei-ping, ZHU Yan-hui, LI Chun-liang, XIANG Hua-zheng, WEN Zhi-qiang
(*Institute of Computer and Communication, Hunan University of Technology, Zhuzhou Hunan* 412008, *China*)

Abstract: Judging the emotional tendencies of Chinese words is the basic work of the semantic emotional tendency study of text. Building a basic emotional lexicon with Chinese emotional words will provide a core subset for identifying emotional words in a special area. It is able to identify and enlarge emotional word set effectively in corpus and also improve the efficiency of classification. A method of calculating the emotional value of Chinese emotional words on the basis of the similarity of Chinese words was provided. And also a Chinese basic emotional lexicon dictionary was constructed based on the HOWNET emotional word set. The emotional tendencies of Chinese texts were judged through the dictionary together with TF-IDF. Experiments show that this method has achieved a satisfying result.

Key words: basic semantic lexicon; orientation analysis; semantic weight; seed word

0 引言

词语的情感倾向分为带有褒义主观色彩和贬义色彩的情感词,而情感词语的情感倾向权值是指该词语的主观色彩强度,主观色彩越强情感倾向权值分数越大。在中文词语集中,有很多词本身就具备很强的主观色彩,比如“喜欢”、“漂亮”、“幸福”等褒义词语在任何主题的文章中,如果不考虑否定前缀词[1]或其他一些副词连词的影响,它们都带有很强的褒义色彩;而“厌恶”、“残酷”、“暴力”等是带有贬义倾向的词语。这些情感词在所有领域的文本情感倾向性分析中都是重要的特征词语,在本文中我们把这些具备跨领域能力的情感词叫做基础情感词。如图 1 所示:三个圆表示在三个不相关的领域文本集合中具有情感倾向的词语集合,它们的交集(阴影区)我们称之为基础情感词语集。对这些情感词的识别以及情感倾向权值的赋值是中文文本情感倾向分析的重要步骤[2]。本文研究从具有情感倾向的中文词语中挑选常用的情感词够成一个基础情感词语集,并采用词语相似度方法[3]计算出每个词的情感倾向权值,构成一个基础情感词词典。

本文对于词语的情感褒贬程度用一个[-1, +1]的实数度量,情感权值小于 0 的为贬义词,大于 0 的为褒义词,情感权值的绝对值越大,情感强度越强[4]。

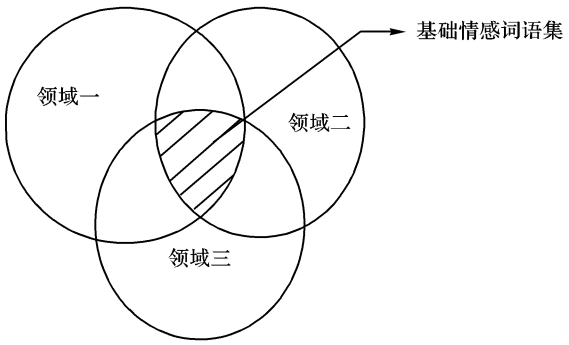

图 1 基础情感词语集

本文使用的基础情感词以 HowNet[5] 发布的情感词语集为基础,通过人工挑选,去掉一些非常用及情感倾向不明显的词语,得到 6196 个情感词,然后计算它们的情感倾向权值,去掉分类不正确的词以及权值过低的中性词,最后得到 5281 个基础情感词。

收稿日期:2009 - 05 - 08;**修回日期**:2009 - 07 - 07。
基金项目:湖南省自然科学基金资助项目(05JJ30122);中国包装总公司科研资助项目(2008-XK13);湖南省教育厅科研资助项目(07B014);湖南工业大学研究生创新基金资助项目(CX0812)。

作者简介:柳位平(1981 -),男,湖南邵阳人,硕士研究生,主要研究方向:文本分类; 朱艳辉(1968 -),女,湖南湘潭人,教授,CCF 高级会员,主要研究方向:智能信息处理、信息检索、文本分类; 栗春亮(1984 -),男,河北邯郸人,硕士研究生,主要研究方向:文本分类; 向华政(1971 -),男,湖南桃源人,副教授,博士研究生,主要研究方向:智能信息处理; 文志强(1973 -),男,湖南湘乡人,副教授,博士,主要研究方向:目标检测、智能信息处理。

