● 周　知[1]，方正东[2]
(1. 西北大学公共管理学院，陕西　西安　710127；2. 数字广东网络建设有限公司，广东　广州　510000)

融合依存句法与产品特征库的用户观点识别研究

摘　要：[目的/意义] 有效分析并利用电子商务网站用户评论数据，发掘海量用户评论数据的价值，识别用户关注的商品属性，在丰富用户行为学理论研究内容的同时，也为相关实践提供参考。[方法/过程] 基于依存句法分析技术以及 Word2Vec 词向量技术构建的产品特征库进行在线评论用户观点的抽取，并通过引入依存词对的词性特征、依存关系组合特征和词汇距离约束等方法，提升用户观点抽取的精度和质量。[结果/结论] 文章所提的基于依存句法和产品特征库的用户观点抽取方法相较于最近距离法和 SBV 极性传递法有更优的实验效果，在准确率、召回率和 *F*1 值上相较于两种基准方法均有较大的提升，证明了所提方法的有效性。

关键词：依存句法；用户观点；商品属性；产品特征；特征挖掘
DOI：10.16353/j.cnki.1000-7490.2021.07.016
引用格式：周知，方正东．融合依存句法与产品特征库的用户观点识别研究[J]．情报理论与实践，2021，44(7)：111-117.

Research on User Opinion Recognition Based on Dependency Syntax and Product Feature Thesaurus

Abstract: [Purpose/significance] Effectively analyzing and using the user review data of e-commerce website, exploring the value of massive user review data, identifying the commodity attributes of users' concern, enriching the theoretical research content of user behavior, and providing reference for related practice. [Method/process] The product feature thesaurus based on dependency parsing technology and word2vec technology is used to extract the user's opinion. The precision and quality of user's opinion extraction are improved by introducing the part of speech feature of dependency word pair, dependency combination feature and words distance constraint. [Result/conclusion] Compared with the nearest distance method and SBV polarity transfer method, the proposed user opinion extraction method based on dependency syntax and product feature thesaurus has better experimental results. Compared with the two benchmark methods in accuracy, recall rate and *F*1 value, the proposed method is effective.

Keywords: dependency syntax; user opinion; product attribute; product feature; feature mining

社会信息化程度不断加深的过程中，用户的消费活动持续从线下实体场景转移至电子商务平台，平台中的用户评价内容研究对商品质量提升、服务水平改善和营销策略制定有重要参考意义，商家也可以在用户评论特征提取的基础上，分析用户倾向与喜好，在为商家提供参考的同时优化平台信息生态环境[1]。通过用户观点识挖掘，可以对用户行为、用户认知模式进行分析与研究，为知识服务水平的优化提供参考，可以看出，对这一社会现象的研究具有明显的理论研究价值和实践意义，用户观点识别也因此逐渐成为情报学、管理科学、信息科学等学科的研究热点[2]。

用户观点识别作为自然语言处理（Natural Language Process，NLP）的重要任务之一，在研究过程中，较多使用到 NLP 领域中的特征识别与提取方法。目前较常见的包括句法分析、语义距离、点互信息、标签传播等方法，其中句法分析作为语言学中理性主义的代表性方法，以其逻辑严谨，规则明显等优势，长期以来成为语义特征提取中的重要方法[3]，成为用户观点识别的重要方法。

然而，目前基于句法规则的用户识别研究在取得了一系列进展，做出重要贡献的同时，也存在一些问题。首先，其方法较多直接使用句法规则本身，或进行简单的修正，较少考虑用户的表达方式和分析对象的领域属性。其次，对于特征词候选集合的构建，往往以名词作为起点，忽略了词汇之间的距离，以及较少考虑句法规则组合为整体效果提升带来的收益。这两方面的问题均限制了用户观点识别结果的价值。

针对这种情况，本文提出一种融合句法特征组合与产品特征库的方法，利用句法结构中词汇距离、领域特征库内容控制等方法对提取的结果进行约束与控制，优化提取准确率的同时保证最小程度牺牲召回率，并以搜索型产品

中的代表性产品——手机为例进行实验，为相关问题的解决提供思路，为理论研究的范畴拓展做出一定贡献。

1 相关研究

用户观点抽取的研究大体可以分为基于规则和模式的观点抽取，以及基于机器学习、神经网络的自动化抽取两大类。以句法规则来看，主要是通过词性分析、句法分析和语义分析等技术解析用户评论，研究者再根据自身经验学识寻找评论中符合要求的规则和模式，然后利用这些规则和模式提取用户评论中的特征观点对。在识别模式构建工作中，主要切入点可以分为基于语义关联的方法和基于句法分析的方法。

1) 基于语义关联的观点识别。这部分研究利用用户自然语言中的语义单元的关联关系，通过外部词典控制、词汇特征、语义距离或半监督学习等方法，构建观点识别模板，本质上仍然是以语言学的特征和规则为起点。

