计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(10):144−149 [doi: 10.15888/j.cnki.csa.005996]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

教材在线评论的情感倾向性分析①

刘若兰[1], 年　梅[1], 范祖奎[2]

[1](新疆师范大学 计算机科学技术学院, 乌鲁木齐 830054)
[2](新疆警察学院 语言系, 乌鲁木齐 830011)

摘　要: 为了充分挖掘和应用电子商务网站中的教材评论信息, 运用细粒度的情感分类算法对用户的在线评论进行分析, 基于教材特征级的情感分析结果, 辅助潜在客户和商家做出合理有效的决策. 本文首先使用爬虫采集教材的在线评论文本, 对其进行去噪、分词和词性标注等预处理; 然后分析产品特征, 在通用情感词典的基础上扩建领域情感词典; 最后基于句法分析结果, 结合教材评论的语言特性, 设计适合教材评论的情感倾向性分析算法, 并通过实验验证了算法的有效性.

关键词: 教材在线评论; 细粒度情感分析; 情感词典; 产品特征

引用格式: 刘若兰,年梅,范祖奎.教材在线评论的情感倾向性分析.计算机系统应用,2017,26(10):144–149. http://www.c-s-a.org.cn/1003-3254/5996.html

Emotional Tendency Analysis of Online Comments on Teaching Materials

LIU Ruo-Lan[1], NIAN Mei[1], FAN Zu-Kui[2]

[1](The Computer Science&Technology Department, Xinjiang Normal University, Urumqi 830054, China)
[2](The Language Department, Xinjiang Police College, Urumqi 830011, China)

Abstract: In order to fully tap and apply the information of textbook reviews on the e-commerce website, we use fine-grained emotional classification algorithm to analyze the user's online comments, based on the sentiment analysis results of product feature level, so as to assist customers and businesses to make reasonable and effective decision. In this article, we first use the crawler tool to collect online comment texts of teaching materials, and carry on some pretreatments such as denoising, segmentation and POS tagging, and then analyze the product features, based on the general emotional dictionary expands domain sentiment dictionary. Finally, based on the syntactic analysis results, combined with the language features of textbook comments, we design an affective tendency analysis algorithm which is suitable for the textbook reviews, and prove the validity of the algorithm through experiments.

Key words: the online reviews of teaching material; fine-grained emotion analysis; emotion dictionary; product features

近年来, 电子商务的迅猛发展潜移默化地改变着人们的购物方式, 网络购物已经成为众多消费者首选的购物方式, 购买商品后, 多数消费者也热衷于在网站上留下对产品或服务的真实看法或体验; 很多消费者也习惯于在购买商品前浏览已购者的评论, 从而帮助自己选择合适的产品. 因此电子商务网站上产品的在线评论信息急剧增长, 教材评论就是其中的一类. 大量的教材评论反映了消费者对教材的整体意见和态度, 具有很高的挖掘和应用价值. 一方面, 评论中表达的观点和情感可以对其他客户的购买意向产生影响; 另一方面, 便于商家对教材的质量或服务进行改进, 提高客户满意度. 但是数量庞大、纷繁复杂的评论信息如果

① 基金项目: 国家自然科学基金(61163064); 教育部人文社会科学工程科技人才培养专项(15JDGC022); 新疆师范大学数据安全重点实验室资助项目; 新疆师范大学计算机应用技术重点学科资助
收稿时间: 2017-01-08; 采用时间: 2017-02-17

不加分析, 将使用户和商家从中提取教材质量的可靠信息变得非常困难, 因此迫切需要借助数据挖掘技术识别大量用户发表的教材评论文本的情感倾向, 从中获取用户对教材的主观意见. 故本文在现有文本倾向性分析技术的基础上, 结合教材评论的特点, 设计适合教材评论的情感倾向性分析算法, 以实现对教材评论信息的挖掘和处理.

