Journal of Computer Applications
计算机应用,2015,35(12):3481 -3486,3505
ISSN 1001-9081
CODEN JYIIDU
2015-12-10
http://www.joca.cn

文章编号:1001-9081(2015)12-3481-06
doi:10.11772/j.issn.1001-9081.2015.12.3481

面向产品评论的细粒度情感分析

刘　丽*,王永恒,韦　航
(湖南大学 信息科学与工程学院,长沙 410082)
(*通信作者电子邮箱270818745@qq.com)

摘　要:针对传统粗粒度情感分析忽略具体评价对象,以及现有细粒度情感分析方法忽略无关评价要素的问题,提出结合条件随机场(CRF)和语法树剪枝的方法对产品评论进行细粒度情感分析。采用基于MapReduce的并行化协同训练(Tri-training)的方法对语料进行半自主标注,利用融合多种语言特征的条件随机场模型,获取评论中的评价对象和正负面评价词。通过建立领域本体和句法路径库实现语法树剪枝,对含有多个评价对象和评价词的文本,去掉无关评价对象的干扰,抽取出正确的评价单元,最后形成可视化产品报告。实验结果显示,提出的方法在两种不同领域数据集上,识别情感要素的综合准确率达89%左右,情感评价单元的综合准确率也达89%左右。实验结果表明,与传统方法相比,结合CRF和语法树剪枝的方法识别准确率更高,性能更好。

关键词:产品评论;细粒度情感分析;MapReduce;协同训练;条件随机场;语法树剪枝
中图分类号:TP391.1　　**文献标志码**:A

Fine-grained sentiment analysis oriented to product comment

LIU Li*, WANG Yongheng, WEI Hang
(*College of Information Science and Engineering, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: The traditional sentiment analysis is coarse-grained and ignores the comment targets, the existing fine-grained sentiment analysis ignores multi-target and multi-opinion sentences. In order to solve these problems, a method of fine-grained sentiment analysis based on Conditional Random Field (CRF) and syntax tree pruning was proposed. A parallel tri-training method based on MapReduce was used to label corpus autonomously. CRF model of integrating various features was used to extract positive/negative opinions and the target of opinions from comment sentences. To deal with the multi-target and multi-opinion sentences, syntax tree pruning was employed through building domain ontology and syntactic path library to eliminate the irrelevant target of opinions and extract the correct appraisal expressions. Finally, a visual product attribute report was generated. After syntax tree pruning, the accuracy of the proposed method on sentiment elements and appraisal expression can reach 89% approximately. The experimental results on two product domains of mobile phone and camera show that the proposed method outperforms the traditional methods on both sentiment analysis accuracy and training performance.

Key words: product comment; fine-grained sentiment analysis; MapReduce; Tri-training; Conditional Random Field (CRF); syntax tree pruning

0　引言

随着Web2.0的盛行,人们表达情感和意见的方式已不再局限于传统的面谈、写日记等方式,而是逐渐转向网络平台,以文本的形式来表达。尤其体现在电子商务上,电子商务的发展,推动消费者网购热情的同时,也产生大量的产品评论。产品评论是消费者分享使用效果、评价商品的重要数据资源,为用户消费习惯的调查、企业制定营销策略和产品满意度调查等提供可靠的数据支持。人工浏览无法应对海量评论资源,如何方便快捷地挖掘出评论中有价值的信息逐渐成为研究热点。情感分析能从评论中获取用户的喜怒哀乐,从而知道用户对商品的喜好程度及意见,对用户而言多了一份选择商品的依据,对商家而言,可以从中改进产品,提升产品受欢迎度。

传统的情感分析,往往是面向篇章和句子级别的粗粒度的分析方法,主要包括情感词典法和机器学习法。

基于情感词典的方法主要是通过统计文章或者句子中正负面情感词的数量来判断情感极性。文献[1]手工构建正负面情感词典,并利用这两个词典对股评文档中的情感词进行统计,按照一个正面情感词记为1、一个负面情感词记为-1、中型词记为0的计分策略,得到文档的情感倾向性,最后分析出股评文档和股票走势的关系。随着大量网络词汇、口语词在评论句子中的出现,现有的情感词典无法识别这些带有情感色彩的词汇,导致分类效果不是很理想。

基于机器学习的方法,分为无监督学习方法和有监督学习方法。文献[2]认为含有副词和形容词的短语为情感词,为此提出一种无监督的学习方法,利用相应的规则识别出这些短语,并分别计算这些短语与规定的正面情感词和负面情

