一种基于多重词典的中文文本情感特征抽取方法

朱艳辉，栗春亮，徐叶强，柳位平

（湖南工业大学 计算机与通信学院，湖南 株洲 412008）

摘 要：情感特征抽取是文本情感分类的重要步骤，正确的选择情感特征并赋予合理的情感权重是保障分类精度的前提。利用基础情感词词典、连词词典及词语距离，提出了一种基于多重词典的中文文本情感特征抽取算法，实验证明该方法优于HM, SO-PMI和词语语义距离等经典的特征抽取算法。

关键词：情感特征；情感权重；多重词典；情感特征抽取算法

中图分类号：TP301 **文献标志码**：A **文章编号**：1673-9833(2011)02-0042-05

A Method of Emotional Feature Extraction in Chinese Text Based on Multiple Lexicons

Zhu Yanhui，Li Chunliang，Xu Yeqiang，Liu Weiping

（School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412008，China）

Abstract：Emotional feature extraction is an important step in text sentiment classification, so choosing emotional feature correctly and giving a reasonable sentiment weight are the premise to guarantee classification precision. A Chinese text emotional feature extraction algorithm is proposed based on multiple lexicons including basic semantic lexicon, conjunction lexicon and word distance. The experiment results show that the algorithm outperforms some classic feature extraction algorithms of HM, SO-PMI and word semantic distance etc.

Keywords：emotional feature；sentiment weight；multiple lexicons；emotional feature extraction algorithm

0 引言

目前国内外对文本主题的分类研究已经比较深入，但是对文本情感分类的研究还处在一个较初级的阶段。近年来，国内外已有不少学者在文本情感分类方面进行了研究，并取得了一些有效的方法，比如贝叶斯、最大熵、支持向量机等机器学习方法。而情感特征抽取是文本情感分类的重要步骤，准确的情感特征以及合适的情感权重是保证情感分类性能的关键。

已有不少学者对文本情感特征抽取方法进行了研究，Hatzivassiloglou和Mckeown提出了经典的HM算法，利用形容词之间的连词存在语言学上的限制（连词连接的2个词表示相同或相反的态度），将语料库中的形容词聚类为正性词汇和负性词汇，以判断形容词的情感特征[1]。Peter D. Turney提出了SO-PMI算法[2]，使用一个词和强烈表示正面倾向的词“excellent”的互信息，减去它和强烈表示反面倾向的词“poor”的互信息，计算这个词的情感倾向。姚天昉等人通过研究发现，在中文文本中只有25.5%的有效句子中出现了关联词，如果使用HM算法，则查全率较低，并且在实验中发现，5.5%的句子中，2

收稿日期：2010-12-16

基金项目：湖南省自然科学基金资助项目（10JJ3002），中国包装总公司科研基金资助项目（2008-XK13）

作者简介：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，主要研究方向为智能信息处理，信息检索，文本分类，E-mail：swayhzhu@163.com

个关联词之间出现了转折词，由此看出现在同一个句子当中的情感词、词语距离相近的情感词，它们的情感倾向未必相同，因此上述算法均有一定的局限性[3]。

本文利用基础情感词典、连词词典结合词语距离算法提出一种基于多重词典的中文文本情感特征抽取方法。采用的方式是：首先利用基础情感词词典初步提出文本情感特征集 C_1，然后利用连词词典提取与 C_1 中的情感特征相关联的情感特征组成情感特征集 C_2，最后将已抽取的情感特征集 C_2 作为种子词，进一步利用词语距离算法提取未识别出情感倾向的情感词，形成最终情感特征集 C。 实验结果表明，使用多重词典的中文文本情感特征抽取方法，其准确率、查全率和 F 值要高于经典的 HM, SO-PMI, 词语距离等算法。