1 相关工作

通过总结在线评论情感倾向性分析的研究发现, 国内外学者分别从粗粒度和细粒度两个层面进行了研究. 粗粒度的情感分类旨在判断篇章或句子级评论文本的整体情感倾向. 但当一个评论语句对产品的多个属性进行评价时, 粗粒度的情感分析方法则无法获知用户具体喜欢或不喜欢哪些属性. 此时就需要使用细粒度情感分析算法, 识别在线评论所涉及产品各属性的情感倾向. 而产品属性会因领域的不同而发生变化, 描述不同属性的评价词也不尽相同, 因此细粒度情感分析是与领域密切相关的, 目前, 细粒度情感分析方法已被应用于汽车[1]、手机[2]、净化器[3]等领域评论数据的研究中, 采用的研究方法主要包括有监督和无监督两类方法.

有监督方法, 把细粒度的情感分析任务转化为词汇的序列标注问题, 如文献[4]将属性词和情感词的抽取视为一个序列标注任务, 利用词汇化的隐马尔科夫模型判别词汇所属的标注类别. 刘丽等人[2]则提出条件随机场 (CRF) 和语法剪枝相结合的细粒度情感分析方法.

无监督方法则是基于句法分析的方法, 如姚天昉[1]等人通过构建汽车领域本体提取主题词, 然后基于句法分析结果, 提出改进的 SBV 算法识别主题—意见词对, 最终确定语句中各主题词的情感极性. 刘鸿宇[5]等人首先借助句法分析结果抽取候选评价对象, 然后利用网络挖掘的 PMI 算法和名词剪枝算法筛选候选评价对象, 最后将情感句划分为四类, 并制定相适合的倾向性判别规则, 最终实现评价对象级的倾向性判别. 例如文献[6-8]也基于句法分析进行在线评论的情感倾向性分析研究.

教材评论信息的挖掘对于教师和学生选择合适的教材具有重要的参考作用, 同时能够为教材编写人员以及出版部门提高教材质量提供可靠建议. 但到目前为止还未见教材评论领域细粒度情感分析研究的成果. 由于细颗粒度的评论分析所挖掘的产品属性和情感词是与领域密切相关的, 因此无法直接使用其他领域或者通用领域的产品特征词以及情感词. 例如“内容”、“排版”、“纸张”等产品特征, 以及“深入浅出”, “醍醐灌顶”等情感词, 在其他领域中几乎很少见. 其他领域细粒度情感分析的资源和算法如果直接应用在教材评论分析中, 必然会影响情感分析的效果, 为此本文进行教材在线评论的情感倾向性分析研究. 在分析算法的选择方面, 考虑到有监督方法通常需要耗费大量人力标注语料, 也不利于领域切换. 而无监督方法无需耗费人力资源标注语料, 并且能够准确描述情感词和评价对象之间的搭配关系, 以及副词与情感词之间的修饰关系. 因此本文采用基于句法分析的方法对教材评论进行特征级的情感分析.

2 基础资源的构建

本文借助细粒度情感分类技术, 对从网络上抓取的大量计算机专业本科教材的评价文本进行情感极性分析, 从而辅助商家和出版社改进教材的质量、制定合理的销售策略, 并为潜在消费者的购买决策提供参考依据.

细粒度情感分析方法的基本流程如图 1 所示, 包括数据采集、数据预处理、产品特征提取以及评论文本倾向性分析四个步骤. 其中, 数据采集、预处理以及产品特征提取为情感倾向性分析算法提供基础资源, 主要包括评论语料资源、领域情感词典资源和产品特征词库资源等. 本文首先介绍语料的预处理与资源的构建, 然后在第 3 节中对核心算法—评论文本情感倾向性分析算法的设计进行详细介绍.