收稿日期:2015-06-23;**修回日期**:2015-08-02。
基金项目:国家自然科学基金资助项目(61371116);湖南省自然科学基金资助项目(13JJ3046)。
作者简介:刘丽(1990-),女,山西临汾人,硕士研究生,主要研究方向:文本分析、数据挖掘;　王永恒(1973-),男,河北霸州人,副教授、博士,主要研究方向:大规模数据库、数据挖掘、物联网复杂事件处理;　韦航(1990-),女,广西柳州人,硕士,主要研究方向:文本分析、数据挖掘。

感词的点互信息值(Pointwise Mutual Information, PMI),两值之差作为该短语的情感倾向,文章中所有短语的情感倾向的平均值视为文章的情感倾向值。文献[3]选择最大熵模型、朴素贝叶斯和支持向量机作为有监督学习的文本分类模型,选择特征bigram、unigram、词性标注以及词的位置作为情感分析的特征。

粗粒度的情感分析没有考虑情感所面向的具体对象,无法满足用户了解产品的各个方面特性的需求,为此,提出细粒度的情感分析方法。

细粒度情感分析是面向评价对象的情感分析,可以分析出一条评论中参与表达情感的各个要素,包括评价对象、正负面评价词等。现有方法通常包括无监督和有监督的方法。

无监督的分析方法往往是基于规则模板的方法,文献[4]中提出一种关联规则法来抽取评价对象和评价词,认为名词和名词短语为评价对象,形容词为候选评价词。文献[5]在文献[4]的基础上引进点互信息法进行改进,通过计算名词或名词短语与规定标识词之间的点互信息值,来确定属于评价对象的可能性,从而去掉不属于评价对象的名词或名词短语,并通过抽取的评价对象和句法关系来辅助抽取评价词。这类方法抽取评论语料中的属性词(评价对象)和评价词,对于罕见属性词和评价词的抽取效果不好,而且是单独抽取属性词和评价词的,忽略了二者之间的关系。

有监督学习方法中,文献[6]采用一种基于词汇化的隐马尔可夫模型(Hidden Markov Model, HMM)的方法,将评价对象和评价词的抽取看作是一个序列标注任务,通过标注类别来确定哪些词属于评价对象和评价词,但它是一种产生式模型,不能很好地融合各种特征。文献[7]提出一种基于依存句法树结构的结合条件随机场(Conditional Random Field, CRF)模型,来联合抽取评价对象和评价词,在线性条件随机场基本点特征的基础上,加入了依存句法树中的树边特征,改善了线性CRF中情感要素长距离语义依赖的问题。文献[8]提出基于语言学结构的CRF模型进行情感分析,以顺序结构、连接词结构、句法树结构、连接词与句法树相融合的结构作为四种语言学结构,分析比较这几种结构下的CRF模型情感分析的性能。这两种方法将语言学结构用于CRF的模型实现中,充分利用各种语言学特征,提高了情感要素识别的准确率。

文献[9]将本体域知识作为CRF训练的一种特征,将评论中每种属性的类型作为一个特征,如手机的内存、屏幕、按键等属于硬件一类,则将这些属性标注为硬件,作为一种本体特征。文献[10]中将CRF和遗传算法相结合,利用遗传算法优胜劣汰的思想,随机从语义特征集中选取了最好的特征进行训练。这两种方法从条件随机场训练所需特征入手,引入了新颖有助于提高识别准确率的特征。文献[11]中提出结合主动学习的CRF模型,用主动学习的方法代替了手工标注语料,提高了语料标注的效率。文献[12]中将协同训练Co-training的思想用在了训练CRF模型上,不仅实现了语料的自主标注,而且完成了CRF模型的训练。这两种方法从标注语料入手,克服了以往人工标注的费时费力的缺点,提高了标注的效率。

现有的细粒度情感分析方法,为了提高情感要素的识别效果,提取了多种多样有价值的分类特征,但是忽略了评论文本的复杂性,评论文本中经常会出现多个评价对象和评价词,有些是与评论主体无关的,一定程度上影响真正情感要素的识别,而且在此基础上利用邻近法抽取的评价单元也不够准确。

为此本文在CRF模型的基础上,引入语法树剪枝的方法,通过剪枝去掉无关评价对象和评价词,不但提高评价对象和评价词的识别准确率,也提高了评价单元的准确率。从而形成可靠的可视化产品报告。此外为了进一步提高语料标注的效率,采用基于MapReduce的并行化协同训练(Tri-training)思想来标注实验语料。

