北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.016

基于情绪因子的中文微博情绪识别与分类

张晶　朱波　梁琳琳　侯敏†　滕永林

中国传媒大学国家语言资源监测与研究有声媒体中心，北京　100024; † 通信作者, E-mail: houmin@cuc.edu.cn

摘要　以情绪因子中的常用情绪词和情绪短语为基础构建情绪词典，并针对特殊的情绪表达形式，结合标点符号和表情符号在情绪分析中的功能，建立情绪规则库。系统通过对情绪词典和情绪规则的匹配和计算，实现对中文微博情绪的识别和分类，并在 2013 年 CCF 第二届自然语言处理与中文计算会议中文微博情绪分析评测中取得较好成绩。测试结果证明该方法有效。

关键词　情绪因子；情绪词典；情绪规则；情绪计算

中图分类号　TP391

Recognition and Classification of Emotions in the Chinese Microblog Based on Emotional Factor

ZHANG Jing, ZHU Bo, LIANG Linlin, HOU Min†, TENG Yonglin

Broadcast Media Language Branch, National Language Resources Monitoring and Research Center, Communication University of China, Beijing 100024; † Corresphoding author, E-mail: houmin@cuc.edu.cn

Abstract　Based on basic emotional words and phrases, an emotional dictionary was bulit. According to the special expressions of emotions and functions of punctuations and emoticons in the emotional analysis, a set of emotional rules were set up. The authors recognized and classified the emotions in microblogs according to the algorithm based on the emotional rules and dictionary, and achieved preferable result in the task of emotional analysis of Chinese microblogs in the 2nd Conference on NLP&CC hosted by CCF in 2013. Experiment results show that this algorithm can work effectively.

Key words　emotional factor; emotional dictionary; emotional rules; emotional computing

情绪是人们日常生活中一种不可或缺的心理活动，但由于情绪本身的复杂性，学者们对情绪的内涵有着不同的看法。孟昭兰[1]认为“情绪是多成分组成、多维量结构、多水平整合，并为有机体生存适应和人际交往而同认知交互作用的心理活动过程和心理动机力量”，她指出情绪的功能和结构；Campos[2]认为“情绪是个体与环境意义事件之间关系的心理现象”，Arnold[3]则认为“情绪是对趋向知觉为有益的、离开知觉为有害的东西的一种体验倾向”，两位学者都注意到了情绪和倾向的关系。情绪和评价同属人的主观意识，情绪句和评价句都属语言中主观句的一部分。正如 Lazarus 等[4]提到的“情绪依赖于短时或持续的评价”，情绪和评价是相互联系的，二者之间没有一道非此即彼的界限，但它们属于不同的范畴，二者之间有着本质的区别：评价是说话人对外在事物价值的评定，而情绪反映的则是人内在的心理状态，其外在的诱因可以存在也可以不存在。

学界不仅尚未对情绪有一个完整、系统的认识，对情绪的分类也是众说纷纭。我国古代就有“七情说”，即喜、怒、哀、惧、爱、恶和欲；《荀子•天论》中认为情绪应分为 6 类，包含好、恶、喜、怒、哀和乐；心理学家林传鼎[5]根据大徐本《说文》将情绪划分为 18 类：安静、喜悦、恨怒、悲痛、哀怜、

国家语委十二五规划重点项目(ZDI125-3)资助
收稿日期: 2013–07–05; 修回日期: 2013–08–26; 网络出版时间: 2013-11-11 10:25

忧愁、忿急、烦闷、恐惧、惊骇、恭敬、抚爱、憎恶、贪欲、嫉妒、骄慢、惭愧和耻辱；法国哲学家笛卡尔则认为人的原始情绪应有惊奇、爱悦、憎恶、欲望、欢乐和悲哀 6 类等。可见，对情绪的分类学界还没有达成共识。本文遵从 2013 年 CCF 第二届自然语言处理与中文计算会议中文微博情绪分析评测的要求，将情绪划分为喜好(lik)、高兴(hap)、悲伤(sad)、厌恶(dis)、愤怒(ang)、恐惧(fea)、惊讶(sur)7 类，并依据这 7 类情绪，对微博情绪进行分析研究。

