ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 160－167
DOI:10.16337/j.1004-9037.2016.01.017

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

中文微博观点句识别及要素抽取研究

王冠群　田　雪　黄德根　张　婧

(大连理工大学计算机科学与技术学院,大连,116024)

摘　要:研究中文微博情感分析中的观点句识别及要素抽取问题。在观点句识别方面,提出了一种利用微博中的情感词和情感影响因子计算微博语义情感倾向的新算法;在观点句要素抽取方面,利用主题词分类及关联规则,辅以一系列剪枝、筛选和定界规则抽取评价对象。通过观点句识别和观点句要素抽取结果的相互过滤,进一步提高召回率。实验数据采用第六届中文倾向性分析评测所发布的数据,结果表明,本文方法在观点句识别和要素抽取方面能够取得较好的效果,观点句识别的精确率、召回率入 F 值分别为 95.62%,54.10%及 69.10%;观点句要素抽取的精确率、召回率以及 F 值分别为 22.07%,12.66%和 16.09%。

关键词:情感分析;语义情感倾向;情感影响因子;主题词分类;关联规则

中图分类号:TP391　　**文献标志码**:A

Opinion Sentence Identification and Element Extraction in Chinese Micro Blogs

Wang Guanqun, Tian Xue, Huang Degen, Zhang Jing

(College of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, China)

Abstract: The research aimes at opinion sentence identification and element extraction in sentiment analysis in Chinese micro blogs. In the aspect of opinion sentence identification, the authors propose a new algorithm to compute the micro blog semantic sentiment orientation using sentiment words and emotional impact factors. In element extraction, subject term classification and the association rule are applied, accompanied with a series of pruning, sifting and delimiting rules to extract evaluative objects in micro blogs. Through mutual filtering of opinion sentence identification and element extraction, the recall rate is improved further. The released data of the sixth Chinese opinion analysis evaluation is adopted as experimental data. The results show that the methods perform well in opinion sentence recognition and element extraction. The precision ratio, recall rate, and F-value of opinion sentence identification are 95.62%,54.10% and 69.10%, respectively. The precision ratio, recall rate, and F-value of element extraction are 22.07%,12.66% and 16.09%, respectively.

Key words: sentiment analysis; semantic sentiment orientation; emotional impact factor; subject term classification; association rule

---

基金项目:国家自然科学基金(61173100,61173101,61272375)资助项目。
收稿日期:2014-09-19;**修订日期**:2014-10-21

引 言

近年来,微博作为一种新的信息发布平台和社交平台越来越受到人们的关注。它蕴含着巨大的政治、商业价值,比如对于热点事件的消息发布可以用于突发事件的检测[1],对于产品的评论可以用于产品的市场调研及反馈等。在微博情感分析方面,目前针对英文微博的分析已经比较成熟,英文微博限制用户发布的文本不得超过 140 个字符,通常是一个包含 7～10 个单词的句子,内容相对简单;而中文微博限制用户发布的文本不得超过 140 个中文字符,通常包含多个句子,且每个句子涉及的主题和情感可能不同,这加大了中文微博情感分析的难度。中文微博情感分析主要包括情感词的识别及分类、观点句的识别及分类和观点句要素抽取等任务,本文主要研究其中的观点句识别和要素(评价对象)抽取。