这方面研究的重要成果包括，张志远等从评论对象的类别出发，基于语义词典运用语义相似度和相关度方法抽取评价对象，最后在SEMEVAL竞赛的电子产品、餐厅和旅馆三个领域数据集上的实验证明了所提特征的有效性[4]；江腾蛟等针对金融评论评价对象复杂繁多，情感词的词性、语义更丰富等特点，提出一种基于浅层语义与语法分析相结合的评价对象—情感词对抽取方法[5]；Rana等针对现有评价目标提取方法较大依赖依存句法解析器性能且需要人工语法规则的干涉的不足，提出一种基于序列模式规则的用户评论评价目标提取方法，以学习用户行为并确定意见和评价目标的关系[6]；Al-obeidat等提出一个用户评论的意见管理框架以帮助商家识别并解决从在线产品评论中提取的客户问题，并构建了一个基于Web的交互原型，帮助企业所有者选择一组具有最佳成本/收益权衡的任务，确保所有任务都能在时间期限内完成[7]；Lin等针对传统抽取方法存在着忽略意见关系、字距限制等不足，提出一种基于主动学习的半监督提取方法，以解决监督方法中人工标注耗时且容易出错的问题[8]。

2) 基于依存句法的观点识别。通过自然语言表达的用户评论，无论内容多复杂，始终通过句法或句法的组合对其关系进行表示、分解或结构性表达。因此利用改进的依存句法规则进行观点识别，是该研究的重要思路。

以依存句法为核心或起点的研究包括，喻影等通过分析句子的逻辑结构，结合词性标注，抽取并加权与情感色彩更相关的词语，提出一种基于依存句法分析的关键词抽取方法[9]；张璞等通过分析商品评论中评价对象与评价短语的词性、依存句法及语义依存，设计核心搭配抽取规则，引入COO算法及改进ATT链算法，提出一种使用规则模板进行评价搭配抽取的方法[10]；李纲等提出一种针对产品网络评论的情感标签抽取模型，将基于依存句法的词对抽取算法与HowNet情感词典的情感计算方法相结合，利用情感极性计算过滤抽取出的情感标签[11]。夏卉等通过引入依存语法关系，对评论模板实现自动分类、过滤、泛化并形成模板库，然后基于模板库、外部词典和筛选过滤机制提取特征标签，该方法的性能优于单纯过滤与泛化的抽取方法[12]。Wawer在综述了当前波兰语单词到短语再到句子层面可用的自然语言处理、情感分析的工具、软件和解决方案后，提出一种评价意见目标的提取方法[13]；Aung等提出了一种无需训练实例的无监督提取意见和产品特征的方法，通过使用StanfordCoreNLP依赖解析器获得产品方面和观点之间的依赖关系，并基于这些关系建立抽取产品特征和观点的规则[14]。

上述两方面研究均取得了一定的效果，但建立的规则和模式较受限于各自的领域，难以跨领域移植。为了避免这些不足，相关学者们尝试使用机器学习等自动化方法抽取用户评论中的特征观点对，但是机器学习需要大量的人力标注模型的训练数据集，且最后模型的性能较大依赖于人工标注的质量。另外，虽然自动化抽取方法的性能可能更优，但难以对抽取的结果进行深入的解释和分析，不利于用户观点抽取领域知识经验的积累。本文使用基于规则和模式的人工方法抽取用户观点，提出一种基于依存句法和产品特征库的用户观点抽取方法，并辅以词性特征、依存距离约束和产品特征库过滤筛选，以提升观点抽取的效果。

2 研究设计

本文提出一种基于依存句法和产品特征库的用户观点抽取方法，以识别并抽取用户评论中的特征观点对。研究设计如图1所示。抽取方法包括三个步骤：特征观点对抽取，特征观点对完善，特征观点对过滤。首先，利用人工设定的规则和模板对特征观点对进行抽取；其次，通过特定的依存关系组合来完善前一步抽取的特征观点对；最后，利用本文构建的多维产品特征库对上一步的结果进行过滤筛选，得到最终的结果。具体的方法如下。

2.1 基于句法规则组合的特征观点对完善

本文选择主谓关系（SBV）、定中关系（ATT）、状中结构（ADV）、动补结构（CMP）和动宾关系（VOB）作为依存关系基本词对，这些关系也是搜索型产品中基于句法分析的特征词提取的重要出发点。基本抽取规则如表1所示。

