基于“属性-情感词”汽车本体的文本情感分析

王连喜[1,2]
(1. 广州市非通用语种智能处理重点实验室,广东 广州 510006;
2. 广东外语外贸大学 信息科学与技术学院,广东 广州 510006)

摘 要:对特定领域网络评论进行情感分析,可以帮助商家更深入地了解用户需求、总结自身产品和服务的优势与不足,也可以帮助消费者了解特定领域产品各方面性能的评价分布,从而优化其消费决策。提出一种面向汽车领域的“属性-情感词”本体构建流程,并在此基础上提出基于“属性-情感词”本体的汽车评论文本观点句情感分析方法。该方法以观点句识别方法为基础,利用“属性-情感词”本体对汽车领域产品的八个维度(属性)进行情感分析,并与经典的朴素贝叶斯情感分类方法进行实验对比。结果表明提出的方法能有效提高属性层面上的情感分析准确率和召回率。但由于汽车领域的细粒度情感分析效果会受到“属性-情感词”本体的完善程度及相关规则的影响,因此需进一步完善“属性-情感词”本体,并构建更为全面的规则。

关键词:汽车评论;网络口碑;属性-情感词本体;观点句识别;情感分析
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2020)08-0193-06
doi:10.3969/j.issn.1673-629X.2020.08.034

Sentiment Analysis Method Based on Attribute-sentiment Ontology in Automobile Domain

WANG Lian-xi[1,2]
(1. Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangzhou 510006, China;
2. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: Sentiment analysis of reviews in specific areas can help businesses to better understand user needs, summarize the advantages and disadvantages of their own products and services, and also help consumers understand the distribution of performance of products in specific areas, thus optimizing their consumption decisions. We propose an ontology construction flow of “attribute-sentiment words” oriented to the automotive field, and propose sentiment analysis methods based on “attribute-sentiment words” ontology. Based on the viewpoint recognition method, this method uses the “attribute-sentiment word” ontology to carry out the sentiment analysis of the eight dimensions (attributes) of the automotive products and compares it with the classical Naive Bayesian sentiment classification method. It is showed that this method can effectively improve the accuracy and recall rate at the attribute level. Since fine-grained sentiment analysis in the field of automobile is affected by the perfection of “attribute-sentiment word” ontology and relevant rules, it is necessary to further improve the ontology of “attribute-sentiment word” and construct more comprehensive rules.

Key words: car reviews; Internet word-of-mouth; attribute-sentiment ontology; opinion sentence recognition; sentiment analysis

0 引 言

随着社会媒体与电子商务技术的快速发展与普及,普通民众已经习惯在网络发布和获取信息。据中国互联网络信息中心(CNNIC)发布的第 43 次《中国互联网络发展状况统计报告》显示,2018 年中国数字经济以电子商务为先导力量获得迅速发展,引领数字产业崛起和产业数字化转型[1]。特别在在线汽车网络评论领域,互联网用户创造了大量蕴含情感色彩的 UGC,准确地对用户主动生产的口碑评论信息进行挖掘和分析,可以有效帮助消费者了解汽车产品各方面性能的评价分布,从而优化其消费决策,同时也可以帮助商家了解用户需求和理解用户消费习惯,总结自身产品和服务的优势与不足。

由于网络上的产品评论大多是以半结构化形式表示的,缺乏对数据本身的描述,也没有规范性的结构,甚至有些评论的情感词在与不同产品属性进行组合时

收稿日期:2019-09-29 **修回日期**:2020-01-20
基金项目:广东省科技计划项目(2015A030401093);国家社会科学基金青年项目(17CTQ045)
作者简介:王连喜(1985-),男,副研究员,博士,硕导,CCF 会员(14733M),研究方向为涉华网络舆情分析、数据挖掘。

会表达出不同的情感倾向。例如,“这款车的油耗高”与“这款车的性价比高”两个评论中都存在情感词“高”,但是前者的情感倾向性是消极的,而后者是积极的。上述问题会对网民获取、利用、分析 UGC 带来一定的困难,因此亟需高效、有用的方法对特定领域的情感词、产品属性对象以及它们组合所蕴含的情感倾向性进行准确识别。