观点句识别任务需要确定每条微博的情感极性,其分析方法大致可以分为两类:基于词典的分析方法[2]和基于机器学习的分析方法[3-5],其中机器学习方法又可以分为有监督的学习方法[3-4]和无监督的学习方法[5]。基于词典的分析方法内容包括:构建含有正向情感词和负向情感词的情感词典,利用情感词典确定文本的情感倾向。文献[2]通过构建多个词典,包括否定词词典、程度词词典以及感叹词词典来综合计算每条微博的情感指数。基于词典的方法重点在于情感词典的构建,其对实验结果产生直接影响,该方法的缺陷在于无法解决未登录词问题。有监督的机器学习方法,包括朴素贝叶斯 NB、最大熵 ME 和支持向量机 SVM 等。文献[3]分别比较了多种分类算法以及各种特征和特征权值选择策略在基于监督学习的情感分类中的效果;文献[4]提出了一种基于 SVM 的层次结构多策略方法,融合表情符号、情感词典以及上下文等多种特征对中文微博进行情感分类,实验准确率达到 67.283%。无监督的机器学习方法利用非标注样本建模,通常使用标注的种子词集来实现无监督分类。文献[5]选取"excellent"和"poor"作为正负向基准情感词,得到每个单词与基准词之间的点互信息后,通过计算它们的差值得到单词的情感倾向性,文献中使用多类英文评论作为语料进行实验,平均正确率达到 74.39%。由于有监督的机器学习方法需要利用已标注的语料,从而耗费大量人力,无监督的机器学习方法精度较低,因此在观点句识别方面本文采用基于词典的分析方法,抽取微博中的情感词、程度副词、否定副词、连词和标点,通过分析它们的语义强度为其赋予权值,用来计算整条微博的语义情感倾向值,并根据倾向值的正负和大小确定观点句。

观点句要素识别不仅需要抽取微博中的评价对象,还要确定其情感倾向。基本方法有基于无监督学习的抽取方法[6-7]和基于有监督学习的抽取方法[8-9]。基于无监督学习的抽取方法有:文献[6]基于关联规则,使用词频和词的位置信息等特征构建启发式规则,实现对英文微博评价对象的抽取,F 值为 75.79%;文献[7]分析微博的特点,提出了一种基于 MB-LDA 模型的微博主题挖掘算法。相对于无监督学习的抽取方法,基于有监督学习的评价对象抽取方法起步较晚。文献[8]将评价对象抽取问题建模成序列标注问题,使用 CRFs 进行学习,获得了很好的抽取效果。文献[9]提出了一种基于浅层语义分析的评价对象抽取方法,将情感描述单元作为谓词,对应的评价对象作为语义角色,将评价对象抽取问题转化为语义角色识别问题,该方法充分利用了句法知识,在文中的多组实验中,F 值均在 55%以上。由于有监督学习的训练语料可能因标注不一致影响结果的精确率,因此本文选择无监督的学习方法。首先利用微博中的主题词进行分类,然后采用关联规则中的 Apriori 算法[10]对各类微博分别抽取其中的名词性频集,最后利用一系列剪枝、定界和筛选规则完成评价对象的抽取。

1 算法模型

1.1 情感词典的建立

考虑到有些情感词需要依赖其上下文才能表达具体的情感倾向,将情感词分为两类,即独立情感词

典和上下文相关情感词典，分别建立情感词典。(1)独立情感词典：其中的词语均能独立表达情感倾向，例如词语“给力”。使用大连理工大学信息检索研究室的情感词汇本体[11]和台湾大学中文通用情感词典。(2)上下文相关情感词典：其中的词语本身并不具有情感倾向，但在特定搭配中显示情感倾向，例如特定搭配“不禁摔”，“不禁”和“摔”这两个词本身不具有情感倾向，但此搭配含有负向的情感倾向，将同样作为情感词处理。从语料中抽取高频＜名词，形容词＞和＜副词，动词＞组合，然后对其进行人工筛选和极性标注。

1.2 观点句识别

观点句识别即判断每一条微博的情感极性。情感词对语义情感倾向起决定性作用，同时程度副词、否定副词、连词和标点会对语义情感倾向产生影响，本文将四者统称为情感影响因子，并将其按作用范围划分为词语级、子句级和句子级：词语级情感影响因子作用于情感词，包括程度副词和否定副词；子句级情感影响因子作用于子句，即连词；句子级情感影响因子作用于整个句子，即标点。在观点句识别之前，首先要对语料进行预处理，主要包括：(1)去噪：去掉测试语料中的句子编号、URL、@信息，提取微博正文。(2)分词和词性标注：采用本实验室 NiHao 分词工具[12]进行分词和词性标注。