1 基于条件随机场的情感要素识别

1.1 条件随机场

条件随机场是一种用于序列标注的概率统计模型,由Lafferty等[13]于2001年首次提出,它结合了最大熵模型和隐马尔可夫模型的特点。用于评论文本情感要素识别时,输入观察序列,即经过分词的评论文本 $X=\{x_1,x_2,\cdots,x_n\}$,就可以计算所有可能的状态序列(即每个词被标注的类别)的条件概率,并输出概率最大时的序列状态 $Y=\{y_1,y_2,\cdots,y_n\}$,计算公式如下:

$$P(Y|X)=\exp\left(\sum_i\sum_k\lambda_k f_k(y_{i-1},y_i,X,i)\right)/Z(X) \quad (1)$$

其中 $Z(X)$ 是归一化因子,它可以确保所有的概率 P 小于1,计算公式如下:

$$Z(X)=\exp\left(\sum\sum\lambda_k f_k(y_{i-1},y_i,X,i)\right) \quad (2)$$

式中:X 是观察序列;Y 就是对应的标注完后的状态序列;$f_k(y_{i-1},y_i,X,t)$ 是一个任意的特征函数,每个特征函数 f 表示为观察序列的实数值特征集合中的一个元素;λ_k 是对应的特征函数的权值。

当条件概率最大时,得到想要的序列状态。

1.2 语料的半自主标注

手动标注语料费时费力,基于MapReduce的Tri-training模型来半自动标注语料提高语料标注的性能,而且结合人工验证,进一步确保标注的准确度。MapReduce是由谷歌开发的一款并行编程模型,可以并行处理大规模数据,分为Map和Reduce两个过程,两个过程的输入和输出都是键值〈key, value〉的形式。Tri-training,是一种半监督学习方法,由Zhou在文献[14]中提出的一种既不需要充分冗余视图也不需要使用不同的分类器的算法,可以利用少部分已标注语料,自主标注大规模语料,融合两种方法对语料标注有很大帮助,但是Tri-training整个过程是一个迭代的过程,每标注一部分语料都要用到上一步已标注的语料,不适合并行处理,但每一步的迭代,是串行且需要处理大量文本的,因此可以把Tri-training的每一步的迭代进行一个并行处理,同样可以减少语料标注的时间开销,主要思路如下:

1)定义标注集,分为四类:评价对象(Target of Opinion, TO)、正面评价(Positive Opinion, PO)词、负面评价(Negative Opinion, NO)词、背景词(Background Word, BW)。具体说明如表1。

2)初始数据集标注。利用随机采样算法从评论语料中

抽取三个数据集 $D1$、$D2$、$D3$,对这些数据进行分词,然后手工对每个词按照 1)中的分类进行标注,得到如下所示初始训练集:

电池/TO 太/BW 垃圾/NO 了/BW,/BW 用/ BW 不/BW 到/BW 半天/BW 就/BW 没/BW 电/BW 了/BW

表 1　标注集说明

标注	说明
TO	评论中的评价对象,如手机的性能、颜色、像素等属性
PO	对评价对象的正面意见、观点
NO	对评价对象的负面意见、观点
BW	除上面三种词之外余下的词

3)朴素贝叶斯分类器作为基分类器,对数据集 $D1$、$D2$、$D3$ 进行训练,得到三个有差异的分类器模型 $M1$、$M2$、$M3$ 。

4) 用 $M1$、$M2$、$M3$ 分别对分词后的未标注文本 X 进行标注,标注过程中,包括两个并行化过程。

①单个分类器进行词标注的并行化。

用分类模型对文本 X 中的每个词进行分类时,需要计算词被标注为每一类(TO,NO,PO,BW)的概率,协同训练计算概率时是依次计算词属于每个标注类别的概率,随着训练集迭代增大,计算过程也趋于复杂。对该过程进行并行处理可大大减少时间开销,过程中:Map 的输入为词和标注类别(TO,NO,PO,BW),其中 key 为词,value 为标注类别,生成的中间结果键值对为〈词,标注概率〉;Reduce 接受中间键值传过来的数据,将 Key 值相同的形成键值对〈词,(概率 1,概率 2,概率 3,概率 4)〉,最后将概率最大的〈词,概率〉提取出,就完成计算词属于每个类别的概率,得到词的标注类别。