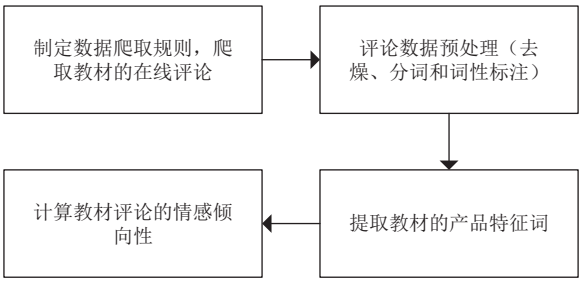

图 1 细粒度情感分析方法的基本流程图

2.1 评论语料资源的预处理

教材评论的情感倾向性分析和产品特征的提取需

要大量评论语料的支撑, 因此本文利用定制爬虫从当当、京东等网站爬取了教材评论文本, 然后对其进行了去噪、分词和词性标注等预处理, 以便为后期工作提供较好的数据资源.

2.1.1 数据去噪

从当当、京东等网站采集的原始评论中存在很多冗余评论, 如: 同一用户针对同一产品发表的多条相同评论, 这类数据会影响教材情感分析结果的准确性. 因此对这种重复数据作删除处理, 最终仅保留其中的一条. 此外, 评论中大于 200 字的长评论大都是对教材的客观介绍, 不具备情感分析条件, 因此本文删除了这类评论. 为了消除评论中的噪音, 还对语料进行了错别字纠正、拼音、英语替换等一系列去噪处理.

2.1.2 语料的分词与词性标注

利用中国科学院计算机所研发的中文分词软件 ICTCLAS2016 对已去噪的评论数据进行分词和词性标注. 由于 ICTCLAS2016 对一些计算机专业名词、网络新词等词汇的切分结果不正确, 因此本文自定义了领域分词词典, 以优化词汇切分效果.

2.2 教材评论情感词典的构建

教材评论情感倾向性分析离不开情感词典资源的支撑, 但目前, 国内还没有一部面向教材评论领域的情感词典. 而教材评论也有别于其他领域的用户评论, 它经常使用的有些情感词是其他领域很少使用或不使用的词语, 例如: “言简意赅”、“妙笔生花”、“深入浅出”等词, 因此通用情感词典难以满足教材评论情感分析研究的需求. 鉴于此, 文本选择了基础情感词典, 构建了领域情感词典、网络情感词典和极性修饰情感词典等资源.

(1) 基础情感词典

目前, 公开发表的中文情感词典资源有知网的 HowNet、台湾大学发布的 NTUSD 以及大连理工大学构建的情感词汇本体库. 上述三个词典中, HowNet 和 NTUSD 仅区分了情感词的极性, 而大连理工大学发布的情感词典不仅区分了词汇的情感极性, 还描述了词汇的情感强度. 为了计算教材评论的褒贬极性及其极性强度, 本文选择大连理工大学的情感词库作为基础情感词典.

(2) 领域情感词典

教材评论中有很多其他领域不使用, 并且通用情感词典不包含的情感词, 因此本文总结了教材评论中经常使用但基础情感词典不包括的情感词, 例如“妙笔生花”、“由浅及深”等词汇, 并人工定义情感极性, 形成了教材评论的领域情感词典, 目前, 该词典共收集了 643 个词汇.

(3) 网络情感词典

网络情感词起源于网络, 并深受网络用户的喜爱, 教材评论中网络情感词的使用频率也很高. 例如“给力”, “爆赞”等词汇, 所以本文收集、整理了这种极性明显、情感强烈的网络情感词, 形成网络情感词典.

(4) 极性修饰情感词典

用户在发表评论时, 通常会使用程度副词和否定副词来表达不同程度的情感态度. 其中程度副词影响情感的强弱, 否定词则影响情感的极性. 为此本文从相关文献中总结了修饰情感词的程度副词和否定副词, 形成极性修饰情感词典. 参考文献[9-15], 从中总结了 132 个程度副词, 并按照其对情感色彩的影响程度, 划分成四个等级, 具体见表 1 所示. 否定词则来源于对文献[10-13]的总结, 最终得到了 62 个否定副词, 如表 2 所示.