需要对部分规则进行组合，以完善特征观点对抽取工作。以依存关系ATT为例对特征组合进行说明。例如，

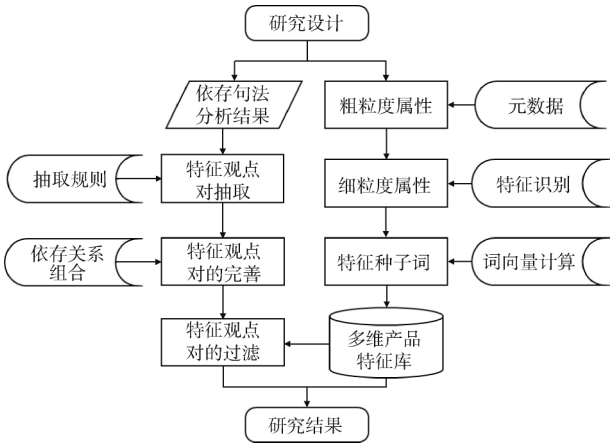

图1 基于依存句法分析和产品特征库的用户观点抽取

表1 依存关系的抽取规则

序号	抽取规则	抽取结果	序号	抽取规则	抽取结果
1	$n \xrightarrow{SBV} a$	<n，a>	7	$d \xrightarrow{ADV} a$	<d，a>
2	$n \xrightarrow{SBV} v$	<n，v>	8	$d \xrightarrow{ADV} v$	<d，v>
3	$n \xrightarrow{SBV} d$	<n，d>	9	$n \xrightarrow{VOB} v$	<n，v>
4	$n \xrightarrow{SBV} i$	<n，i>	10	$v \xrightarrow{CMP} d$	<v，d>
5	$v \xrightarrow{SBV} a$	<v，a>	11	$v \xrightarrow{CMP} a$	<v，a>
6	$n \xrightarrow{ADV} v$	<n，v>	12	$v \xrightarrow{ATT} n$	<v，n>

“手机充电速度很快”这句话的依存句法分析结果如图2所示。

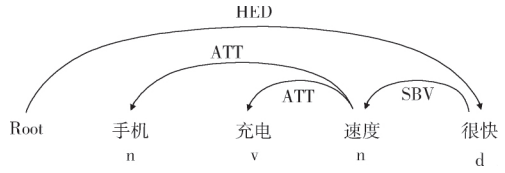

图2 “手机充电很快”的依存句法分析结果

“充电速度”作为统一整体的名词短语被拆成动词词性的“充电”和名词词性的“速度”两个部分，且“充电”和“速度”这二者的依存关系为ATT。若能对ATT关系的“<充电，速度>”和SBV依存关系的“<速度，很快>”进行整合，就可以获得完整的特征观点对“<充电速度，很快>”。故通过ATT+SBV依存关系能够还原部分本该为统一整体的特征观点对中的产品特征。

除了“ATT+SBV”这两种依存关系的组合可以进一步合并完善外，还有以下几种依存关系的特征观点对可以进一步完善，以获得语义更加清晰完整、情感更加突出的特征观点对。

1) “SBV+ADV”：对于SBV是主谓关系，ADV是状中结构，状中结构中的词语可以对主谓关系中的观点词进行修饰，这个修饰对于后续进行用户观点词的情感倾向跟强度的计算来说非常重要，所以在抽取观点时需要考虑状中结构。

2) “ATT+SBV+ADV”：是“SBV+ADV”组合的特殊情况，即此时SBV关系特征观点对中的评价对象被分词工具切成两部分，而且这两部分又满足ATT定中关系，所以也需要将ATT定中关系中的另一部分评价对象考虑进去以完善评价对象，使得最终的特征观点对更加完整、语义更加完善。

3) “SBV+VOB”：VOB动宾关系的搭配是对SBV主谓关系中谓语动词的补充和完善，即评价短语本应是“主语—谓语—宾语”的简单结构，但分词后的单独部分均是语义不完整的搭配，需要将其合并成一个部分来看，才能使得语义更加完整。

4) “ATT+SBV+VOB”：同样，这也是上面“SBV+VOB”的一种特殊情况，即简单句式“主语—谓语—宾语”中的主语不完整，被分词工具切割成了两个部分，需要用ATT定中关系完善主语评价对象。

5) “SBV+CMP”：CMP动补结构是对SBV主谓关系中谓语动词的补充，需要将这二者依存关系的部分搭配合并为一个整体，如依存句法中的SBV（小水滴，看着）和CMP（舒服，看着），只有将二者合并为<小水滴，看着，舒服>这个完整的特征观点对后才能获得更加明确的用户意见。

综上所述，形成表2所示的6种依存关系组合，完善前面抽取的特征观点对，获得观点态度更加明确的特征观点对。

表2 完善特征观点对的6种依存关系组合

序号	依存关系组合	组合结果
z-1	$ATT(w_1,w_2)+SBV(w_2,w_3)$	$<w_1,w_2,w_3>$
z-2	$SBV(w_1,w_2)+ADV(w_3,w_2)$	$<w_1,w_3,w_2>$
z-3	$ATT(w_1,w_2)+SBV(w_2,w_3)+ADV(w_4,w_3)$	$<w_1w_2,w_4,w_3>$
z-4	$SBV(w_1,w_2)+VOB(w_3,w_2)$	$<w_1,w_2,w_3>$
z-5	$ATT(w_1,w_2)+SBV(w_2,w_3)+VOB(w_4,w_3)$	$<w_1w_2,w_3,w_4>$
z-6	$SBV(w_1,w_2)+CMP(w_2,w_3)$	$<w_1,w_2,w_3>$