基于此,文中以汽车评论文本为研究对象,通过构造面向汽车领域的“属性-情感词”本体,提出基于“属性-情感词”本体的观点句情感分析方法,以期准确识别出情感词与不同产品属性对象组合所表达的情感倾向,从而提高汽车产品的细粒度情感分析效果。

1 相关研究

近年来,本体技术[2-3]已经被广泛应用于评论情感分析研究中。许多学者通过构建通用型情感词汇本体或情感词典来辅助情感分析研究,也有部分学者尝试结合领域本体技术与产品特征来提高特定领域评论情感分析的准确性。2008 年,徐琳宏等人通过整理和标注了多种词典和语义资源,构建了中文情感词汇本体库[4]。该情感词汇本体由三元组来描述,并通过计算情感词汇与给定的 20 类标准词汇在语料中的互信息来确定情感强度和极性。该情感词汇本体已成为目前被广大研究人员借鉴或使用最多的工具。Lau 等人[5]提出一种用模糊领域本体的实例化版本来表示情感知识,重点关注领域特征、领域情感词以及它们之间的对应关系抽取,能够较好地应用于上下文敏感的意见挖掘。郭冲等人[6]定义了一种用于细粒度意见挖掘的情感本体树结构,并结合细粒度意见要素抽取技术提出基于本体树的自动构建方法。

在领域本体构建及产品舆情分析方面,目前也产生了许多有价值的成果。杜嘉忠等人[7]提出了一种基于领域专用情感词的网络评论情感分析方法,该方法通过构建并利用特征-情感词本体对网络上的产品评论进行情感分析。王晓东等人[8]在现有情感词汇本体的基础上,结合规则集和词类组合模型提出了一种基于语料库的情感词汇本体扩展算法。刘丽珍等人[9]构建了产品领域情感本体,并利用领域情感本体的先验情感知识消除情感词的领域依赖性,有效识别了暗含的产品特征,能够提高在线产品评论情感分析的性能。唐晓波等人[10]以情感词典为基础,根据手机产品特征及其评论特点,构建了手机产品领域的本体,并实现了手机产品特征的抽取、分类与情感分析。尹裴等人[11]从特征词与观点词的语义关系入手,根据领域本体判断特征观点对的极性,并通过加权平均方法计算整个产品的极性。郑丽娟等人[12]结合基于语义和基于统计的方法,通过抽取特征观点对和观点词情感判断,构建相应的情感本体,提出了一种基于情感本体的在线评论情感极性及强度分析方法。何有世等人[13]通过构建手机产品领域本体实现了产品属性的提取与层次划分,并提出了基于领域本体的产品网络口碑信息多层次细粒度情感挖掘方法。

以上研究都偏向于用逻辑推理和情感计算的方法实现产品评论领域本体构建。相对其他方法,领域本体对于特定领域的网络舆情分析、属性词提取和观点抽取等内容更具专业性和针对性。除基于领域本体的情感分析方法外,还有基于情感知识的情感分析方法和基于机器学习的情感分析方法。

基于情感知识的方法通常使用一些已有的各类情感词典、领域词典以及主观文本的情感极性组合评价单元对主观文本的极性进行计算[14-17]。常用的知识有 WordNet、情感属性、位置属性、关键词属性、词性搭配关系等。尽管这一类方法可以较为充分地利用文本情感的先验知识,能够较好地解决规范性文本的情感分析问题,但由于忽视了文本分布的信息,所以容易出现经验偏置,难以解决新兴语言表达以及隐式表达的形式。

基于机器学习的方法一般是先采用机器学习方法对文本特征进行识别、提取和选择,然后构建相应模型完成相关情感分析任务。Pang 等[18]将 n-gram 词语和词性作为特征,分别采用朴素贝叶斯、最大熵和支持向量机等机器学习方法来解决文档级情感分类的问题。基于机器学习的方法由于能充分利用文本特征的分布信息,对规范化和非规范化的文本都能有效处理,但容易忽略与情感相关的先验语义特征,所以其分类性能仍存在较大提升空间。陈炳丰等人[19]通过构建汽车情感词典,提出了基于条件随机场模型的情感实体识别和情感倾向分类方法,结果表明该方法能够应用于汽车领域的网络舆情分析。