目前，以语言工程为目标的情绪分析研究尚处起步阶段，Aman 等[6]采用基于知识的方法实现句子的情绪分析，Quan 等[7]基于情绪词实现对中文情绪语料库(Ren-CECps)中句子的情绪识别。本文认为，情绪词仅仅是语言中情绪表达的一种载体，语言中表达情绪的手段是多样的，可以将语言中所有表达情绪的载体统称为“情绪因子”。本文的目标是详细分析情绪因子，并在此基础上构建情绪词典和情绪规则库，从而实现对中文微博情绪的识别和分类。

1 情绪因子及其表现手段

情绪因子是语言中表达情绪的载体，也是进行情绪识别和情绪分类的主要依据。根据语言单位大小和表现手段不同，可以将情绪因子分为情绪词、情绪短语、情绪表达式、微博表情符号和标点符号 5 种类型。

1.1 情绪词

情绪词指的是能反映人内在的心理反应与感受的词语，大多是名词、动词、形容词。根据词义表达情绪的直接和间接，情绪词还可分为直接情绪词和间接情绪词两类。

直接情绪词指的是词义直接描绘情绪倾向的词语，如“快乐”(hap)、“钟爱”(lik)、“心酸”(sad)、“恼怒”(ang)、“纳闷”(sur)、“害怕”(fea)、“痛恨”(dis)等。这类情绪词应直接添加进情绪词典。间接情绪词指的是词义不是描绘情绪倾向，而是描述一种动作、行为或事物，在语用中才表现出某种情绪倾向的词语。如“控告”，作为一个动词，它不直接描绘情绪倾向，但由于“控告”多是对他人或集体不当行为的告发，其中蕴含着不满情绪，因此，在语境中它往往会传递出 ang 的情绪倾向。类似的词还有“强占”(ang)、“一手遮天”(dis)、“灾难”(fea)、“免单”(hap)等。这些词也属于情绪词，需添加到情绪词典。

情绪词是情绪因子的最基本内容，也是表达情绪最直接的手段。因此，提取情绪词并判断其情绪类别是情绪识别的首要工作。

1.2 情绪短语

情绪短语指的是能够表达一定情绪倾向的词组。情绪短语也是情绪因子的重要组成部分。因为短语大于词，表达语义更精准，表达情绪更明确，所以它在某种程度上还可以修正、改变或消除情绪词的情绪倾向。例如①:

例1 给脸不要脸，说轮流充电的是你们，等老娘充了电就说自己没用电。

例2 有时候，选择快乐，更需要勇气。

例 1 中没有情绪词，但是该句是一个情绪句，其情绪类别是 ang。例 2 中虽然有情绪词“快乐”，但是该句不具有情绪倾向。

根据构成成分的不同，情绪短语可以分为 3 种类型:

1) 修饰词+情绪词。

当修饰词是程度副词时，情绪短语的情绪类别与情绪词通常是一致的，只是所表达情绪的强烈程度有所增强或者减弱。如“特别开心”，情绪类别仍为 hap。但也存在着情绪转变的情况，如“真是讨厌”，“讨厌”的类别是 dis，前加修饰词“真是”后，其情绪类别是 ang，与“讨厌”不一致。

当修饰词是否定词时则会使情绪类别发生转变或消除情绪。如“不开心”的情绪类别是 sad 而不是 hap。“不必担心”则消除了“担心”的 dis 情绪，不再具有情绪倾向。

2) 修饰词+普通词。

有些情感短语是由修饰词和普通词组合而成的，如“特别热闹”(hap)是由修饰词“特别”与普通词“热闹”组成的情绪类别是 hap 的偏正短语，类似的还有“好消息”(hap)、“不好的预兆”(fea)等。

3) 普通词+普通词。

普通词和普通词组合一般是没有情绪倾向的，但在一定语境下，一些普通词的组合会产生情绪倾向。如“神马逻辑”，由普通词“神马”和“逻辑”组合成表达 ang 的情绪短语。类似的还有，“什么效率”(ang)、“想都不敢想”(fea)、“人品爆发”(hap)、“搞关系”(dis)、“容颜老去”(sad)、“我的亲娘啊”(sur)等。

① 本文中的例子均引自 2013 年 CCF 第二届自然语言处理与中文计算会议中文微博情绪分析评测提供的微博样例数据。

情绪短语作为情绪因子的核心内容, 不仅在情绪表达中发挥十分重要的作用, 同时也是提高句子情绪识别准确率的关键, 更是分析和研究中文微博情绪的重点。

1.3 情绪表达式

情绪表达式是指由一组非连续词语构成的能够表达情绪倾向的结构。情绪短语考察的是带有情绪倾向的连续出现的词构成的词组, 有时一组非连续出现的词语搭配也可用于表达情绪。例如:

例3 原来我的偶像肌肉男霍华德只比我高一个头啊。

上述例子中虽然不含情绪词和情绪短语, 但是有情绪倾向, 其情绪类别是 sur。分析表明, “原来……只”这种表达式使得语句带有了情绪倾向。虽然例子中“原来”和“只”中间隔着一些词, 但它们之间是有联系的, 并且共同决定句子的情绪类别。针对这一情况, 可以根据情绪表达式中词语搭配的特点制定情绪规则, 以实现对此类情绪句的识别。

1.4 微博表情符号

表情符号流行于网络交际, 它以简单图形或彩色图像甚至动画等来表情达意, 与语言的体态语相类似, 已形成一种显式的、固定的表达情绪方式。在微博中, 表情图像转写成文字, 并用“[ ]”将其框起。这类表情符号也是微博情绪分析的重要内容。例如:

例4 今天晒死了, 做张面膜先……[嘻嘻];

例5 心形的叶子呀[太开心];

例6 他最近顽皮得要命呀,经常乘着没人在家就来找东西咬,每次回家都要收拾一番[怒]。

上面 3 个例子都不含情绪词, 但都是情绪句, “开心”“愤怒”的情绪主要是通过表情符号“[嘻嘻]”“[太开心]”“[怒]”表达出来的, 如果将这 3 句中的表情符号去掉, 就很难理解并判断其情绪类别了。但微博语言是复杂的, 单纯依靠表情符号去判断有时也会产生偏差。例如:

例7 [嘻嘻][怒][爱你][爱你][哈哈]。

例7中表情符“[嘻嘻][哈哈]”的情绪类别是 hap, “[爱你]”的情绪类别是 lik, 而“[怒]”的情绪类别是 ang。看来, 当不同情绪类别的表情符号混用时, 仅仅是微博作者的一种随意表达, 句子不带有情绪倾向, 不具有分析和研究的价值。

1.5 标点符号

标点符号是辅助文字记录的语言符号,用来表示说话人的语气, 在帮助文字记录语言内容的同时, 起到补充句子信息的作用。标点符号在微博情绪识别中同样起到关键作用, 属于情绪因子的一部分, 也是理解和判断微博情绪的重要手段。例如:

例8 还敢再难吃点么!

例9 还敢迎接这样的太阳的光辉照耀吗?

上述两个例子都以副词“还敢”开头, 但是这两句的情绪类别却不一样, 分别是 ang 和 fea, 这与句末点号不同有关。“!”表示强烈语气, 突出强调了对饭菜不好吃的不满情绪; 而“?”表示疑问语气, 常用于质疑和怀疑, 在这里表示的是对大气污染造成的太阳光对人体危害的恐惧。

此外, 一些情绪类别中的标点符号是该句成为情绪句的关键。例如:

例10 鲁能跟国安的比赛据说不设客队球迷区??

例11 鲁能跟国安的比赛据说不设客队球迷区。

以上两句使用词语完全一样, 但是两者表达的内容却有本质的差别。第一句是说话人对该场足球比赛不设客队球迷区的消息感到惊讶, 体现的是一种 sur 的情绪; 第二句则没有明显的情绪倾向。这表明, 标点符号“? ? ”在表达情绪中发挥了重要作用, 也是第一句成为情绪句的关键。

当然, 微博中标点符号使用随意, 存在一些不规范的形式, 因此, 在情绪识别中, 不仅要充分重视和利用标点符号在表达情绪中所发挥的作用, 在处理时也要考虑全面。

2 微博情绪的计算方法

针对不同的情绪因子, 应采用不同的处理策略。我们运用情绪词典与情绪规则互动的方法, 通过对情绪词典及情绪规则的匹配和计算实现对微博情绪的识别。

2.1 情绪词典

情绪词典包含情绪词和情绪短语两部分, 是由人工建立并标注完成的。其来源主要包含 3 个方面: 1) 从许小颖等[8]对情感系统的分类归纳成果中吸收了一些直接情绪词; 2) 对大连理工大学建立的情感词汇本体词典[9]进行调整, 吸收其中一部分具有明显情绪倾向且常用的词语; 3) 通过对微博语料的分析研究, 增加微博中具有情绪倾向的网络用语和情

绪短语，如谐音词和短语(兴混、灰常难受)、繁体字词(欢喜)以及英文字母词(SB、high)等。目前，该情绪词典共收录2928个词语条目，各情绪类别包含的情绪词和情绪短语的具体数目如表1所示。

2.2 情绪规则

情绪规则主要处理情绪表达式、情绪符号以及特定词语与标点符号的组合。例如:

例12 难道真的是暗箱操作?