需要指出，依存关系中词语的距离远近会影响抽取词对的质量，特别是在句子比较长的情况下，词对间会形成较多的无效依存关系。Hu等[15]指出特征观点对词汇间的位置距离会影响观点抽取的结果；刘海涛研究后发现[16]，相较于英语和日语，汉语的依存距离为2.81。因此，本文取3个词汇单位的距离为特征观点对抽取的极限，即当依赖词与核心词的距离小于等于3个词汇长度时，视其为有效的依存关系，将其抽取出来放入待处理的特征观点对数据集中，否则视作无效的依存关系，不进行抽取。

2.2 基于Word2Vec的产品特征库的构建

产品特征库的构建相当于构建一个商品种子词集，利用种子词集对2.1节的句法分析结果进行约束，即如果某特定词出现或未出现在词集中，则保留或过滤该词的策略也有所不同。具体的利用方法在1.3节提出。本节主要说明特征库的构建方法，拟结合具体领域为例进行说明，以搜索型产品中的经典代表性手机为例。首先，参照专业的IT信息网站中关村在线手机频道（https://mobile.zol.com.cn/）中手机产品的相关参数设置以及天猫、淘宝、京东等主流网购平台的手机商品参数设置，并结合用户评论的评价特征，将手机产品的粗粒度属性主要划分为价格、屏幕、外观、网络与通话、性能、相机、电池、硬件与配置和服务这9个，每个粗粒度属性下包含的细粒度特征见表3。

表3 手机产品的粗细粒度属性特征

粗粒度属性	细粒度特征
价格	定价、性价比、价保
屏幕	显示、性质、解锁
外观	尺寸、颜值、配色
网络与通话	网络、通话
性能	运行、游戏、操控、影音、功能测评、系统应用
相机	镜头、模式技术、拍照录像
电池	充电、续航
硬件与配置	存储、CPU、扬声器、导航红外
服务	配送、客服运营、赠品配件、售后

用户在线评论的评价对象多为名词、名词短语和动词，而观点词则多为形容词，还有少部分动词。本文在前面分词和词性标注的基础上，使用参考文献[17]的方法，获取各细粒度特征下的代表词汇，即统计词性为动词和名词的特征词的词频，然后选取词频大于3的词汇，并依据表3为每个细粒度特征人工挑选具有代表性的3个特征代表词汇，共计获得30个细粒度特征的90个代表词汇，如表4所示。

运用Word2Vec词向量技术对代表词汇挑选后的剩余词汇进行向量空间映射并依据相似度高低进行特征归类。首先，将各个词汇转换成对应的词向量，计算待归类词汇与上述30个细粒度特征下的代表词集的相似度；其次，根据该相似度均值的高低，将待归类词归类到相似度最高的细粒度特征中。待归类词汇与代表词集相似度的计算公式如公式（1）所示。

$$\mathrm{Sim}(w,\mathrm{seed}_i)=\frac{\sum_{k=1}^{3}(\mathrm{sim}(w,\mathrm{seed}_{ik}))}{3} \quad (1)$$

式中，$i\in[1,30]$，且为正整数；$\mathrm{Sim}(w,\mathrm{seed}_i)\in[0,1]$，$\mathrm{seed}_{ik}$表示第$i$个细粒度特征下的代表词汇集中的第$k$个代表词；$\mathrm{sim}(w,\mathrm{seed}_{ik})$表示该词汇与第$k$个代表词的相似度，且$\mathrm{sim}(w,\mathrm{seed}_{ik})\in[0,1]$。

表4 各细粒度特征下的代表词汇

细粒度特征	代表词汇	细粒度特征	代表词汇
定价	价格、价钱、价位	功能测评	小爱、AI键、跑分
性价比	性价比、价格比、性比价	系统应用	系统、应用、MIUI
价保	价保、保值、降价	镜头	镜头、焦距、超广角
显示	显示、分辨率、屏幕色彩	模式技术	夜景、人像、微距
性质	全面屏、屏占比、防划	拍照录像	照相、像素、录像
解锁	解锁、人脸、指纹	充电	充电、快充、无线充
尺寸	尺寸、手感、握持	续航	续航、电量、待机时间
颜值	颜值、外形、外观	存储	存储、运行内存、内存
配色	颜色、渐变色、色彩	CPU	CPU、骁龙、处理器
网络	网络、上网、断网	扬声器	扬声器、外放、喇叭
通话	通话、语音、听筒	导航红外	导航、GPS、红外
运行	运行、死机、发热	配送	配送、快递、收货
操控	触控、操作、反应	客服运营	客服、卖家、服务态度
游戏	游戏、掉帧、延迟	赠品配件	赠品、赠送、钢化膜
影音	视频、音乐、画质	售后	售后、维修、退货