1.2.1 情感影响因子的抽取

情感影响因子的抽取包含如下几个方面：

(1)词语级情感影响因子的抽取

情感词附近(窗口为 8)出现的程度副词和否定副词被认为是该情感词的词语级情感影响因子，其中程度副词根据语气强弱分为 3 个等级。部分词语级情感影响因子如表 1 所示。

(2)句子级情感影响因子的抽取

对一条微博按句切分，得到若干句子，各句句末标点被认为是句子级情感影响因子。对表示加强语气的句末标点(如“!”等)设定权重为 1.5，对表示不确定语气的句末标点(如“?”等)设定权重为 0，其他标点权重默认为 1。

(3)子句级情感影响因子的抽取

一些频繁出现的连词常表现出对后面内容的情感有削弱或加强功能，例如：“这部电影不好看，虽然场景挺华丽，但是剧情很无聊”。在这个句子中连词“虽然”引导的子句情感被削弱，整个句子的情感更倾向于“但是”引导的子句情感。因此，将句子按连词切分，连词是切分后相应子句的子句级情感影响因子。经过对分词词典中连词的人工筛选，得到情感削弱型连词 22 个，情感加强型连词 44 个。部分词语级情感影响因子如表 2 所示。权值的设定是根据文献[13]对情感影响因子的分级，分析它们在微博中的语义情感强度得到的，实验证明这种设定方法可以获得较好的效果。

表 1 部分词语级情感影响因子

Tab. 1 Several word level emotional impact factors

类 别	词 语	权 值
程度副词	极其、尤其、亘古、最、相当…	1.5
	很、更、好、特、非常…	1.2
否定副词	稍微、挺、极少、较、不大…	0.8
	未、不、不要、毫不、无、没有…	−0.8

表 2 部分子句级情感影响因子

Tab. 2 Several clause level emotional impact factors

类 别	词 语	权 值
情感加强型	但是、总之、综上所述、甚至、进而、果然…	1.5
情感减弱型	虽然、如果、即使、尽管、倘若、纵使…	0.8

1.2.2 语义情感倾向算法

语义情感倾向算法(Sentiment orientation algorithm，SO)的基本思想是利用句子中的情感词和情感影响因子判断句子的情感倾向。算法步骤如下：

(1)根据情感词极性 P_i 和词语级情感影响因子权重 AW_r 计算词语级语义情感倾向 SO_W_i

$$SO\_W_i = P_i \times \prod_{r=1}^{Ma} AW_r \tag{1}$$

式中:正向情感词的极性 $P_i=1$,负向情感词的极性 $P_i=-1$,Ma 表示修饰该情感词的所有词语级情感影响因子的总数。

(2)根据已计算出的词语级语义情感倾向 SO_W_i 和连词权重 CW_j 计算子句级语义情感倾向 SO_C_j

$$SO\_C_j = CW_j \times \sum_{i=1}^{Mw} SO\_W_i \tag{2}$$

式中:Mw 表示该连词作用的子句中情感词的总数。

(3)根据已计算出的子句级语义情感倾向 SO_C_j 和句末标点权重 PW_k 计算句子级语义情感倾向 SO_S_k

$$SO\_S_k = PW_k \times \sum_{j=1}^{Mc} SO\_C_j \tag{3}$$

式中:Mc 表示该句子中的子句总数。

(4)根据已计算出的句子级语义情感倾向 SO_S_k 和微博中情感词在全部词语中所占比例计算微博的语义情感倾向值 SO

$$SO = \frac{n}{N} \times \sum_{k=1}^{Ms} SO\_S_k \tag{4}$$

式中:n 和 N 分别表示微博的情感词总数和全部词语总数,Ms 表示微博中的句子总数。

由于当一条微博的词语总数一定时,情感词数量越多,这条微博带有情感倾向的可能性越大,因此将它们的比例作为语义情感倾向值的一个乘积因子。

(5)根据微博的语义情感倾向值可以得出其情感倾向。若值为正,则微博表达正向情感;若值为负,则微博表达负向情感。语义情感倾向值的绝对值大小反映情感倾向的强度。

1.3 观点句要素抽取

观点句要素抽取的主要任务是针对每一个观点句,抽取其中的评价对象并判断其倾向性。COAE2014 中文倾向性分析评测(为叙述方便,以下简称 COAE 评测)中,对于评价对象的要求仅限于产品(例如三星手机)以及产品的属性(例如质量)。