例13 难道真的是像他们说的，我太单纯了吗?

以上两个例子中都不含有情绪词，但是这两个句子是有情绪倾向的，其情绪类别是sur。分析表明，是短语“难道真的”和标点符号“？”搭配后形成的表达式使得语句带有了情绪倾向。在大规模语料中检索验证后，针对这种情况，制定如下情绪规则:

难道/% 真的/% # ？|?/w = #1:sur。

上述规则表示，当词语“难道真的”越过若干项，与句子末尾的中文或英文形式的“？”搭配出现时，确定该句子具有sur的情绪倾向。

情绪规则除了可以给未出现情绪词的情绪句赋予情绪类别外，也可以消解非情绪句中情绪词的情绪倾向。例如:

例14 我们之所以会痛苦，就是追求的太多。

例15 我们之所以不快乐，就是计较的太多，不是我们拥有的太少。

上述两个例子都含有情绪词，但都不是情绪句，需要借助情绪规则将这两句中的情绪词“痛苦”“快乐”的情绪倾向消解掉。针对这种情况，制定如下情绪规则:

之所以/% 会|不/% */emo = #3:0。

上述规则表示，当情绪词(emo)出现在“之所以会”或“之所以不”后时，规则左部第3项(#3)即情绪词的情绪值为0，即没有任何情绪。

情绪词典和情绪规则两者不是对立的，而是互动的。这种互动，不仅表现在词典要为规则准备好需要的各种资源，还表现在，有时一种语言现象，可以在词典里处理，也可以在规则中处理，这时需要分析各自的利弊，找到一条最合理的路径。一般来说，规则倾向于处理共性的问题，词典倾向于处理个性的问题。

微博语句情绪识别是微博情绪识别的基础，当进行完句子的情绪计算后再针对微博进行情绪计算，如一个微博中只有一个情绪句，则将其情绪类别作为该微博情绪的类。比如，如果整条微博中仅有一个情绪句，其情绪类别是sur，那么该微博的情绪类别则是sur。其他情况与此大致类似，不再赘述。目前，情绪规则库中共包含146条短语规则，46条句子规则以及4条微博规则，正是通过对这些情绪规则的匹配和计算，实现了中文微博的情绪识别任务。

3 系统及其工作流程

运用上述策略,构建一个基于情绪词典和情绪规则的情绪分析系统CUCeas。系统由分词标注模块和情绪计算模块两部分组成。分词标注模块不仅包含普通分词词典和规则，同时出于情绪分析的需要，又加进几部和情绪分析相关的词典，主要的两部是:一部用户词典Usreas，专门储存情绪分析需要切分出来的词和短语并赋予词性；一部包括情绪词、情绪短语并标有情绪类别(如sur、hap等)标记的情绪词典Diceas，用于给待分析语料中的情绪因子进行标记。情绪计算模块则是根据情绪计算规则(包括短语规则、句子规则以及篇章即微博规则)对句子和微博进行情绪计算，从而实现对微博和微博中句子情绪类别的判断。该系统工作流程见图1。

4 实验结果与分析

4.1 微博样例语料实验及结果

运行系统CUCeas，以2013年CCF第二届自然语言处理与中文计算会议中文微博情绪分析评测提供的微博样例语料为对象，对4000条微博样例语料进行了微博情绪识别的实验，并使用正确率、召

表1 情绪词典中各情绪类别包含的情绪词和情绪短语的数量

Table 1 Numbers of emotional words and phrases in emotional dictionary

情绪类别	情绪词	情绪短语	总计	情绪类别	情绪词	情绪短语	总计
hap	97	126	223	dis	343	465	808
lik	267	162	429	ang	124	245	369
sur	61	145	206	sad	182	151	333
fea	174	386	560	总计	1248	1680	2928