使用Python的gensim包进行评论语料中词汇的向量化，基于该词向量模型进行待归类词汇与各细粒度特征下代表词汇间相似度的计算。在计算出待归类词汇与评价特征代表词汇两两之间的相似度后，根据公式（1）计算待归类词汇与细粒度特征代表词集的平均相似度。以此为依据，将待归类词汇归入相似度最高的细粒度特征下的词集中。例如，对于反映电池续航的“耗电量”这个待分类词，它与续航、游戏、充电、运行和系统应用这几个较相关特征的代表词集的相似度分别为：

$$\mathrm{Sim}(\text{耗电量},\mathrm{seed}_{\text{续航}})=\frac{\mathrm{sim}(\text{耗电量},\text{续航})+\mathrm{sim}(\text{耗电量},\text{电量})+\mathrm{sim}(\text{耗电量},\text{待机时间})}{3}=0.7891$$

$$\mathrm{Sim}(\text{耗电量},\mathrm{seed}_{\text{游戏}})=\frac{\mathrm{sim}(\text{耗电量},\text{游戏})+\mathrm{sim}(\text{耗电量},\text{掉帧})+\mathrm{sim}(\text{耗电量},\text{延迟})}{3}=0.7677$$

$$\mathrm{Sim}(\text{耗电量},\mathrm{seed}_{\text{充电}})=\frac{\mathrm{sim}(\text{耗电量},\text{充电})+\mathrm{sim}(\text{耗电量},\text{快充})+\mathrm{sim}(\text{耗电量},\text{无线充})}{3}=0.7020$$

$$\mathrm{Sim}(\text{耗电量},\mathrm{seed}_{\text{运行}})=\frac{\mathrm{sim}(\text{耗电量},\text{运行})+\mathrm{sim}(\text{耗电量},\text{死机})+\mathrm{sim}(\text{耗电量},\text{发热})}{3}=0.6853$$

$$\mathrm{Sim}(\text{耗电量},\mathrm{seed}_{\text{系统应用}})=\frac{\mathrm{sim}(\text{耗电量},\text{系统})+\mathrm{sim}(\text{耗电量},\text{应用})+\mathrm{sim}(\text{耗电量},\text{MIUI})}{3}=0.6350$$

“耗电量”与续航的3个代表词汇的相似度最高，所以将“耗电量”这个评价特征归入“续航”这个细粒度特征下，这也较为符合常识和语言习惯。同时，通过设定相似度的最小阈值以过滤无意义评价特征，即需要设定

最大相似度的最低值，以排除如“男朋友”“女朋友”“家人”等无意义的评价对象。当某个待归类评价特征的最大相似度低于该值时，将该待归类词汇剔除，不进行细粒度特征的归类，本文将该最大相似度的最小阈值设为0.5。

在基于Word2Vec的评价特征归类方法的基础上，将所有的未归类词汇进行细粒度特征的归类处理，并最后用人工对归类结果进行二次校验，修正其中不合理的归类，最后总计获得30个细粒度特征下的633个用户评价特征词汇，以此形成手机产品的产品特征库，所构建的产品特征库如表5所示。

表5 小米9手机的产品特征库

粗粒度属性	细粒度特征	评价特征
价格	定价	价格、价钱、价位、便宜、贵…
	性价比	性价比、价格比、性比价、性價比、划算…
	价保	价保、保值、降价、价格保护、掉价…
屏幕	显示	显示、显示效果、屏幕色彩、屏幕显示、分辨率…
	性质	全面屏、屏占比、防划、康宁大猩猩、耐摔…
	解锁	解锁、人脸、指纹、识别率、手纹…
外观	尺寸	尺寸、手感、握持、大小、单手操作…
	颜值	颜值、外形、外观、款式、外型…
	配色	颜色、渐变色、色彩、蓝色、白色…
网络与通话	网络	网络、上网、断网、网速、信号…
	通话	通话、语音、听筒、打电话、失声…
性能	运行	运行、死机、发热、黑屏、速度…
	操控	操控、触控、反应、响应、滑动…
	游戏	游戏、掉帧、延迟、满帧、吃鸡…
	影音	视频、音乐、画质、电影、追剧…
	功能测评	小爱、AI键、跑分、测评、分身…
	系统应用	系统、应用、MIUI、程序、内设…
相机	镜头	镜头、焦距、超广角、摄像头、主摄…
	模式技术	夜景、人像、微距、月亮模式、防抖…
	拍照录像	照相、像素、录像、照片、自拍…
电池	充电	充电、快充、无线充、充电速度、车充…
	续航	续航、电量、待机时间、耗电、存电…
硬件与配置	存储	存储、运行内存、内存、储存、ROM…
	CPU	CPU、骁龙、处理器、高通、芯片…
	扬声器	扬声器、外放、喇叭、失音、外音…
	导航红外	导航、GPS、红外、定位、NFC…
服务	配送	配送、快递、收货、发货、物流…
	客服运营	客服、卖家、服务态度、分期、现货…
	赠品配件	赠品、赠送、钢化膜、屏保、插针…
	售后	售后、维修、退货、换货、保修…