②投票法文本标注的并行化。

经过 $M1$、$M2$、$M3$ 的处理,每个词都会有三个标注,需要投票决定词属于哪一类。Map 阶段输入词和分类模型,其中 key 为词,value 为分类模型($M1$, $M2$, $M3$),产生的中间结果为〈词,标注类别〉,Reduce 阶段将 key 值相同的键值对形成列表〈词,类别 1,类别 2,类别 3〉,若类别 1 和类别 2 一样,即 $M1$ 和 $M2$ 的分类结果一样,则该词就被标注为类别 1,并加入到 M3 所在的训练集 $D3$ 中,如此形成 $M3$ 的新训练集,$M1$ 和 $M2$ 的训练集也是按这种方法扩充。

5)记录标注后的 X,然后 $D1$、$D2$、$D3$ 重新训练,继续标注下一条文本,如此重复迭代,直到未标注数据完全标完。最后进行人工验证,确保所有语料标注的准确性。

1.3　特征选择

特征选择的好坏直接影响情感要素识别的效果,因此本文选择词、词性、依存关系特征,并加入领域本体特征和评价信息特征,具体如下:

词特征:指经过分词后评论中的每个词,是情感分析的主体,是需要标注的序列。

词性特征:指当前词的词性,要识别的情感要素都有着一定的词性,如评价对象一般是名词,评价词一般是形容词,这些特征在识别情感要素时起到至关重要的作用。

依存句法特征:由于评论句子的结构往往趋于复杂,仅靠词和词性特征,识别效果不是很好,故采用文献[15]中提到的依存句法特征,这个特征表明句子中各个词之间的依赖关系,具体特征含义如表 2。

领域本体特征,该特征主要是为了识别评价对象而提出的,通过构建领域本体知识,可以将评论中每个词分为属性类、产品品牌类、其他等三类。

表 2　依存句法特征表示

依存句法特征	具体含义
父节点	当前词在依存关系中的父节点词
父节点的词性	当前词的父节点词的词性
依赖关系	当前词与父节点词之间的依存关系

评价信息特征,为了更好地区分出正面评价和负面评价,通过 Hownet 中文评价词典匹配法判断评价词是正面评价还是负面评价。

这两种特征都采用三元特征值表示法,具体如表 3。

表 3　特征表示

特征	特征信息	表示方法
领域本体特征	属性类	1
	产品品牌类	-1
	其他	0
评价信息特征	正面评价词	1
	负面评价词	-1
	其他	0

2　评价单元的抽取

实现评价对象和评价词的抽取后,还需要抽取评价单元,评价单元抽取是指将评价词语及其所修饰的评价对象作为一个单元抽取出来。文献[16]中提出以评价词为中心,评价对象的识别仅考虑围绕着评价词在给定窗口范围内进行查找的方法,获取评价单元。由于该方法窗口大小的限制,以及评论文本中无关评价对象的影响,使得抽取到的评价单元准确率不是很高。因此在抽取评价单元前先进行语法树剪枝,流程如图 1 所示。

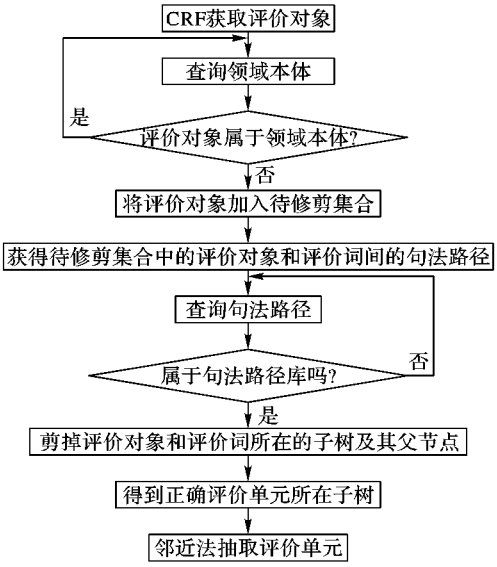

图 1　评价单元抽取流程

2.1　领域本体构建

本体是一种共享概念模型的明确的形式化规范说明。用于描述特定领域知识的专门本体叫作领域本体,可以形式化地描述领域中概念及相互关系,以及该领域所具有的特性和

规律。领域本体的构建需要确定领域本体的专业领域和范畴,以及领域中的重要术语。本文采用文献[17]中提到的形式概念分析(Formal Concept Analysis, FCA)法构建相关领域本体。图 2 给出了关于手机的领域本体。

在这个本体域图中,“手机”为概念,上方表示手机内涵,即手机具有“性能”“系统”“内存”等属性,下方指手机外延,如“小米”“三星”等具体品牌的手机。其中“智能休眠”等属性属于“三星”所特有的属性,作为“三星”的特有属性集与其相关联。在商品或服务领域,同一个概念的外延中的各个实体,是存在竞争关系的实体,如“小米”与“苹果”“三星”等。