表 1 程度副词列表

级别	程度副词						数量
极量(1.5)	最	最为	极其	极为	极度	百分之百	31
高量(1.25)	太	非常	十分	满	更	更加	61
中量(0.75)	较	还	较比	不太	不甚	比较	15
低量(0.5)	略微	略	略为	稍	稍为	稍微	25

表 2 程度副词列表

否定副词							数量
无	从没	不	从未	不要	并非	不想	62
非	绝非	未	从不	甭	没有	不曾	

2.3 教材产品特征词库的构建

细粒度情感分析, 需要获取在线评论中用户评价的产品特征. 产品特征一般是名词和名词性短语, 因此特征提取则转化为对名词和名词性短语的提取和筛选. 由于本文选择的分词工具 ICTCLAS2016 仅能标注出名词, 但不能标注出名词性短语. 为此, 本文在分词结果的基础上, 制定了以下 3 条规则识别文本中的名词性短语, 这样就能较为完整地提取教材的候选产品特征.

(1) “名词+名词”规则: 相邻两个名词直接连接构成的名词词组很可能是一个产品特征. 例如评论句: 纸张/n 材质/n 一般/uyy. 其中“纸张材质”这个产品特征在

分词标注时往往分解成两个名词: “纸张”和“材质”.

(2) “名词+的+名词”规则: 结构助词“的”连接的两个名词, 也很有可能是一个产品特征, 例如评论句: 书/n 的/ude1 质量/n 不太/d 好/a. 产品特征“书的质量”由“的”连接的两个名词“书”和“质量”构成.

(3) “动词+名词”规则: 分析图书评论数据发现, 很多产品特征由相邻的一个动词和一个名词组成. 例如评论: 印刷/v 质量/n 非常/d 好/a, 快递/v 速度/n 也/d 快/a.

按照上述三条规则, 本文以评论语料中的句子为单位, 找出句子中的名词性短语, 并将其转换为名词, 最后将文本中的名词作为候选特征词提取出来, 经过人工校验再确定是否加入到产品特征词库. 具体的实现步骤如下:

(1) 按图 2 所示流程从语料中提取名词性短语.

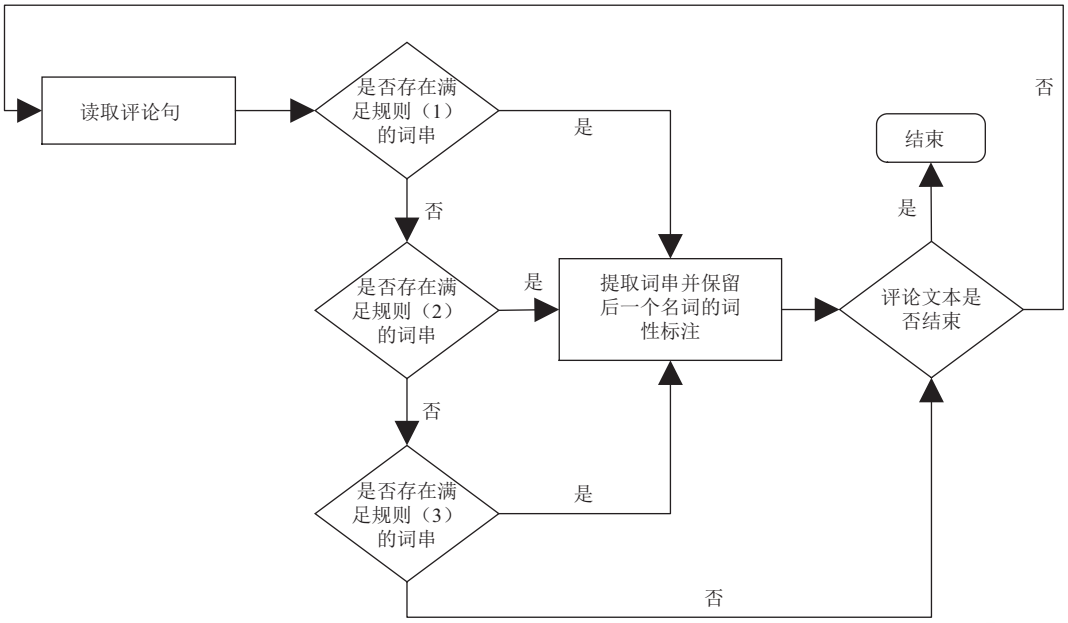

图 2 名词性短语的提取流程

(2) 人工判别提取的名词性短语是否为产品特征, 如果是, 则加入用户自定义分词词典和产品特征词库中.