1 基础情感词倾向性分析及词典构建

1.1 基础情感词倾向性分析

本文的研究侧重于给每一个基础情感词赋予一个语意倾向的度量值,我们采用知网的语意相似度计算公式计算两词语之间的相似度。知网中词语相似度的计算是以词的义原为基础得来的。在知网中将同类的义原组成一棵树,因而把义原的相似度计算转化为义原之间在树中的语义距离计算。假设两个义原在这个层次体系中的路经距离为 d,则这两个义原的语义距离为式(1) 所示[3]:

$$Sim(p_1,p_2) = \frac{\alpha}{d+\alpha} \tag{1}$$

其中:p_1、p_2表示义原;α 是一个可调节的参数。

知网中一个词有几个义原,因而在此基础上计算词语的相似度时,取义原之间相似度最大值作为词语的相似度。对于两个中文词语 w_1,w_2,假设它们分别有多个义原,w_1 的义原为 s_{11},s_{12},…,s_{1n},w_2 的义原为 s_{21},s_{22},…,s_{2m},因而它们的相似度计算如式(2) 所示:

$$Similarity(w_1,w_2) = \max_{i=1\ldots n,j=1,\cdots,m}(s_{1i},s_{2j}) \tag{2}$$

情感词的情感权值大小由这个词与种子词的语义关联的紧密程度有关,这里的种子词是指褒贬态度非常明显、强烈,具有代表性的词语。与褒义种子词联系越紧密,则词语的褒义倾向越强烈。与贬义种子词联系越紧密,则词语贬义倾向越明显[6]。我们假设选定用 *Key_p* 表示褒义种子词,褒义种子词数目为 M,*Key_n* 表示贬义种子词[7],贬义种子词数目为 N,M 和 N 可以相等,也可以不等。情感词语 w 的情感倾向值用 *SO-IR*(w) 表示,以 0 作为默认阈值,最终倾向值大于阈值为褒义,小于阈值为贬义。*SO-IR*(w) 的数值表示 w 的情感强度,值越大情感强度越强。我们提出计算情感词语 w 的情感倾向权值如式(3) 所示:

$$SO\text{-}IR(w) = \frac{\sum_{i}^{M} Similarity(Key\_p_i,w)}{M} - \frac{\sum_{i}^{N} Similarty(Key\_n_i,w)}{N} \tag{3}$$

实验中,式(3) 中的 *Similarity*(key,w) 采用式(2)进行计算。

本文实验步骤如下。

1)选择基础情感词:本文使用的基础情感词以 HowNet 发布的情感词语集为基础,通过人工挑选去掉一些不太常用或者情感倾向不很明显的词语,比如:“零 ”、“怿”、“凹凸“等词语[5],最后得到褒义词 3 219 个,贬义词 2 905 个,最终的基础情感词词典包含 6 196 个基础情感词。

2)设置测试集:以第一步建立的基础情感词作为测试词集,实验中共建立了 5 组测试集(如表 1 所示)。测试集 1 使用的是第 1)步建立的基础情感词集合。测试集 2 ~ 5 中的情感词语是在测试集 1 的基础上随机产生的。

3)选择种子词:将 5 组测试集中的词按 Google 搜索返回的 hits 数进行排序,选择 hits 数最高的词为种子词,在种子词的数量选择上,我们按照测试集中情感词的比率进行选择,褒义词和褒义词分别选择了 hits 数最高的 5%,10%,15%,20%,25%, 30%,40%, 50%,60%,70%,80%,90%,100% 作为种子词集合,对每一个测试集使用上文的式(3)进行实验。

表 1 测试集

测试集编号	褒义词数	贬义词数	总词数
1	3 291	2 905	6 196
2	1 778	402	2 180
3	1 768	862	2 630
4	342	1 251	1 593
5	1 737	678	2 415