1.3.1 微博的主题词分类

由于语料具有明显的领域性,因此首先对语料按照主题词进行分类。部分主题词如表 3 所示。

表 3 部分主题词

Tab. 3 Several subject terms

类 别	主题词
汽车	奔驰、宝马、奥迪、保时捷
电子产品	三星、小米、爱疯、手机
牛奶	光明、伊利、蒙牛、牛奶
银行	建行、招行、银行、信用卡
保险	人寿、保险、社保、人保
翡翠	翡翠、挂件、手镯、戒指

(1)主题词抽取。抽取语料中出现频率较高的名词性成分,词性为 COM-NOUN(普通名词)、ELSE-PRONOUN(其他专名)和 ORG(机构名),并对抽取的结果进行人工筛选和类别标注,得到一个主题词表,词表中共含有 54 个主题词。

(2)根据主题词进行分类。统计每条微博中含有每类主题词的个数,个数最多的类别即为该微博所在类别。若一条微博中不含主题词,则其不参与后续的观点句要素抽取任务。

1.3.2 产品的抽取

(1)关联规则挖掘代表产品的频集

关联规则挖掘使用 Apriori 算法挖掘数据中蕴含的规则,Apriori 算法分为两个步骤:(a)寻找数据中的所有频繁项集,项集中的项目数不定。(b)从频繁项集中寻找生成规则。仅使用 Apriori 算法的第一步,设置最小支持度为所有微博数量的 0.5%,针对微博中的名词性成分,获取每条微博的 k-频集,每个频集代表一个产品,其可能由一个词语构成,也可能是几个词语构成的短语。鉴于产品名不可能过长,因此将 k 的最大值设为 3。为了提高效率,对 Apriori 算法进行了改进。在寻找 k-候选项集的过程中并非按照 Apriori 算法中的描述,即对所有(k-1)-频集进行自连接,而是仅对在同一条微博中的(k-1)-频集进行自连接,这样大大减少了无用 k-候选频集的数量。

(2)频集剪枝

(a)紧密度剪枝。针对 k- 频集($k \geqslant 2$),剪掉无法形成产品的频集。能够形成产品的频集满足的条件如下:频集中每两个词之间不超过 3 个词,且以这样的形式至少出现在 5 条微博中;频集中每两个词之间的词语词性不能是介词等不能存在于名词短语中的词性。

(b)冗余度剪枝。针对 k- 频集($k \leqslant 2$),剪掉冗余频集。非冗余频集满足的条件如下:频集独立出现在至少 3 条微博中。所谓独立出现,即微博中不存在该频集的超集;若某 1-频集中的词语词性为 COM-NOUN,其在整个语料中出现的频率至少是所有微博数量的 5%。这个规则主要为了删除在语料中出现频率较大的干扰名词。

(3)产品的筛选和定界

经过剪枝后,一条微博可能存在多个代表产品的频集,这时需要通过筛选得到最有可能成为产品的频集,并通过定界得到最终的产品。

(a)筛选规则

一条微博中频集数量超过 8 个则认为是干扰微博,分析时不考虑;多元频集优先:多元频集成为产品名的概率更高,所以优先考虑一条微博中项目最多的频集;词性优先:由于 ELSE-PRONOUN,ORG 这两个词性构成产品名的可能性更大,因此优先考虑含有这两个词性的频集。