2.3 基于产品特征库的特征观点对过滤

在特征观点对的抽取后，其中会存在部分无任何实际意义的特征观点对，如<手机，很好>这种太过宽泛而缺乏任何实际对象的搭配，运用1.2节中构建的产品特征库来对上述的抽取结果进行过滤和筛选，以提高观点抽取的准确率。

将特征观点对的过滤筛选操作放在完善操作之后的原因是，较长的特征观点对会被分词软件切分成多个部分，通过依存句法分析可以得到这些部分之间的联系，如果在合并完整的评价对象之前就根据产品特征进行过滤的话，会造成部分表面上不包含产品特征而实际上与其他的词对进行合并可以得到完整评价搭配的词对被过滤掉。为了尽可能多且准确地抽取特征观点对，本文将特征观点对的过滤操作放在完善操作后。具体方法如下。

运用2.2节所构建手机的产品特征库对2.1节的句法规则组合进行约束时，主要有下列4种情况:

1) 若w_1是表5中的词汇，即为产品特征库中的评价特征，则可以直接保留。

2) 若w_2是表5中的词汇而w_1不是，保留当前搭配。

3) 若w_3是表5中的词汇而w_1和w_2不是，保留当前搭配。

4) 若w_1、w_2和w_3均不是表5中的词汇，则需要进一步地拆分词汇，重新判断，具体有以下3种: ①若w_1词汇中截取的前两个字符或后两个字符是表5中的词汇，即为产品特征库中的评价特征，则可以直接保留; ②若w_2词汇中截取的前两个字符或后两个字符是表5中的词汇而w_1不是，保留当前搭配; ③若w_3词汇中截取的前两个字符或后两个字符或后一个字符是表5中的词汇而w_1和w_2不是，保留当前搭配。

除了上述所列的几种情况外的特征观点对，一律剔除。符合上述几种情况的特征观点对会被保留，作为最终用户观点抽取结果。

3 实验过程与结果分析

实验部分包括过程描述和结果分析，过程描述部分包括对用户评论的抓取与预处理、分词及词性标注、依存句法分析及最终结果抽取。结果分析部分，将实验抽取结果和最近距离法、SBV极性传递法进行对比，在准确率、召回率、$F1$值三个数据上进行效果对比与讨论。

3.1 实验过程

基于依存句法和产品特征库的用户观点抽取实验的具体流程如图3所示，主要分为用户评论抓取、评论语料预处理、评论分词和词性标注、句法分析以及观点抽取6个环节。

1) 用户评论抓取。本文以2019年2月上市的小米9手机为研究对象，爬取京东商城小米9手机评论，且总数不低于1000的店铺作为对象，使用“Python3 + WebDriver + Selenium”来模拟用户浏览行为爬取评论。受限于京东商城的反爬虫机制，每家店铺最多可以爬取100页每页10条即1000条评论数据，累计爬取10713条用户评论。所获取的评论样本及其字段如表6所示。

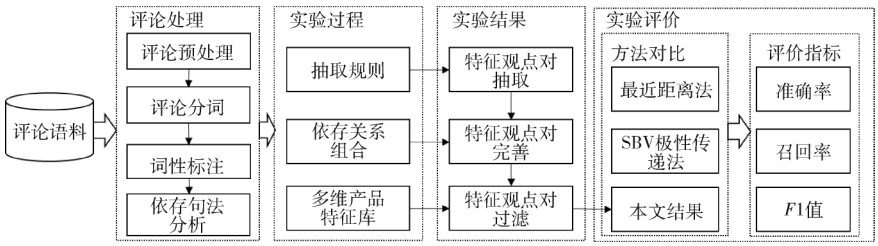

图3 基于依存句法和产品特征库的用户观点抽取实验流程图

表6 实验语料样例表（局部）

用户ID	时间	评分	评价内容
jd_336-0	2019/4/27	4	用起来挺不错的，物流非常流畅就是电量有待提高，总体来说不错
海阔天空	2019/4/27	5	手机很不错很漂亮性能好是我想要的会继续关注小米手机
淘气响响	2019/4/27	5	绝对正品，物流也快，包装也非常好
…	…	…	…
下午_茶_	2019/4/29	5	退役的米6，战斗的米9，4800W像素，不说了，哎呀，真香!