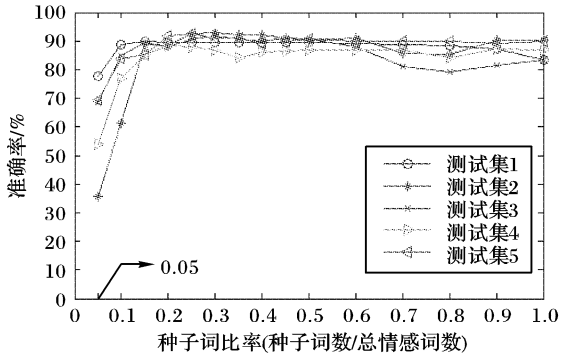

图 2 种子词数量对基础情感词倾向性判别准确率的影响

实验结果如图 2 所示。横坐标表示种子词数占总情感词数的比例,纵坐标表示词语倾向性判断的准确率,实验结果表明,五个测试集均在种子词数量为总情感词数量的 15% 左右时情感倾向性判断准确率达到峰值 90% 左右,并且在达到峰值后准确率都趋向稳定。在上述实验中,实验的数据量规模较小,因为常用的具有较强的词性的词语数量有限(实验数据选择知网情感词集),而且我们想建立的基础情感词典只是想找出一个比较精简的不同领域的公共核心子集,所以数据量的使用上有一定限制。为了弥补数据规模小而使实验结果可能出现的偏差,我们通过设置多组实验的方法进行论证。共设置了五组不同的实验数据进行实验,得到类似的结论,有力的证明了该实验的结论。

1.2 基础情感词词典的构建

测试集 2 ~ 5 是测试集 1 上的子集,因而我们采用测试集 1 作为建立情感词词典的实验数据,在 1.1 节的实验中,种子词数占总数 15% 时情感倾向判断准确率达到最大值 90.16%,因此我们使用种子词数占总数 15% 时计算得到的基础情感词权值作为实验结果,在此基础上去掉分类判断不正确的词语,得到正确的褒义词数 3 093 个, 贬义词 2 480 个,总词语数 5 573 个。

在实验中得到的情感词语的情感权值普遍偏小。使用线性方法对其值进行重新规划。计算公式如下:

$$d' = \frac{d - d_{\min}}{d_{\max} - d_{\min}} \tag{4}$$

其中:d 是根据式(3) 计算得到的情感权值;d' 是规划后的情

感词情感权值;d_{min} 表示式(3)计算出的所有情感权值中的最小值;d_{max} 为最大值。

进行数值规划之后有一些权值较小的偏向中性的词,如:“甘愿 0.038 232 6”、“凝神 0.047 556”、“回礼 0.044 343 1”等,这一类词语的极性不是很强对情感分类的贡献也较小,所以我们将情感权值的绝对值小于 0.05 的词语全部去除,不包含在基础情感词典中。最后基础情感词典中包含正面词语 2 807个,负面词语 2 474 个,总词数 5 281 个。摘取部分情感词典如表 2 所示。

表 2　情感词典部分列表

正面词汇	权值	负面词汇	权值
辉煌	0.881 735	罪恶	-0.981 231
美妙	0.878 707	诅咒	-0.941 896
漂亮	0.878 707	责备	-0.675 592
俱佳	0.815 118	丑陋	-0.675 898
动听	0.810 184	丑恶	-0.675 898
体面	0.800 897	藐小	-0.646 756
淳美	0.780 568	累赘	-0.652 433

2　基础情感词词典的应用

建立基础情感词词典的目的是为了使用该字典对文本进行情感分类。我们使用 SVM 方法对中文文档进行情感倾向性分类实验,实验使用文献[8]中的语料集,正面文章共 1 000 篇,其中用作训练集 700 篇,测试集 300 篇。负面文章 1 000 篇,其中训练集 700 篇,测试集 300 篇。使用上文中的基础情感词词典作为文本特征,分类特征词的权值采用 TF-IDF 计算公式与情感词汇的情感权值相结合,我们对 TF-IDF 计算公式改进如下:

$$w(t,\bar{d}) = \frac{tf(t,\bar{d}) \times \log(N/n_t + 0.01) \times \overline{w_t}}{\sqrt{\sum_{t \in d} [tf(t,\bar{d}) \times \log(N/n_t + 0,01) \times \overline{w_t}]^2}} \tag{5}$$