(b)定界规则

多元频集直接连接生成产品名。例如产品名“三星双卡双待手机”,在获取频集时得到的是{“三星”,“手机”},直接连接即可生成产品名;1-频集中的词语词性为 ELSE-PRONOUN,ORG,向后结合。例如产品名“三星 note3”,在获取频集时得到的是{“三星”},结合后面的字符串得到产品名;1-频集中的词语词性为 COM-NOUN,向前结合。例如产品“冰种飘蓝花翡翠”,在获取频集时得到的是{“翡翠”},结合前面的名词或形容词得到产品名。对于没有成为产品的频集,转换为产品的候选属性,将在后续属性抽取部分使用。

1.3.3 产品的属性抽取

属性通常位于情感词附近,因此属性抽取需要用到观点句识别的中间结果,该中间结果需包含的情感词信息有:位置、所在分句的起始及结束位置、所在分句的情感倾向。属性抽取的步骤如下:(a)在情感词所在分句中获取所有名词性词语。(b)若(a)中抽取的词语属于候选属性,则直接将其作为产品的属性。(c)若(a)中抽取的词语均不属于候选属性,则将与产品距离最近的词语作为产品的属性。(d)若情感词所在分句中没有名词性词语,则认为不存在产品的属性,设为 null。

1.3.4 属性与产品的对应及情感倾向判断

抽取出产品的属性后,需要确定属性属于哪个产品并判断属性的情感倾向。(a)微博中只含有一个产品:所有抽取的属性均为该产品的属性,属性(情感词)所在分句的情感倾向即为对评价对象的情感倾向。(b)微博中含有多个产品且距离较远:属性属于与其距离最近的产品,属性(情感词)所在分句的情感倾向即为对评价对象的情感倾向。(c)微博中含有多个产品且距离较近:属性属于所有产品,通过连

词判断产品之间的关系是并列还是比较,进而确定评价对象的情感倾向。

1.4 干扰微博的过滤

对语料中的干扰微博,需要结合观点句识别及要素抽取结果对其进行过滤,需要过滤掉的微博满足以下两个条件:(1)微博中不含评价对象。(2)微博的语义情感倾向值的绝对值小于某一个相对阈值。在计算出每条微博的语义情感倾向值并抽取每条微博的评价对象后,利用要素抽取结果过滤掉不含评价对象的微博,然后对微博的语义情感倾向值的绝对值进行排序,获取其中的 Top n(n 为结果集中微博总数)作为最终的观点句识别和要素抽取结果。

2 实验设计及结果分析

2.1 实验数据集及评价标准

实验数据集和标准答案集选择 COAE 评测任务四和任务五发布的数据和答案。数据集中共有 40 000条微博消息,涉及电子产品、汽车、牛奶以及翡翠等多个类别,其中包括 7 000 条标注微博和 33 000条干扰微博,评价时只针对标注的 7 000 条微博进行评判,干扰微博不在评价的范围内。标准答案集有两个,分别是评测方使用的 7 000 条标注答案和评测方发布的 5 000 条部分答案。提交 10 000 条微博的结果,结果集仅包含编号在标准答案集中的结果,评价标准采用精确率 P、召回率 R 和 F 值。

$$P=\frac{\text{结果集中正确的结果总数}}{\text{结果集中的结果总数}} \tag{5}$$

$$R=\frac{\text{结果集中正确的结果总数}}{\text{标准答案集中的结果总数}} \tag{6}$$

$$F=\frac{2\times P\times R}{P+R} \tag{7}$$

2.2 观点句识别的实验

观点句识别共 3 组对比实验,并将实验结果与评测结果进行比较。实验 1:通过微博中正向和负向情感词数目的差值确定微博的情感倾向性,此实验作为实验的基线系统。实验 2:使用语义情感倾向(SO)算法确定微博的情感倾向性。实验 3:在方法 2 的基础上根据观点句要素抽取结果过滤干扰微博。观点句识别实验结果如表 4 所示。评价时标准答案集采用 5 000 条部分答案。提交结果使用实验 3 的方法,仅使用独立词典识别情感词。观点句识别结果比较如表 5 所示。评价时标准答案集采用7 000条标注答案。