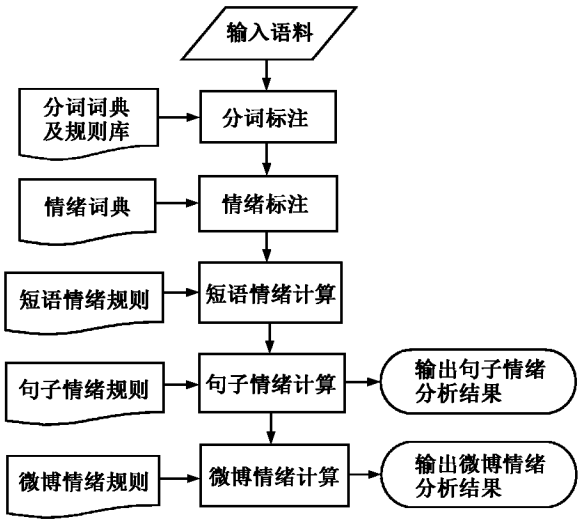

图 1　CUCeas 系统工作流程图

Fig.1　Working flow diagram of CUCeas

回率和 F 值来评估实验结果的性能。表 2 为实验结果。

表 2 数据表明，在含有情绪表达的微博中，各情绪类别的识别效果存在一定的差异，情绪“ang”的 F 值高于 0.55，而情绪“dis”和“fea”的 F 值均低于 0.35。究其原因，一是与不同情绪类别的情绪词典和情绪规则建设水平不一致相关；二是不同情绪类别的情绪强烈度不同，表达情绪手段的多寡性及复杂度也不一样。“dis”“fea”以及“sad”是相对温和的情绪，有一个较宽的中间地带，所以识别难度也比较大。此外，非情绪表达的微博识别效果最好，其 F 值高于 0.65，表明基于情绪因子的微博情绪识别和计算方法在识别非情绪表达时也是有效的。

4.2　微博测试语料实验及结果

4.2.1　测试语料

测试语料来自 2013 年 CCF 第二届自然语言处理与中文计算会议中文微博情绪分析评测提供的微博测试语料，共 50000 条微博，175911 个微博句子。

评测共分为 3 个子任务，即微博情绪判断(判断微博中是否含有情绪表达)、微博情绪识别(判断微博整体情绪类别)以及对情绪句情绪的识别(识别句子的主要情绪和次要情绪)。该评测根据资源受限情况分为 close 和 open 两类，针对 3 个子任务分别使用正确率、召回率和 F 值，宏平均和微平均的准确率、召回率、F 值以及严格和宽松指标的平均精度来评价各个参赛队伍提交结果的性能。

表 2　微博情绪识别的实验结果

Table 2　Recognition of emotion in microblog

类别		正确率	召回率	F 值
情绪表达	hap	0.4562	0.5472	0.4976
	lik	0.3562	0.5477	0.4317
	dis	0.4108	0.2329	0.2973
	ang	0.5696	0.5574	0.5634
	sad	0.4489	0.3402	0.3871
	fea	0.25	0.4286	0.3158
	sur	0.4107	0.4107	0.4107
非情绪表达		0.6849	0.6297	0.6561

4.2.2　评测结果

CUCeas 参加该评测任务的全部 3 个子任务的 open 测试，并取得较好的成绩。表 3~5 是 3 个子任务中 CUCeas 系统的成绩以及参赛所有单位的中位成绩及最好成绩。

4.2.3　评测结果分析

CUCeas 系统在此次评测的 open 测试中，微博情绪判断任务获得第 2 名(共 9 支队伍参加)，微博情绪识别任务获得第 1 名(共 9 支队伍参加)，情绪句情绪识别任务获得第 3 名(共 5 支队伍参加)，并且 3 个子任务的成绩均高于中位成绩。看来，针对中文微博语料，通过对微博语料情绪识别中有效特征的细致分析，基于情绪因子的微博情绪识别和计算方

表 3　微博情绪判断任务结果

Table 3　Judgment of emotions in microblog

单位	正确率	召回率	F 值
CUCeas	0.6363	0.7616	0.6933
中位成绩	0.6549	0.7044	0.6788
最好成绩	0.6413	0.8435	0.7286