综上所述,对于领域依赖性和属性关联性的产品网络舆情分析研究来说,如果能将描述产品属性和情感倾向的词汇进行结合和映射,这样或许能得到更准确的属性评论倾向。基于此,文中针对汽车领域评论文本的网络舆情分析,提出采用基于规则的方法构建“属性-情感词”本体,并以此识别汽车属性及关于属性的评论倾向,然后将该方法与观点句识相结合实现汽车领域的网络口碑信息的情感分析。

2 基于“属性-情感词”本体的汽车领域口碑情感分析方法

汽车领域网络舆情分析是一个非常复杂的文本信息处理和建模的过程,在这个过程中不仅要构建领域

词典或本体,还需要借助机器学习方法构建相关的情感分析模型。在进行情感分析之前,首先需要获取网络论坛中的汽车产品评论,同时需要借助外部数据源收集并提取有关于汽车的属性和专有名词,然后利用数据预处理方法识别和提取评论中的属性词和情感词,并提出基于四元组的“属性-情感词”本体构建方法,最后在上述过程的基础上结合观点句识别方法提出基于“属性-情感词”本体的情感分析方法。具体实现过程如图 1 所示。

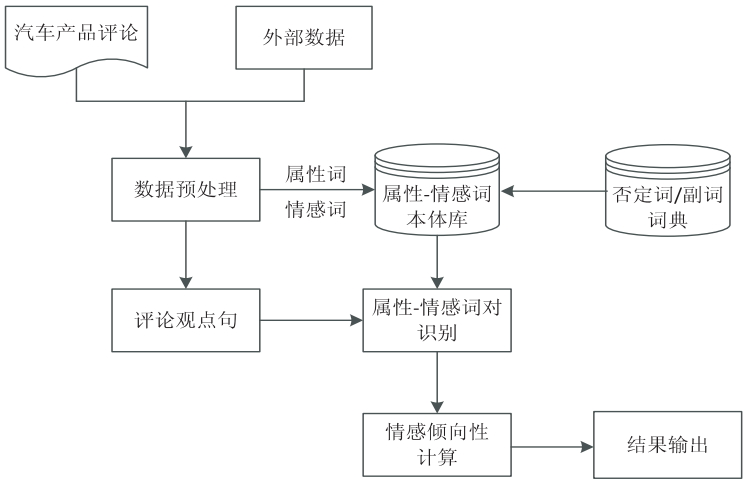

图 1　基于“属性-情感词”本体的汽车领域文本情感分析框架

如图 1 所示,提出的基于“属性-情感词”本体的汽车领域网络舆情分析方法主要包括三个过程:基于“属性-情感词”的本体构建、观点句识别以及情感分析。

2.1　基于四元组表示的“属性-情感词”本体构建

文中构建的汽车领域“属性-情感词”本体是一个包含汽车属性、情感词以及情感极性的知识模型,可以将其定义为一个四元组,即:$O = \{C,N,S,\mathrm{pol}(N,S)\}$,其中,$C$ 表示汽车属性类别,如“性价比”、“油耗”等,N 表示汽车属性关键词,如“质量”、“价格”等,S 表示情感词,如“上乘”、“宽敞”等,$\mathrm{pol}(N,S)$ 表示属性关键词-情感词对的极性,如“1”表示正向,“-1”表示负向。由该定义可知,“属性-情感词”本体可用于识别相同情感词与不同产品属性对象组合所表达出的情感极性。在具体实现过程中,可采用基于规则的方法构建“属性-情感词”本体方法(如图 2 所示)。