1　基于多重词典的中文文本情感特征抽取算法

1.1　基于基础情感词词典的情感特征抽取

基础情感词词典[4]是由中文中常用的情感词语构成，这些词语具有强烈的情感倾向，因此基础情感词常用作文本情感分类的情感特征。由于中文文本与英文文本的表达方式不同，在对中文文本进行特征提取时，首先利用中科院的 ICTCLAS[5]对文本进行分词，利用基础情感词典抽取情感词，具体方法如算法 1 所示。

算法 1　基于基础情感词词典的情感特征抽取方法

// c_i 表示已经抽取的情感特征项，初始为空

输入：文档 d_i，$c_i\{\ \}$

输出：情感特征集 $c_i\{\ w_1,w_2,w_3,\cdots,w_k\}$

Begin

1）使用分词系统对 d_i 分词

2）对分词后的 d_i 进行预处理

去掉除句号、问号、感叹号的所有标点符号

去掉人名、地名、时间以及助动词

3）对于分词后的每一词语 w_i

if w_i 属于基础情感词典

将 w_i 加入到 c_i 中

endif

4）循环执行第 3 步，直到文本中所有词语都得到处理

end

例如：“首先是让我满意的酒店接机服务，这一点是和大家取得共识的了。我是来酒钢办事的，去酒钢办事在这里住还是非常方便的，没有车服务员会热情周到的帮我订车。让我比较感动的是第二天要去办事，前一天却发现衣服挂了口子，因为着急出去，没时间处理，而且本人缝纫手艺也比较差。在情急之中找到了服务员，回来的时候衣服补好了，并且皮鞋也擦亮了，一点没耽误事。在这里对她们的服务表示小小的感谢！”该实例来自谭松波的“中文情感挖掘语料 -ChnSentiCorp”[6]中关于酒店评论的一篇语料。首先进行文本分词，依据基础情感词词典对该文本抽取情感特征，在该实例中抽取的情感特征为：C_1 = {满意，方便，热情，周到，感动，差，耽误，感谢}。抽取的这些特征都是存在于基础情感词典的情感词，它们具有强烈的情感特征，能够在一定程度上表达该语料文本的情感倾向。

1.2　基于连词词典扩展情感特征

1.2.1　连词词典的构建

在中文的表达方式中，连词起到连接词与词、短语与短语以及句与句的作用。如果一个连词连接的 2 个词语中，有一个是带有情感倾向的词语，那么可以判定与它连接的另外一个词语也带有一定的情感倾向。利用这一方法在已知的部分情感特征词语的基础上进一步查找未知的情感特征项。收集这一类相关的中文连词，构建连词词典。

能够用来抽取文本情感特征的连词只有 3 类：转折，递进和并列。这 3 类连词在中文文本中对文章的主观情感有较大的影响而且有着各自不同的作用。转折连词连接的 2 个情感词，它们具有相反的情感倾向；递进连词连接 2 个情感倾向相同的情感词语，并且连词后的情感倾向要强于连词前的情感倾向。同时递进连词与转折连词一样，不同的递进连词有不同的语气程度；并列连词连接 2 个情感倾向相同的词语，与前面 2 类连词不同的是，并列连词连接的两词语情感权重相同。

根据连词的上述特性使用转折，递进和并列 3 类连词构建连词词典，如表 1 所示。

表 1　整理得到的 3 类连词集

Table 1　Three types of organized conjunction collections

连词类型	连词
转折	但、可是、然而、不过，却、但是、偏偏、只是、不过、至于、不料、岂知
递进	而且，更加，甚至，不如、不及，乃至，并且，乃至，况、况且
并列	和、跟、与、同、及、何况

依据连词的特性，连词词典不定义权重值。但为

了区分3类连词，对这3类连词定义标志值，转折连词标志值设置为-1，并列连词为0，递进连词为1。

1.2.2 基于连词词典扩展情感特征

基于基础情感词词典能够抽取文章中具有强烈情感倾向的情感特征，但是在中文句子表达中，除了基础情感词之外还有其它能够表达情感倾向词语。比如："你的想法很好，但是不可行。"在该句中存在基础情感词"好"，具有正面的情感倾向，但如果在此例中只以基础情感词作为情感特征，明显会误判。该例句的情感倾向由词"不可行"表达，从整个结构来讲，该例是负面的情感倾向。