表 4　微博情绪识别任务结果

Table 4　Recognition of emotions in microblog

单位	宏平均			微平均		
	正确率	召回率	F 值	正确率	召回率	F 值
CUCeas	0.2842	0.348	0.3129	0.3232	0.3868	0.3521
中位成绩	0.2474	0.2528	0.2501	0.2145	0.3911	0.2771
最好成绩	0.2842	0.348	0.3129	0.3232	0.3868	0.3521

表 5　情绪句情绪识别任务结果

Table 5　Recognition of emotions in emotional sentence

单位	平均精度(宽松指标)	平均精度(严格指标)
CUCeas	0.2908	0.2818
中位成绩	0.2878	0.2809
最好成绩	0.365	0.3484

法是有效的, 并具有一定的优势。

但是目前微博情绪分析整体准确率仍比较低, 距离实用化的目标还有相当距离, 究其原因, 除了工作中的失误, 如情绪词典不够完善、规则不能覆盖全部语言现象外, 还有以下 3 个方面。

1) 情绪本身的复杂性。一些情绪之间的界限并不是十分明晰, 存在着相互交叉的现象和模糊地带。如“不爽”“受不了”体现的情绪是 ang 还是 dis? 很难划清。如果硬去划分, 难免见仁见智。

2) 语境对情绪倾向的影响。语境是影响句子情绪倾向的重要方面。首先是上下文语境, 微博中存在着一些反语的现象, 需要结合上下文语境去理解。如“我真是个幸运的孩子。”如果不结合它的上文“一早上就那么惊讶地发现掉钱了。”就会错误地将其理解为是一个 hap 情绪句。其次是社会文化语境, 背景知识和生活常识也是理解句子情绪倾向性的关键。如“一条微博要发 *n* 次, 用一个小时!”就需要结合社会文化语境去理解微博作者所表达的 ang 的情绪。在处理这些问题时, 系统明显表现出知识不足, 目前尚未找到合适的解决方法。

3) 微博语言的不规范性和随意性。微博中错别字、标点符号的滥用和误用现象比比皆是, 这些不规范现象会影响分词和规则的匹配。同时微博中还存在着大量的谐音词, 如: 10 在受不了(实在受不了)、有桑(忧伤)、玛德(妈的)等, 这些随意使用的谐音词给研究工作带来很大的挑战, 如果不能有效地处理这些带有情绪倾向的谐音词, 就会降低情绪句识别和判断的准确度。

5 结语

情绪因子的 5 种表现手段在中文微博情绪分析中都发挥着重要作用。情绪词、情绪短语和情绪表达式是情绪因子的基本组成部分, 也是微博情绪判断和识别的关键因素。表情符号和标点符号作为表达情绪的重要手段, 对微博情绪的识别、分类起到重要的辅助作用。测试结果表明, 以情绪因子为基础的情绪识别方法取得了较好的效果, 具有可行性, 但也存在一定问题, 需要进一步解决、完善和提高。例如, 语用因素也应是一种情绪因子, 社会文化常识在情绪识别中也会起到相当的作用, 但目前还找不到一种合适的形式化描述方法, 这将是今后的努力方向。

致谢 中国传媒大学何伟老师、邹煜老师对本文提出了宝贵的修改建议, 在此表示衷心的感谢。

参考文献

[1] 孟昭兰. 情绪心理学. 北京: 北京大学出版社, 2005

[2] Campos J. Sicioemotional Development // Mussen P. Handbook of Child Development, 1983: 4th. Vol.2

[3] Arnold M. Emotion and Personality. Columbia University press, 1960

[4] Lazarus R, Folkman S. Stress, appraisal, and coping. New York: Springer, 1984

[5] 林传鼎. 社会主义心理学中的情绪问题. 社会心理科学, 2006, 21(1): 37, 62

[6] Aman S, Szpakowiczm S. Identifying expressions of emotion in text. Lecture Notes in Computer Science, 2007, 4629: 196–205

[7] Quan Changqin, Ren Fuji. Sentence emotion analysis and recognition based on emotion words using Ren-CECps . International Journal of Advanced Intelligence, 2010, 2(1): 105–117

[8] 许小颖, 陶建华. 汉语情感系统中情感划分的研究 // 第一届中国情感计算及智能交互学术会议论文集. 北京, 2003: 199–205

[9] 徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造. 情报学报, 2008, 27(2): 180–185