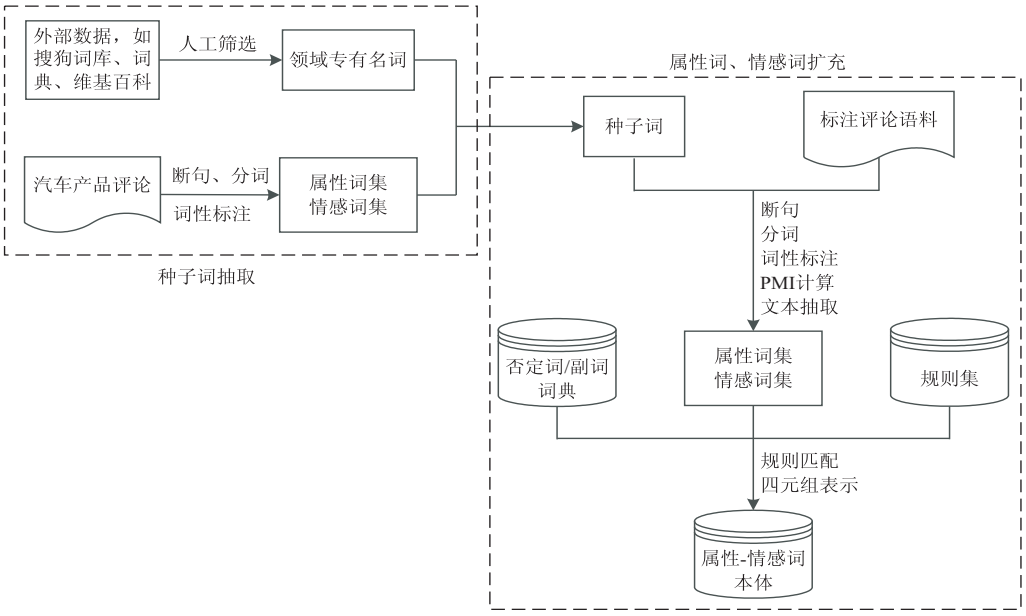

图 2　汽车领域“属性-情感词”本体构建流程

由图 2 可知,汽车领域的“属性-情感词”本体构建过程主要包括两大模块:一是种子词抽取,二是属性词、情感词扩充。在种子词的识别和抽取方面,一方面通过从企业官方网站、搜狗词库、微博、汽车门户网站等相关资源中获取汽车名称术语及部件术语,构建汽车专有名词本体库;另一方面,从汽车之家论坛中采集已对操控、空间、动力、内饰、舒适性、外观、性价比、油耗等八个方面进行评论的规范语料,并通过断句、分

词、词性标注、词频统计、文本抽取等处理过程形成属性词和情感词的种子词。在属性词和情感词的扩充方面,首先选取一定数量的正负向语料作为训练集,然后对训练语料进行预处理(断句、分词、词性标注、PMI 计算、文本抽取),并结合否定词/副词词典及相关规则匹配属性关键词-形容词对,最终在进行四元组表示的基础上形成汽车领域的“属性-情感词”本体。

表 1 属性关键词-情感词对匹配规则集

序号	类型	规则描述	规则表示	备注
1	词对位置关系	[n,a]	n+a	属性关键词(n、n_1、n_2)与形容词(a)之间的位置关系(词距离不超过 4)
2		[n_1+(“和”、“与”、“跟”、“同”)+n_2,a]	n_1+a 和 n_2+a	
3		[n_1+n_2,a]	n_1+a	
4		[n_1+“的”+n_2,a]	n_2+a	
5	否定词	句子存在 N 个否定词	$s*(-1)^N$	s 为词对情感极性
6	句子级情感	M 个词对情感极性之和	$\sum_{i=1}^{M} s_i$	S_i 表示第 i 个词对的情感极性

由“属性-情感词”本体的定义可知,该四元组既包含汽车属性,也包含了描述该属性的具体关键词及其情感倾向。但在识别、抽取和判断评论中的属性关键词与情感词对的极性时,需要遵循如表 1 所示的规则:如果评论中存在否定词,则根据否定词数量对属性关键词-情感词对的情感极性进行计算;如果评论中存在多个相同属性的属性关键词-情感词对,则对它们进行线性求和。最后,结合属性关键词的类别,得到“属性-情感词”本体的四元组 O = {属性类别,属性关键词,情感词,情感极性}。