利用连词词典，在算法1抽取的情感特征基础上进一步抽取情感特征项。依据连词的特性来抽取，并列连词连接的情感词情感特征权重相同，转折连词连接的情感倾向相反，递进连词依据其所处在连词的位置而确定，递进连词后的情感词倾向强于前面情感词倾向。基于连词词典扩展情感特征算法如算法2所示。

算法2 基于连词词典扩展情感特征算法

// c_i 表示算法2中抽取的情感特征集，$O(w)$ 设为词 w 的权重

输入：文档 d_i，阈值 t, $c_i\{w_1,w_2,w_3,\cdots,w_m\}$

输出：$c_i\{w_1,w_2,w_3,\cdots,w_n\}$

Begin

1）在文本 d_i 中查找候选情感词 w

2）if w不存在于c_i中

（1）求 w与种子词的关联度

SO-PMI(*w*)

（2）If *SO-PMI*(*w*)> t // t一般取0.05

（ⅰ）判断 w 的情感倾向

（ⅱ）计算 w 的情感权重

（ⅲ）将 w 及权重加入到 c_i 中

3）Repeat step 1 until to the end of document d_i

End

对1.1中的例子进一步抽取情感特征，在实验中发现递进连词"而且"，在其后查找到一个形容词"差"存在于特征情感词集中，所以进一步在"而且"的前面查找出"着急"，因此将该词加入到情感特征集中，由于该词处于递进连词的前面，所以将它的权重设置为"差"的一半。在该实例中进一步抽取的情感特征为 C_2= {满意，方便，热情，周到，感动，差，耽误，感谢，着急}。

1.3 基于词语距离进一步扩展情感特征

词语距离判断法的基本思想是：如果一篇文章表现出一种情感倾向，那么在该文章中情感词会群集出现，它们共同决定一篇文章的情感倾向，也就是说一篇文章当中出现的情感词语的情感倾向大部分是一致的，并且情感词群之间是共现的，词语之间的距离越近，情感倾向也就越相似[7]。在计算候选情感词与种子词之间的关联度时采用SO-PMI（semantic srientation-sointwise mutual information）算法。

SO-PMI算法是在PMI上演化而来，PMI是用来计算2个元素的相似性[8]。假设2个词 w_1 和 w_2 之间的PMI定义如式（1）所示，

$$PMI(w_1,w_2)=\log_2\left(\frac{p(w_1\ \&\ w_2)}{p(w_1)p(w_2)}\right), \quad (1)$$

式中：$p(w)$表示词语 w 出现的概率；

$p(w_1\&w_2)$表示词 w_1 和 w_2 共现的概率。

SO-PMI是在PMI的基础上判断词语的情感倾向。假设要计算词语 w 的情感倾向，用词 w 分别对每一个种子词使用PMI算法计算相似度，然后计算词 w 和 k 个正面词的关联权重以及与 k 个反面词的关联权重之差，计算如式（2）所示，

$$SO\text{-}PMI(w)=PMI(w,\{K_p\})-PMI(w,\{K_n\}), \quad (2)$$

式中：$\{K_p\}$表示正面种子词集；

$\{K_n\}$表示负面种子词集。

SO-PMI(w) > 0，则词 w 的情感倾向为正面，反之为负面。由于种子词是选择情感极性较强的词语，因此 *SO-PMI*(w)数值越大，词 w 的情感极性越强。

通过算法1和算法2，从文本中选取了一部分情感特征项：

1）利用了基础情感词典，文本中所有的基础情感词都是要选择的情感特征项，该类词对于文本的情感倾向分类的正确与否贡献度最大。

2）根据连词的特性，利用连词词典，进一步扩充文本的情感特征项，并且确定了其情感权重。

接下来利用词语距离的方式从剩余的候选集中提取情感特征集，选用算法2中提取的情感特征作为种子词集。具体步骤如下：

1）对于文本 d 中的候选情感词 w，使用公式（2）计算 w 与种子词之间的关联度 $SO\text{-}PMI(w)$，如果 *SO-PMI*(w)>T（T一般取0.05）[7]，则将 w 加入情感特征集。