2）评论预处理。对10712条用户评论语料进行预处理。剔除如“系统默认好评”“超赞”等不包含评价对象的无意义短评，在进行去重、清洗操作后，获得6301条评论数据，作为后续实验的评论数据集。对于用户评论的观点抽取实验，由于需要对用户评论进行人工标注以比对实验效果，从上述“干净”的数据集中随机抽取600条评论作为观点抽取实验的数据集，作为效果评估数据集。

3）评论分词。本文选用结巴分词来进行评论分词。用户评论的口语化、多样性以及复杂性，加大了中文分词的难度，如表达积极情感的“牛皮”“快的一匹”网络用语，很难被识别为一个整体，往往被拆成多个部分，如“牛皮”被拆成“牛”“皮”，破坏了原有组合代表的新语义。为保证分词的准确率，需要添加用户自定义词典，将这些独特、新颖的网络用语、评价口语短语分别作为一个整体添加到自定义词典中，以保证评论语料中的相应评价短语被准确分词。

4）词性标注。在此使用哈工大的Pyltp来进行词性标注（863词集），并在此基础上进行依存句法分析。考虑到评论语料中存在着大量难以被准确标注词性的网络用语、短语等，如表达赞扬的形容词“nb”，表达消极意义的“呵呵了”“辣鸡”等网络用语，使用Pyltp的自定义词性标注词典，将这些难以或无法被准确标注的词语、短语添加到自定义词典中，提高词性标注的准确率，为后续的依存句法分析奠定基础。

5）依存句法分析。使用哈工大语言云的工具包Pyltp进行依存句法分析，获得用户评论中词对之间的依存关系。需要指出，当用户评论语句很长，且句式语法不够规范时，依存句法分析的结果往往会存在大量的误差，需要使用相关规则模板并结合词性特征、依存距离约束等对其进行完善和补充，以获得语义完整、正确的评价搭配。

6）用户观点抽取。在得到各评论语句的依存句法分析结果后，首先根据1.1节中的抽取规则模板从分析结果中提取依存关系为SBV，ATT，ADV，VOB和CMP且词性特征也满足条件的词对，作为初步的观点抽取结果；然后根据1.2节中的“ATT+SBV”“SBV+ADV”“SBV+VOB”等6种依存关系组合对初步抽取的结果进行合并和完善，以抽取更加完整、语义更明确的特征观点对；最后根据1.3节中所构建的手机多层次产品特征库对完善后的特征观点对进行过滤，获得最终的抽取结果。

3.2 结果分析

本文选用准确率、召回率和$F1$值作为本文所提的观点抽取方法的评估指标，并将本文方法与基准方法的实验结果进行对比以验证本文方法的有效性。对于用户在线评论的观点抽取实验来说，准确率、召回率以及$F1$值三者的计算公式分别如公式（2）~公式（4）所示：

$$\text{准确率(Precision)} = \frac{\text{抽取正确的特征观点对数目}}{\text{抽取出来的特征观点对总数目}} \tag{2}$$

$$\text{召回率(Recall)} = \frac{\text{抽取正确的特征观点对数目}}{\text{语料数据中的特征观点对总数目}} \tag{3}$$

$$F1 = \frac{2 * \text{准确率} * \text{召回率}}{\text{准确率} + \text{召回率}} \tag{4}$$

本文选取下列两种方法作为对比的基准方法，来验证本文所提的基于依存句法分析和产品特征的评论观点抽取方法的有效性。

1）最近距离法。文献[15]首先挖掘评论语料中频繁出现的名词或名词短语将其作为待选的特征词汇，并将出现在3条及3条用户评论以上的特征词汇作为抽取的评价对象，然后抽取评论语句中距离评价对象3个词汇范围内的形容词作为其观点词，进而得到<评价对象，观点词>的特征观点对。

2）SBV极性传递法。文献[18]首先抽取评论语句中依存关系为SBV（主谓关系）的词对作为候选的特征观点对，然后借助ATT链算法识别完善的评价对象，进而完善上一步抽取的特征观点对，得到用户评论的最终抽

取结果。

对随机抽取的600条用户评论进行人工标注，获得817个特征观点对，分别运用三种实验方法进行用户观点抽取实验得到抽取结果，计算出三种实验方法在准确率、召回率和*F*1值三个评估指标上的值，结果如表7所示。

表7 三种观点挖掘方法的评估结果

实验方法	准确率/Precision	召回率/Recall	*F*1值
最近距离法	0.62	0.35	0.45
SBV极性传递法	0.39	0.38	0.38
本文方法	**0.70**	**0.72**	**0.71**

可以看出，本文所提方法在三个评估指标上的表现均优于另外两种对比方法，且在召回率和*F*1值上相较于两种基准方法有较大的提升，证明了本文方法的有效性。对于最近距离法，该方法的产品特征只考虑了名词，忽略了用户评论中部分动词和习语也可以充当特征观点，如“运行”“充电”等，限制了最终抽取结果的召回率。SBV极性传递法仅仅考虑SBV和ATT两种依存关系，排除了评论中存在的蕴含大量用户观点信息的ADV、CMP和VOB等依存关系，也没有对主语和谓语的词性做进一步挖掘，导致准确率及召回率受到限制。