2.2 汽车评论观点句识别

由于汽车领域评论语料中包含大量客观信息,这些信息并不表达用户对汽车或属性的评价。太多客观信息会增加情感分析的工作量,也会影响情感分析的结果,因此在对语料进行情感分析前,需要对语料进行主客观分类,即对评论语料进行观点句识别。

针对汽车评论的观点句识别问题,文中采用融合基于特征模板和基于 SVM 分类的观点句识别方法,其主要过程包括:特征提取和 SVM 分类器构造。在识别观点句之前,设计了如表 2 所示的特征模板,该模板包含两个一元特征和三个二元特征,用于匹配和提取评论中的有用特征。

在基于特征模板的特征提取的基础上,结合基于 SVM 分类方法构建观点句识别模型。该模型的构建步骤如下:首先,对训练语料进行断句、分词、词性标注,并根据特征模板匹配并提取出训练语料中的相关特征,同时利用向量空间模型将语料向量化;然后,利用 Libsvm 软件中的 C-SVC 模型构造 SVM 分类器;最后,利用 SVM 分类器对测试语料进行观点句识别。

表 2 特征模板

序号	类型	特征模板	规则描述	模板规则	说明
1	一元特征	名词特征	属性关键词	属性关键词	n. 表示名词; ? 表示前后两特征之间的距离; v. 表示动词
2		情感特征	情感词	情感词	
3	二元特征	名词特征+情感特征	属性关键词+情感词	n. +? +情感词	
4		情感特征+语气特征	情感词+语气词	情感词+? +语气词	
5		动词特征+程度特征	动词+程度副词	v. +? +程度副词	

2.3 情感分析

在“属性-情感词”本体构建和观点句识别的基础上,文中提出基于“属性-情感词”本体的情感分析方法。该方法主要是基于特征匹配和映射得出评论中的属性关键词-情感词对,并以“属性-情感词”本体判定句子情感倾向性,其过程如下:

输入:汽车评论语料、“属性-情感词”本体;

输出:汽车评论情感分析结果。

Step1:对语料进行断句(以句号、分号、感叹号等作为断句的依据)、分词、词性标注等预处理;

Step2:建立并利用规则对评论中情感词进行识别,同时计算评论中属性-情感词对的情感极性;

Step3:识别并提取评论中的汽车属性关键词,并利用“属性关键词-情感词”对匹配规则对属性关键词及其邻近的词语进行匹配;

Step4:若匹配成功,则提取相应的情感词并根据“属性-情感词”本体规则构建四元组;

Step5:对语料中的所有句子按属性进行情感极性累加,即对具有相同属性的四元组进行分类求和。

在情感分析过程中,如果匹配过程中出现词语情

感极性无法判定的情况,则可以通过其与对应属性关键词在训练集正负向语料中的共现频率大小来判断其情感极性。具体判断规则如下:

$$\text{pol}=\begin{cases}1, p(N_i,W_i\mid \text{pos}) > p(N_i,W_i\mid \text{neg})\\ -1, p(N_i,W_i\mid \text{pos}) < p(N_i,W_i\mid \text{neg})\end{cases}$$

其中,$p(N_i,W_i\mid \text{pos})=\dfrac{N_+(N_i,W_i)}{N_+}$,$p(N_i,W_i\mid \text{neg})=\dfrac{N_-(N_i,W_i)}{N_-}$,$W_i$ 表示待测词语,pol 表示待测词语 W_i 的情感极性,pos、neg 分别表示训练集正、负向语料,N_i 表示属性关键词,$p(N_i,W_i\mid \text{pos})$、$p(N_i,W_i\mid \text{neg})$ 分别表示待测词语 W_i 与属性关键词 N_i 在正向语料 pos 和负向语料 neg 中的共现概率,$N_+(N_i,W_i)$、$N_-(N_i,W_i)$ 分别表示在正向语料 pos 和负向语料 neg 中待测词语 W_i 与属性关键词 N_i 共同出现的评论数,N_+、N_- 分别表示正向语料 pos 和负向语料 neg 所包含的评论总数。

3　实验与结果分析

3.1　实验数据

实验中所用到的语料均来自于太平洋汽车网和汽车之家,且都是经由三名专业人员进行人工标注而成的。语料规模为 3 200 条评论句子,其中用于描述操控、空间、动力、内饰、舒适性、外观、性价比、油耗等八个属性的正负向语料各 200 条评论句子。