表 4 观点句识别实验结果

Tab. 4 Results of opinion sentence identification %

实验方法	精确率 P	召回率 R	F 值
实验 1	95.86	32.44	48.48
实验 2	96.83	46.54	62.86
实验 3	95.62	54.10	69.10
提交结果	94.60	52.10	67.10

表 5 观点句识别结果比较

Tab. 5 Result comparison of opinion sentence identification %

实验方法	精确率 P	召回率 R	F 值
提交结果	90.2	41.9	57.2
最好	96.2	54.7	68.1
中等	85.7	30.5	45.0

实验的结果分析如下:

(1)通过比较实验 1 和实验 2 的结果可知,语义情感倾向算法对观点句识别的 F 值提升较大。SO 算法的主要优势在于全面分析微博中情感词和情感影响因子对情感倾向性的影响,降低了正负情感词相互抵消的可能性;同时,利用微博中情感词在全部词语中所占比例修正的语义情感倾向值具有很好的区分度,通过它的大小能够判断微博的语义情感强度。

(2)通过比较实验 2 和实验 3 的结果可知,根据观点句要素抽取结果过滤干扰微博,对结果的召回

率提升较大,从而使 F 值上升了 6.24%。但由于要素抽取结果存在较多错误,在一定程度上影响了观点句识别的精确率。

(3)提交结果的 F 值较中位结果高出 12.2%,可知 SO 算法在观点句识别方面存在优势;同时,提交结果与最好结果还是存在一定的差距,主要是由于带有模糊情感倾向的微博没有识别出来,而对于一些带有反讽和反问语气的微博也无法做到准确识别。

2.3 观点句要素抽取的实验

观点句要素抽取共 3 组对比实验,并将实验结果与评测结果进行比较。实验 1:不进行主题词分类,针对所有微博利用关联规则进行观点句要素抽取,在所有能抽取到评价对象的微博中随机选取 10 000条作为答案。实验 2:使用实验 1 进行观点句要素抽取,根据观点句识别实验 2 的结果选择语义情感倾向值较高的 10 000 条微博作为答案。实验 3:在实验 2 的基础上预先对所有微博根据主题词进行分类,然后对每个类别分别使用关联规则进行观点句要素抽取,针对不同类别,使用不尽相同的规则进行抽取。结果如表 6 所示。评价时标准答案集采用 5 000 条部分答案。观点句要素抽取结果比较如表 7 所示。评价时标准答案集采用 7 000 条标准答案,提交结果为实验 2 结果。

表 6 观点句要素抽取实验结果

Tab. 5 Results of element extraction %

实验方法	精确率 P	召回率 R	F 值
实验 1	19.14	8.75	12.01
实验 2	19.58	11.84	14.76
实验 3	22.07	12.66	16.09

表 7 观点句要素抽取结果比较

Tab. 7 Result Comparison of element extraction %

实验方法	精确率 P	召回率 R	F 值
提交结果(实验 2)	20.1	8.9	12.3
最好	44.1	17.7	23.9
中等	17.4	7.0	9.7

实验的结果分析如下:

(1)通过比较实验 1,2 的结果可知,根据语义情感倾向值选取观点句要素抽取结果可以有效提高结果的召回率。结合观点句识别实验 2 和实验 3 的结果,可以得出观点句识别和要素抽取具有较强的相关性,即在一般情况下,语义情感倾向值高的微博通常含有评价对象,同时含有评价对象的微博通常都是观点句。因此二者的相互过滤对彼此的结果都具有积极的影响。

(2)通过比较实验 2,3 的结果可知,使用主题词分类可以提高观点句要素抽取的精确率和召回率。通过主题词分类,相似的微博能够集中在一起进行处理,关联规则的最小支持度在获取频集时可以发挥更大的作用,并且分类后,可以针对不同类别的微博,提供不太相同的定界规则,进一步提升结果的正确率。同时,对微博进行分类后,由于每次关联规则挖掘的微博总数以及获取的频集总数都下降了一个数量级,因此评价对象抽取时间降低到原来的 1/10 以下。