(3) 利用新的用户分词词典, 再次对评论数据进行分词及词性标注, 然后再对新的标注结果执行步骤 (1)-(2), 直到没有新的产品特征加入用户分词词典为止.

(4) 提取教材评论最终标注结果中的名词, 判断其中不属于产品特征词库中的名词是否为教材产品的特征或属性, 如果是, 则合并到产品特征词库中.

经过上述步骤, 最终构建的特征词库共包括 1321 个特征词.

3 教材评论情感倾向性分析算法的设计

本文采用基于句法分析的极性判别算法实现教材评论文本的情感倾向性分析, 句法分析工具使用了哈工大社会计算与信息检索研究中心研发的语言技术平台 (LTP), 通过该平台对优化后的分词结果进行依存关系分析. 极性判别方法则借鉴了文献[2]中的情感分析算法, 首先使用 SBV、VOB、ATT 和 ADV 这四种依存关系, 从教材评论中抽取特征-意见对, 然后根据意见词及其修饰词的极性值确定特征-意见对的情感值. 其中 SBV、VOB、ATT 和 ADV 依存关系的具体说明见表 3.

表 3 LTP 依存关系标记说明

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我送她一束花 (我<--送)
动宾关系	VOB	直接宾语, verb-object	我送她一束花 (送<--花)
定中关系	ATT	attribute	红苹果(红<--苹果)
状中关系	ADV	adverbial	非常美丽(非常<--美丽)
动补结构	CWP	complement	做完了作业(做-->完)

利用文献[2]提供的依存关系对教材评论中的特征-意见对进行抽取后发现, 有些情感搭配对还无法提

取出来, 其原因在于依存分析时有些产品特征被标注成表 3 中的动补结构 (CWP), 为此本文提出增加 CWP 结构提取部分特征-意见对. 此外, 由于教材评论语句的随意性, 存在很多省略评价对象的评论语句, 例如评论句: 还不错哦, 推荐. 对于此类情况, 仅利用文献[2]的情感分析算法则无法抽取. 通过对大量教材评论的分析发现, 省略评论句大多数是对“书”的整体进行评价, 因此当评论句中情感词所修饰的评价对象被省略时, 本文提出补充“书”作为其所修饰的评价对象.

基于第 2 节生成的产品特征词库和情感词典资源, 本文的情感倾向性分析算法可描述如下: 其中 *ProductValue* 表示产品特征的情感值, *SentimentValue* 代表情感词的极性值, *AdverbStrength* 则代表副词的强度值, 这里的副词包括极性修饰情感词典中的程度副词和否定副词.

(1) 寻找一条评论语句中所有的 SBV 结构对, 对于每个 SBV 对, 记主语为 *subject*, 谓语为 *predicate*, 如果主语是产品特征, 谓语是情感词, 转步骤 (2), 如果主语不是产品特征, 但谓语是情感词, 则转步骤 (3).

(2) 继续查找谓语是否有 ADV 结构.

a) 如果有, 并且 ADV 结构中的副词 *adverb* 在副词列表中, 则:

ProductValue(*subject*)=*SentimentValue*(*predicate*)***AdverbStrength*(*adverb*)

b) 否则产品特征的极性值:*ProductValue*(*subject*)=*SentimentValue*(*predicate*)

(3) 查找谓语是否有 VOB 关系对, 如果有, 并且 VOB 关系中的宾语 *object* 是产品特征, 则该产品特征的情感值:*ProductValue*(*object*)=*SentimentValue*(*predicate*)

(4) 查找评论句中的全部 ATT 结构对, 如果其中的名词 *noun* 是产品特征, 并且定语 *attribute* 是情感词, 则判断该情感词是否有 ADV 结构.