2）判断 w 的情感倾向，计算方式如式（3）。

$$SO\text{-}DIST(w)=\frac{1}{\mu}\sum_i^m\log_2\frac{N-1}{Dist(w,p\_key_i)}-\frac{1}{\upsilon}\sum_i^K\log_2\left(\frac{N-1}{Dist(w,n\_key_i)}\right), \quad (3)$$

式中：$\mu = \sum_{i}^{m} Dist(w, p\_key_i)$；　　（4）

$$\upsilon = \sum_{j}^{K} Dist(w, n\_key_j)；　　（5）$$

p_key_i 表示算法 2 提取的正面情感特征；

n_key_i 表示同一文本中负面情感特征；

$Dist(w, p_key_i)$为 w 与正面情感特征项 i 的距离；

对于词语 w，如果 $SO\text{-}DIST(w)$大于 0 则判定其为正面的情感倾向，否则为反面。

3）依据词语距离算法原则，使用如下方式计算 w 的情感权重：如果 w 判定为正面倾向则使用文本中正面的情感词汇作为其权重的计算依据，反之使用负面词汇，计算方式如式（6）。

$$O(w)' = \sum_{i}^{k} \frac{d_i}{D} * O(seed_i)，　　（6）$$

式中：$D = \sum_{i}^{k} d_i$；

$seed_i$ 表示情感词集中与 w 相同倾向的情感词；

$O(seed_i)$表示情感特征集中情感词的情感权重；

d_i 表示候选特征 w 与 $seed_i$ 的词语距离。

算法如下。

算法 3　基于词语距离的情感特征选择

// c_i 表示算法 1 中抽取的情感特征集，$O(w)$ 设为词 w 的权重

输入：文档 d_i，$c_i\{w_1, w_2, w_3, \cdots, w_k\}$

输出：情感特征集 $c_i\{w_1, w_2, w_3, \cdots, w_m\}$

Begin

1）在文本中查找连词

2）if 存在连词 $conj(w)$

（1）在 $conj(w)$所在句中查询 2 个相同词性的词语 w_1, w_2

（2）if w_1, w_2 中任意一个存在于 c_i

// 假设 w_1 存在于 c_i 中

（i）w_2 加入 c_i

（ii）if $conj(w)$是并列连词

赋予 w_1, w_2 相同的情感权重

else if $conj(w)$是递进连词

if w_2 在 $conj(w)$前

$O(w_2) = O(w_1)/2$

else $O(w_2) = O(w_1)*2$

else if $conj(w)$ 是转折连词

$O(w_2) = -O(w_1)$

endif

3）循环执行步骤 2，直到文本结束

End

对 1.1 中的文本进一步抽取情感特征，使用算法 3 提取后情感特征为：C = {满意，方便，热情，周到，感动，差，耽误，感谢，共识，着急，没时间，情急，补好，擦亮，服务}。

2　实验结果分析

实验语料采用谭松波的“中文情感挖掘语料－ChnSentiCorp”中关于酒店评论的语料，总共4 000篇，正负语料各 2 000 篇，从这 4 000 篇语料中随机抽取 200 篇文章，正、负面文本各 100 篇作为训练语料集，手工对其进行词语级的情感词标注，标注的方式为：按照词语所处的上下文环境进行语义标注，如“不开心”，标注为负面情感词。

实验利用基于多重词典的情感特征抽取算法识别出测试语料中的情感词，实验中将多重词典的情感特征抽取算法与 HM 算法，SO-PMI 算法，词语语义距离算法进行了比较，基于多重词典取得了较好的效果，准确率达到 0.940，覆盖率达到了 0.893。实验结果如表 2 所示。

表2　情感特征识别实验结果比较

Table 2　Comparison of the proposed experiment results and other semantic orientation identification methods