相较于基线方法，本文在准确率、召回率和*F*1值上均取得更优的效果，不仅考虑了用户评论语句中的句法依存关系，同时兼顾了特征观点对的词性特征，通过引入词对间3个词汇距离的依存距离约束，来排除依存句法分析结果中的无效和干扰词汇组合，克服了过长的评论语句会造成句法分析结果引入过多干扰依存关系，保证了抽取结果的较高准确率；另外，通过评论语句中的ATT句法传递关系对观点抽取结果中的复合评价对象进行完善，以及使用ADV、VOB和CMP依存关系综合抽取部分复杂句式中的特征观点对，进一步提升了抽取结果的准确率和召回率。

4 结束语

针对传统句法分析方法在用户评论特征词对提取中的固有短板和问题，本文提出一种融合句法规则组合和产品特征库的方法，在有效解决特征提取工作的同时，有效建立了产品的特征库，并以京东商城上小米手机为例进行实验，本文的策略组合有效地提升了词汇提取的准确率、召回率，最终的实验结果表明，本文所提方法在准确率、召回率和F1值上与最近距离法和SBV极性传递法两种基准方法相比，均有较大的提升，证明了本文所提的观点抽取方法的有效性。

本文也存在某些局限性，仅以手机为代表性对象的搜索型产品特征提取，未考虑到其他类型的产品或体验型产品，一定程度上限制了方法的可拓展性，另外，特征库的构建过程中也过于集中地考虑了手机产品的构建方法与流程，对其他产品的通用能力缺乏更深入的探讨，这些问题也是今后研究的重点和方向，在依存句法的强化利用和特征库的深度结合上，做出进一步探索。□

参考文献

[1] ZENG Ziming，ZHOU Zhi，MU Xiangming. User review helpfulness assessment based on semantic analysis [J]. The Electronic Library，2020，38 (2)：337-351.
[2] 周清清，章成志. 在线用户评论细粒度属性抽取 [J]. 情报学报，2017，36 (5)：484-493.
[3] 宗成庆. 中文信息处理研究现状分析 [J]. 语言战略研究，2016，1 (6)：19-26.
[4] 张志远，赵越. 基于语义和句法依存特征的评论对象抽取研究 [J]. 中文信息学报，2018，32 (6)：80-87，97.
[5] 江腾蛟，万常选，刘德喜，刘喜平，廖国琼. 基于语义分析的评价对象—情感词对抽取 [J]. 计算机学报，2017，40 (3)：617-633.
[6] RANA T A，CHEAH Y. Sequential patterns rule-based approach for opinion target extraction from customer reviews [J]. Journal of Information Science，2019，45 (5)：643-655.
[7] AL-OBEIDAT F，SPENCER B，KAFEZA E. The opinion management framework：identifying and addressing customer concerns extracted from online product reviews [J]. Electronic Commerce Research and Applications，2018，27：52-64.
[8] LIN Yuming，JIANG Xiangxiang，LI You，ZHANG Jingwei，CAI Guoyong. Semi-supervised collective extraction of opinion target and opinion word from online reviews based on active labeling [J]. Journal of Intelligent and Fuzzy Systems，2017，33 (6)：3949-3958.
[9] 喻影，陈珂，寿黎但，陈刚，吴晓凡. 基于关键词和关键句抽取的用户评论情感分析 [J]. 计算机科学，2019，46 (10)：19-26.
[10] 张璞，李逍，刘畅. 基于规则的评价搭配抽取方法 [J]. 计算机工程，2019，45 (8)：217-223.
[11] 李纲，刘广兴，毛进，叶光辉. 一种基于句法分析的情感标签抽取方法 [J]. 图书情报工作，2014，58 (14)：12-20.
[12] 聂卉，杜嘉忠. 依存句法模板下的商品特征标签抽取研究 [J]. 现代图书情报技术，2014 (12)：44-50.
[13] WAWER A. Sentiment analysis for Polish [J]. Poznan Studies in Contemporary Linguistics，2019，55 (2)：445-468.
[14] AUNG S S. Analysis on opinion words extraction in electronic product reviews [J]. International Journal of Systems and Software Security and Protection (IJSSSP)，2019，10 (1)：47-61.
[15] HU Minqing，LIU Bing. Mining and summarizing customer reviews [C] //Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York，USA：ACM Press，2004：168-177.
[16] 刘海涛. 依存语法的理论与实践 [M]. 北京：科学出版社，2009.
[17] 卢玲，杨武，刘旭，李言. 基于实体情感演化置信网的观点检测方法 [J]. 计算机应用，2017，37 (5)：1402-1406.
[18] 顾正甲，姚天昉. 评价对象及其倾向性的抽取和判别 [J]. 中文信息学报，2012，26 (4)：91-98.

作者简介：周知 (ORCID：0000-0002-5530-2968)，男，1989年生，博士，讲师。方正东 (ORCID：0000-0001-8118-2923)，男，1995年生，硕士，工程师。

录用日期：2021-02-02