3.2　实验过程

文中使用 protégé 工具包,通过从企业官方网站、搜狗词库、微博、汽车门户网站等相关资源获取汽车名称术语及部件术语构建汽车评价对象本体库。将汽车产品评论分为操控、空间、动力、内饰、舒适性、外观、性价比、油耗这八个属性,分别构建了这八个属性的关键词表,然后在此基础上构建“属性-情感词”本体。下面以实验中某个句子的分析处理为例,详细说明提出的基于“属性-情感词”的情感分析过程(如图 3 所示)。

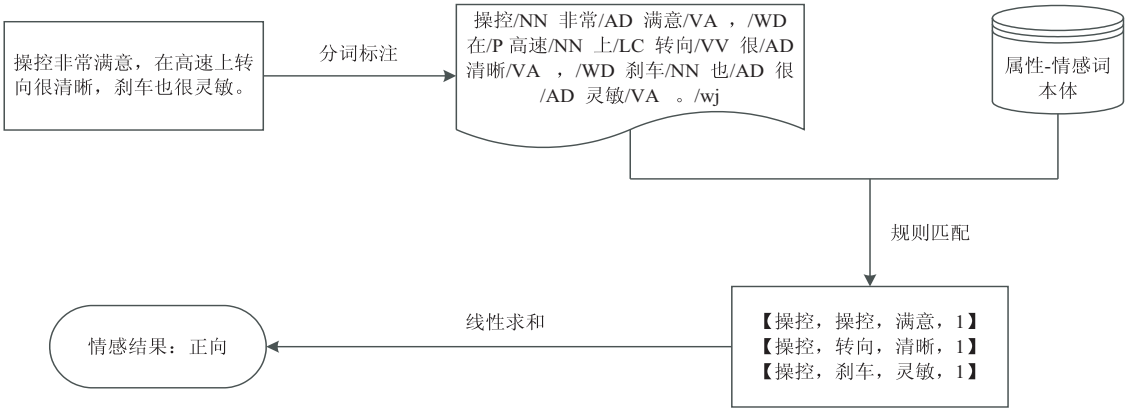

图 3　情感分析示例

3.3　评价指标

采用准确率、召回率和 F_1 值来评价情感分析方法的性能,其计算方法如下:

$$\text{准确率}=\frac{\text{算法正确判断情感倾向的句子数量}}{\text{算法能判断情感倾向的句子数量}}$$

$$\text{召回率}=\frac{\text{算法正确判断情感倾向的句子数量}}{\text{句子总数量}}$$

$$F_1=\frac{2\times \text{准确率}\times \text{召回率}}{\text{准确率}+\text{召回率}}$$

3.4　结果分析

为了验证提出方法的有效性,选择经典的朴素贝叶斯方法作为对比算法。在实验过程中,选取标注语料的三分之二作为训练集,训练出情感分类模型,剩下的三分之一作为测试集。表 3 列出了基于“属性-情感词”本体的情感分析方法和基于朴素贝叶斯的情感分析方法的实验结果。

表 3　对比实验结果　%

方法	准确率	召回率	F_1
属性-情感词本体方法	81.30	71.26	75.95
朴素贝叶斯方法	69.67	70.82	70.24

从表 3 可以看出,提出的基于“属性-情感词”本体的情感分析方法比朴素贝叶斯情感分类方法的效果更好。这是因为,朴素贝叶斯分析方法忽略了与情感相关的先验语义特征,同时也没有结合语境进行分析,即没有考虑到情感词在不同的语境表达中可能会出现不同情感的问题。

而文中方法则可以将情感词与特定的语境相结合,有效解决了情感词在描述不同属性关键词时情感倾向可能不同的问题。例如:文中方法可以正确判别“空间大”为正向情感,“车内噪音大”为负向情感。但由于构建的本体规模不够大,使用的规则不够完善,该方法在召回率方面还有待改进。