(3)提交结果的 P,R,F 值仅为最好结果的一半,说明方法在观点句要素抽取方面整体性能不高,分析其原因主要有两个方面,一是只抽取了高频产品,没有对低频产品进行有效抽取;二是在方法中只单纯使用词性、词频和词的位置信息,而没有用到句法知识,导致在产品和属性的对应关系方面以及对评价对象的倾向性判断方面产生偏差。

3 结束语

本文主要研究中文微博情感分析中的观点句识别及要素抽取,提出了一种新的微博语义情感倾向算法,在观点句识别方面有较好的表现;同时提出了一种基于关联规则的评价对象抽取方法,方法中使用了主题词分类及一系列规则,使 F 值有明显提高。本文的方法存在很多问题,在如下两个方面还有待深入研究:(1)构建反讽和反问触发词库,用于判断带有反讽或反问语气的微博的情感倾向性。(2)考虑低频产品的处理方法以及如何利用句法知识过滤不正确的产品和属性组合。

参考文献:

[1] 杨亮,林原,林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报,2012,26(1):84-90.
Yang Liang, Lin Yuan, Lin Hongfei. Hot event discovery in micro-blog based on emotional distribution[J]. Journal of Chinese Information Processing, 2012,26(1):84-90.

[2] Shen Yang, Li Shuchen, Zheng Ling, et al. Emotion mining research on micro-blog[C]// Web Society,SWS'09,1st IEEE Symposium on. Lanzhou, China:IEEE, 2009:71-75.

[3] Pang B, Lee L, Vaithyanathan S. Thumbs up:Sentiment classification using machine learning techniques[C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10. Stroudsburg, USA:Association for Computational Linguistics, 2002:79-86.

[4] 谢丽星,周明,孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报,2012,26(1):73-83.
Xie Lixing, Zhou Ming, Sun Maosong. Hierarchical structure based hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction[J]. Journal of Chinese Information Processing, 2012, 26(1):73-83.

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

[6] Hu M, Liu B. Mining and summarizing customer reviews[C]// Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2004:168-177.

[7] 张晨逸,孙建伶,丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展,2011,48(10):1795-1802.
Zhang Chenyi, Sun Jianling, Ding Yiqun. Topic mining for microblog based on MB-LDA model[J]. Journal of Computer Research and Development, 2011, 48(10):1795-1802.

[8] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA:Association for Computational Linguistics, 2010:1035-1045.

[9] Li S, Wang R, Zhou G. Opinion target extraction using a shallow semantic parsing framework[C]// Twenty-Sixth AAAI Conference on Artificial Intelligence. Toronto, Canada:AAAI, 2012:1671-1677.

[10] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]// Proc 20th Int Confon Very Large Databases. Santiago, Chile:VLDB, 1994, 1215:487-499.

[11] 徐琳宏,林鸿飞,潘宇,等. 情感词汇本体的构造[J]. 情报学报,2008,27(2):180-185.
Xu Linhong, Lin Hongfei, Pan Yu, et al. Constructing the affective lexicon ontology[J]. Journal of the China Society for Scientific and Technical Information,2008,27(2):180-185

[12] Degen H, Deqin T. Context information and fragments based cross-domain word segmentation[J]. China Communications, 2012, 9(3):49-57.

[13] 张桂宾. 相对程度副词与绝对程度副词[J]. 华东师范大学学报:哲学社会科学版,1997 (2):92-96.
Zhang Guibin. The relative degree adverbs and the absolute degree adverbs[J]. Journal of East China Normal University:Philosophy and Social Sciences, 1997 (2):92-96.

作者简介:

王冠群(1990-),女,硕士研究生,研究方向:自然语言处理与机器翻译,E-mail:esythan@mail.dlut.edu.cn。

田雪(1991-),女,硕士研究生,研究方向:自然语言处理与机器翻译。

黄德根(1965-),男,教授,博导,研究方向:自然语言处理与机器翻译。

张婧(1987-),女,博士研究生,研究方向:自然语言处理与机器翻译。