a) 如果有, 并且包含在 ADV 结构的副词在副词列表中, 则:

ProductValue(*noun*)=*SentimentValue*(*attribute*)***AdverbStrength*(*adverb*)

b) 否则产品特征的情感值:

ProductValue(*noun*)=*SentimentValue*(*attribute*)

(5) 对于评论中所有的 CMP(动补) 结构, 如果依存关系对右边的动词 *verb* 是产品特征, 同时左边的补语 *complement* 是情感词, 则判断该情感词是否有 ADV 结构.

a) 如果有, 并且 ADV 结构中的副词在副词列表中, 则:

ProductValue(*verb*)=*SentimentValue*(*complement*)***AdverbStrength*(*adverb*)

b) 否则产品特征的情感值:*ProductValue*(*verb*)=*SentimentValue*(*complement*)

(6) 记录步骤 (1)-(5) 中处理过的情感词, 并打上 *Marked* 标签, 查找句子中未标记的情感词 *sword*, 为情感词添加“书”作为其描述的产品特征, 然后判断情感词是否有 ADV 结构.

a) 如果有, 并且 ADV 结构中的状语 *adverbial* 在副词列表中, 则产品特征“书”的情感值:*ProductValue*(书)=*SentimentValue*(*sword*)***AdverbStrength*(*adverb*)

b) 否则: *ProductValue*(书)=*SentimentValue*(*sword*).

4 实验

为了证明本文面向教材评论构建的情感词典和提出的情感倾向性分析算法的有效性, 从构建的评论语料资源中选取了 4000 个句子作为实验语料, 并人工标注了语料中产品特征和情感修饰项的关系以及产品特征的情感极性, 最终标注了 6095 个产品特征-意见对.

为了评估本文算法的性能, 采用目前常用的准确率、召回率和 *F* 值 (*F-measure*) 作为评价指标, 其计算方式定义如下:

准确率: $P=\frac{A}{B}\times 100\%$

召回率: $R=\frac{A}{C}\times 100\%$

F 值: $\mathrm{F-measure}=\frac{2\times P\times R}{P+R}\times 100\%$

其中, *A* 为算法挖掘出的正确特征-意见对的数量, *B* 为算法挖掘出的特征-意见对的数量, *C* 为测试语料中特征-意见对的人工标注数量.

在实验中, 首先使用了本文构建的教材评论情感词典, 测试算法在产品特征-情感描述项和产品特征极性方面的识别结果; 然后再使用大连理工大学发布的通用情感词典, 获得相同语料下的识别效果; 最后以人工标注结果为基准, 分别计算使用这两部情感词典的识别结果的准确率、召回率和 *F* 值, 结果如表 4 所示.

表 4 文本算法利用两部词典的情感倾向性分析结果

	准确率(%)	召回率(%)	*F*值(%)
大连理工大学发布的通用情感词典	80.4	59.1	68.1
教材评论领域情感词典	81.6	88.8	85.1

由表 4 可知, 基于本文的教材评论情感词典进行教材评论倾向性分析的准确率、召回率和 F 值, 均高于利用大连理工大学发布的通用情感词典的实验结果. 其中召回率的提高幅度最明显, 其主要原因在于本文在构建教材评论情感词典时, 考虑到用户评论中经常使用一些口语化的情感词、网络情感词以及评论教材时使用的特殊情感词, 故对这类情感词进行了归纳总结, 然后将其纳入教材评论情感词典中, 这在一定程度上扩大了情感词典的覆盖面, 提高了情感词典在教材评论领域的适用性, 因此使用该词典的倾向性判别性能明显提升, 充分证明了文本构建的教材评论情感词典的实用性和有效性.

此外, 为了比较本文情感分析算法的性能, 还将本文方法和文献[2]的方法进行了比较, 实验结果如表 5 所示.