其中:$w(t,d)$ 为词 t 在文本 d 中的权重,而 $tf(t,d)$ 为词 t 在文本 d 中的词频;N 为训练文本的总数;n_t 为训练文本集中出现词 t 的文本数;$\overline{w_t}$ 为特征词情感权值。

我们在文本情感倾向分类实验中做了三组对比实验,第一组采用普通的 TF-IDF 方法进行分类处理,即在式(5)中不含 $\overline{w_t}$ 项。第二组利用了情感词词典,但将所有的正面情感词汇的情感权值赋值为 1,负面情感词语的权值赋为 -1。第三组采用了 1.2 节所计算的情感词权值并代入式(5)进行特征词权值计算。这三组实验的实验结果如表 3 所示。

表 3　利用基础情感词词典进行文本情感倾向分类的实验结果比较

实验集标号	第一组	第二组	第三组
准确率/%	46.37	70.62	82.10

实验结果表明,单纯利用 TF-IDF 方法来进行中文文本的情感倾向分类所得到的分类精度很低,仅 46.37%,而利用情感词汇作为特征后所得到的精度有比较大的提高,特别是用到基础情感词词典并使用情感权值时精度进一步提高,准确率为 82.1%。总体来说我们建立的基础情感词典在很大程度上提高了分类精度,但是分类准确率还有进一步的提升空间。在分析了该预料库数据后发现。该预料库中每一篇文章的内容都比较短,而我们建立的是一个核心的较小基础情感词典,所以能提出的情感特征项比较少,甚至我们在实验中发现有正面和负面的文章最后提出的特征向量是相同的,所以影响了分类效果,针对这一问题,我们在以后的工作中会在本文的基础上针对某一领域利用基础情感词典来扩建领域情感词典,再加上否定、肯定转折、递进以及一些程度副词等来进一步提高分类效果。

3　结语

本文所提出的方法是作者在中文文本情感倾向分析中所做的一些基础研究工作,文章对情感词权值计算及基础情感词词典的建立进行了研究,提出的情感词权值计算方法不要求种子词数量相等,研究和实验结果表明,基础情感词典的构建和利用对中文文本情感倾向分类的精度有很大的提高。

本文建立的基础情感词词典还有一定的局限性,比如,在情感词语的选择上是以 HowNet 为基础,其精度对 HowNet 情感词集的依赖性较强;情感权值计算的合理性还有待进一步完善和例证,文本的分类准确率还有进一步的提升的空间。在后续的研究工作中将进一步完善该基础情感词词典,并利用该词典进一步探索针对某一特定领域的大规模预料库中建立[9]专用情感词典的方法,以进一步提高分类效果。

参考文献:

[1] KU L-W, LO Y-S, CHEN H-H. Using polarity scores of words for sentence-level opinion extraction [C]// Proceedings of the 6th NTCIR-6 Workshop Meeting. Toyko, Japan: [s. n.], 2007: 316 - 322.

[2] 王秉卿, 张姝, 张奇. 中文情感词识别[C]// NCIRCS2008: 第四届全国信息检索与内容安全学术会议. 北京:[出版社不详], 2008: 63 - 69.

[3] 刘群, 李素建. 基于《知网》的词汇语义相似度的计算[J]. 中文计算语言学, 2002, 17(2): 59 - 76.

[4] 王克, 张春良, 朱慕华, 等. 基于情感词词典的中文文本主客观分析[C]// NCIRCS2008: 第四届全国信息检索与内容安全学术会议. 北京:[出版地不详], 2008: 56 - 62.

[5] 知网[EB/OL]. [2009 - 03 - 12]. http://www.keenage.com.

[6] 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.

[7] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 2002: 417 - 424.

[8] 谭松波. 中文情感挖掘语料—ChenSentiCorp [EB/OL]. (2008 - 12 - 19) [2009 - 03 - 12]. http://www.searchforum.org.cn/tansongbo/corpus-senti.htm.

[9] KAJI N, KITSUREGAWA M. Building lexicon for sentiment analysis from massive collection of HTML documents [C/OL]// EMNLP-CoNLL 2007: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007: 1075 - 1083 [2009 - 03 - 08]. http://www.aclweb.org/anthology/D/D07/D07-1115.pdf.