算 法	准确率 /%	查全率 /%	F-measure/%
HM	82.3	57.5	67.7
SO-PMI	84.6	80.4	82.5
词语语义距离	83.2	72.0	77.2
基于多重情感词典	94.0	89.3	91.6

3　结语

本文利用基础情感词典，连词词典以及词语距离算法提出一种基于多重词典的情感特征抽取方式。首先利用基础情感词典对文本抽取基础情感词作为文本的情感特征，然后在已抽取的情感特征基础上利用中文连词特性提取与情感特征连用的情感词作为情感特征，第三步利用词语距离算法在未知的情感倾向的候选情感词集中进一步提取情感特征。实验证明该方法优于经典的 HM, SO-PMI 和词语距离等算法。

参考文献：

[1] Hatzivassiloglou V，McKeown K R. Predicting the Semantic Orientation of Adjectives[C]//Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL. Stroudsburg，PA，USA：Association for Computational Linguistics，1997：174-181.

[2] Turney Peter. 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. USA：Association for Computational Linguistics，2002：417-424.

[3] 姚天昉，娄德成. 汉语情感词语义倾向判别的研究[C]//第七届中文信息处理国际会议论文集（ICCC2007）. 北京：电子工业出版社，2007：221-225.
Yao Tianfang, Lou Decheng. Research on Semantic Orientation Distinction for Chinese Sentiment Words[C]//7th International Conference On Chinese Computing（ICCC2007）. Beijing：Publishing House of Electronics Industry，2007：221-225.

[4] 柳位平，朱艳辉，栗春亮，等. 中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29(11)：2882-2884.
Liu Weiping，Zhu Yanhui，Li Chunliang, et al. Research on Building Chinese Basic Semantic Lexicon[J]. Journal of Computer Applications，2009，29(11)：2882-2884.

[5] ICTCLAS项目组. ICTCLAS汉语分词系统[EB/OL]. [2008-09-03]. http://ictclas.org/news_ictclas_publish.html.
ICTCLAS Project Group. ICTCLAS Chinese Word Segmentation System[EB/OL]. [2008-09-03]. http://ictclas.org/ news_ictclas_ publish.html.

[6] 谭松波. 中文情感挖掘语料库-ChnSentiCorp[EB/OL]. [2010-06-29].http://www.searchforum.org.cn/tansongbo/corpus-senti.htm.
Tan Songbo. Chinese Sentiment Mining Corpus[EB/OL]. [2010-06-29]. http://www.searchforum.org.cn/tansongbo/corpus-senti.htm.

[7] 宋 乐，何婷婷. 中文情感词句识别及文本观点抽取研究[C]//第二届中文倾向性分析评测会议. 上海：中国中文信息学会信息检索专业委员会，2009：30-37.
Song Le，He Tingting. Research on Sentiment Terms′ Polarities Identification and Opinion Extraction[C]//The Second Chinese Opinion Analysis Evaluation. Shanghai：Chinese Information Processing Society, Information Retrieval Technical Committee，2009：30-37.

[8] 李荣陆. 文本分类及其相关技术研究[D]. 上海：复旦大学，2005.
Li Ronglu. Research on Text Classification and Its Related Technologies[D]. Shanghai：Fudan University，2005.

[9] 王 鹏，樊兴华. 中文文本分类中利用依存关系的实验研究[J]. 计算机工程与应用，2010，46(3)：131-133.
Wang Peng，Fan Xinghua. Study on Chinese Text Classification Based on Dependency Relation[J]. Computer Engineering and Applications，2010，46(3)：131-133.

[10] 焦庆争，蔚承建. 一种可靠信任推荐文本分类特征权重算法[J]. 计算机应用研究，2010，27(2)：472-474.
Jiao Qingzheng，Wei Chengjian. Reliable Trust Recommendation Model for Feature Weighting in Text Categorization[J]. Application Research of Computers，2010，27(2)：472-474.

（责任编辑：罗立宇）