表 5 本文算法和文献[2]算法的情感倾向性分析结果

	准确率(%)	召回率(%)	F值(%)
文献[2]	75.6	74.1	74.8
本文算法	81.6	88.8	85.1

从表 5 的结果表明, 利用本文算法进行情感倾向性分析的实验结果比文献[2]中算法的实验效果好, 其准确率、召回率、F 值的性能指标都有所提升, 这证明本文提出的教材评论情感倾向性分析算法的有效性. 同时也进一步表明, 在进行细粒度的情感分析时, 没有一种万能的倾向性分析算法能够适应所有领域的评论数据, 因此面向不同领域的评论数据进行极性分析时, 需要依据评论数据的特殊性, 提出适合于该领域评论数据的领域词典和倾向性分析算法.

5 总结

本文利用当前与日俱增的教材评论信息进行情感倾向性分析研究, 通过构建评论语料库、产品特征词库和教材领域情感词典资源, 借鉴已有的情感倾向性分析技术, 结合教材评论的不同之处, 最终提出适合教材评论的细粒度情感倾向性分析算法, 从教材评论中提取特征-意见对, 挖掘读者评论对产品特征的褒贬评价, 从而帮助消费者优化购买决策, 也可为商家改进产品、制定销售方案提供有效依据. 最后通过实验验证了本文算法的有效性. 虽然本文提出的方法在一定程度上实现了教材评论产品特征级的情感分析, 但也存在一些不足之处, 如: 产品特征和情感词典构建的自动化程度不够高, 还需要依靠人工筛选. 在情感倾向性分析方面, 特征-意见对提取的查全率还有待进一步提高. 这些将成为本文下一步的主要研究任务.

参考文献

1 姚天昉, 娄德成. 汉语语句主题语义倾向分析方法的研究. 中文信息学报, 2007, 21(5): 73–79.

2 刘丽, 王永恒, 韦航. 面向产品评论的细粒度情感分析. 计算机应用, 2015, 35(12): 3481–3486, 3505. [doi: 10.11772/j.issn.1001-9081.2015.12.3481]

3 占文平. 面向产品评论的情感分析技术研究[硕士学位论文]. 杭州: 浙江工商大学, 2015.

4 Jin W, Ho HH. A novel lexicalized HMM-based learning framework for web opinion mining. Proc. of the 26th Annual International Conference on Machine Learning. New York, NY, USA. 2009. 465–472.

5 刘鸿宇, 赵妍妍, 秦兵, 等. 评价对象抽取及其倾向性分析. 中文信息学报, 2010, 24(1): 84–88, 122.

6 陈豪, 刘功申, 黄晨. 基于句法分析的商品情感倾向性分析. 信息安全与通信保密, 2013, (2): 68–70.

7 冯时, 付永陈, 阳锋, 等. 基于依存句法的博文情感倾向分析研究. 计算机研究与发展, 2012, 49(11): 2395–2406.

8 万常选, 江腾蛟, 钟敏娟, 等. 基于词性标注和依存句法的Web 金融信息情感计算. 计算机研究与发展, 2013, 50(12): 2554–2569. [doi: 10.7544/issn1000-1239.2013.20130875]

9 邸鹏. 基于句子情感权值合成算法的篇章情感分析[硕士学位论文]. 太原: 太原理工大学, 2015.

10 郭书彤. 基于 Web 文本的图书评论倾向性分析方法的研究[硕士学位论文]. 长春: 东北师范大学, 2015.

11 周城. 面向中文 Web 评论的情感分析技术研究[硕士学位论文]. 长沙: 国防科学技术大学, 2011.

12 刘玉娇, 琚生根, 伍少梅, 等. 基于情感字典与连词结合的中文文本情感分类. 四川大学学报 (自然科学版), 2015, 52(1): 57–62.

13 徐叶强. 基于情感分类的产品评论垂直搜索引擎的研究[硕士学位论文]. 株洲: 湖南工业大学, 2012.

14 张成功, 刘培玉, 朱振方, 等. 一种基于极性词典的情感分析方法. 山东大学学报 (理学版), 2012, 47(3): 47–50.

15 陈国兰. 基于情感词典与语义规则的微博情感分析. 情报探索, 2016, (2): 1–6.