4 结束语

随着汽车行业的快速发展,不同汽车品牌的竞争日趋强烈。通过对用户使用评论的分析和利用,对汽车企业的发展和走向有重要意义。但是,在大数据时代,用户评论中存在大量噪音,使得企业对信息的获取成本大大增加。在此背景下,期望通过基于属性-情感词本体的评论情感挖掘对汽车领域产品的八大属性进行细粒度情感分析,从而给汽车企业与消费者带来一定的参考价值。

但是,该研究目前还存在很多的不足。例如,在该方法中由于依赖人工方式构建本体和情感词典构建的工作量非常大,所以属性关键词和情感词的抽取准确率仍然有待提高;该方法在处理成分残缺句子时的健壮性较差,导致评论分析的召回率比较低。

在未来的研究中,可考虑引入情感强度的计算,从而帮助解决成分残缺句子属性关键词的匹配映射以及比较要素的抽取问题。

参考文献:

[1] 中国互联网络信息中心.《第 43 次中国互联网络发展状况调查统计报告》[EB/OL]. [2019-04-17]. http://www.cac.gov.cn/wxb_pdf/0228043.pdf.

[2] 周立柱,贺宇凯,王建勇. 情感分析研究综述[J]. 计算机应用,2008,28(11):2725-2728.

[3] 任飞亮,沈继坤,孙宾宾,等. 从文本中构建领域本体技术综述[J]. 计算机学报,2019,42(3):654-676.

[4] 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.

[5] LAU R,LAI C,MA J,et al. Automatic domain ontology extraction for context-sensitive opinion mining[C]//Proceedings of the international conference on information systems. Phoenix:[s. n.],2009:35-53.

[6] 郭 冲,王振宇. 面向细粒度意见挖掘的情感本体树及自动构建[J]. 中文信息学报,2013,27(5):75-83.

[7] 杜嘉忠,徐 健,刘 颖. 网络商品评论的特征-情感词本体构建与情感分析方法研究[J]. 现代图书情报技术,2014(5):74-82.

[8] 王晓东,王 娟,张 征. 基于情感词汇本体的主观性句子倾向性计算[J]. 计算机应用,2012,32(6):1678-1681.

[9] 刘丽珍,赵新蕾,王函石,等. 基于产品特征的领域情感本体构建[J]. 北京理工大学学报,2015,35(5):538-544.

[10] 唐晓波,兰玉婷. 基于特征本体的微博产品评论情感分析[J]. 图书情报工作,2016,60(16):121-127.

[11] 尹 裴,王洪伟. 面向产品特征的中文在线评论情感分类:以本体建模为方法[J]. 系统管理学报,2016,25(1):103-114.

[12] 郑丽娟,王洪伟. 基于情感本体的在线评论情感极性及强度分析:以手机为例[J]. 管理工程学报,2017,31(2):47-54.

[13] 何有世,何述芳. 基于领域本体的产品网络口碑信息多层次细粒度情感挖掘[J]. 数据分析与知识发现,2018,2(8):60-68.

[14] DAVE K,LAWRENCE S,PENNOCK D. Mining the peanut gallery:opinion extraction and semantic classification of product reviews[C]//Proceedings of the 12th international conference on world wide web. [s. l.]:ACM,2003:519-528.

[15] KIM S,HOVY E. Determining the sentiment of opinions[C]//Proceedings of the 20th international conference on computational linguistics. Geneva:Association for Computational Linguistics,2004:1367-1373.

[16] 林 政,谭松波,程学旗. 基于情感关键句抽取的情感分类研究[J]. 计算机研究与发展,2012,49(11):2376-2382.

[17] 廖 健,王素格,李德玉,等. 基于观点袋模型的汽车评论情感极性分类[J]. 中文信息学报,2015,29(3):113-120.

[18] PANG B,LEE L,VAITHYANATHAN S. Thumbs up?:sentiment classification using machine learning techniques[C]//Proceedings of the conference on empirical methods in natural language processing (EMNLP). Philadelphia:Association for Computational Linguistics,2002:79-86.

[19] 陈炳丰,郝志峰,蔡瑞初,等. 面向汽车评论的细粒度情感分析方法研究[J]. 广东工业大学学报,2017,34(3):8-14.