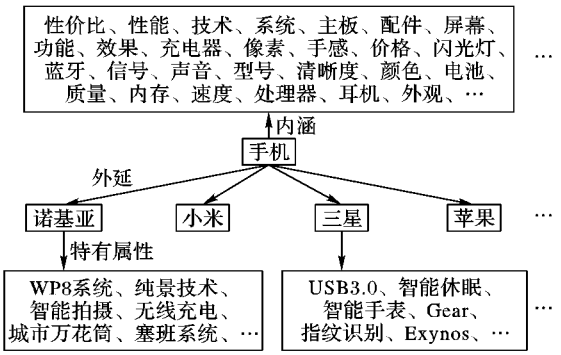

图 2 手机领域本体

2.2 构建句法路径库

句法路径库的构建是指获得语法树上评价对象与评价词语之间的句法路径,也就是指这两个词之间的句法结构。如图 3 所示的语法树中,评价对象和评价词之间有四条句法路径,其中“快递”为一个评价对象,“坑爹”为一个评价词语,两节点之间的句法路径(箭头所示)为:NN→NP→IP→VP→VP→VV。

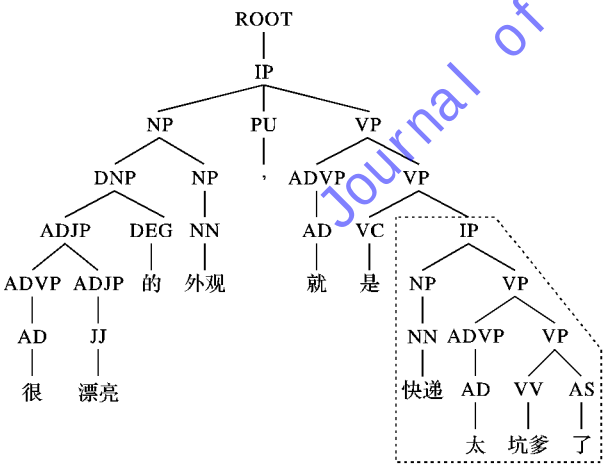

图 3 句子 1 的语法树剪枝示例

文献[18]中提出了句法路径库构建的方法,认为评价词语与其修饰的评价对象之间的句法路径是具有一定的规律的、可总结的。通过统计大规模语料库中的句法路径,发现评价对象与其真正相对应的评价词之间的句法路径出现频率较多,而与其没有正确对应关系的评价词之间的句法路径出现频率较少。因此这些句法路径的统计能较为真实地反映句法路径的正确性和实用性。故采取该方法来构建句法路径库,包括句法路径的生成和泛化。

通过 CRF 抽取到评价对象和评价词之后,统计评价对象与评价词之间的句法路径,并进行泛化处理,将句法路径中出现的连续相同的句法标签用一个标签代替,也就是合并句法路径中连续的相同成分,如上面的句法路径 NN→NP→IP→VP→VP→VV 可以泛化为 NN→NP→IP→VP→VV。按照频率多少将句法路径排序,设定阈值 m,将前 m 个句法路径加入句法路径库,其他频率的句法路径被去掉。

2.3 语法树剪枝

领域本体和句法路径库构建之后,进行语法树剪枝,语法树剪枝是针对含有多个评价对象和评价词的产品评论,将与产品主体无关的评价对象及评价词从语法树中剪除,得到与产品主体相关的评价单元。详细剪枝的流程如下:

1)通过 CRF 抽取的评价对象和评价词,构成评价对象集合和评价词集合。

2)查询评价对象集合,找出领域本体中不存在,且与产品主体无关的实体、属性特征等加入待修剪评价对象集合;对于存在于领域本体中,但与产品主体存在竞争关系的实体,也加入待修剪集合。

3)评价对象集查询完后,针对得到的待修剪评价对象,获取这些评价对象与评价词之间的语法树上的句法路径,匹配 2.2 节构建好的句法路径库,识别出对其进行修饰的相对应的评价词,从而找出无关的评价单元。

4)找出无关评价单元中评价对象和评价词在语法树上对应的共同父节点,剪去父节点下方含有该评价词和评价对象的子树,若剪枝后,该父节点下没有其他子树,则剪除该父节点。

实验中从京东商城获取了苹果手机和索尼相机的产品评论,下面分别以这两类产品评论中的两个句子为例,详细说明剪枝过程。

句子 1 关于索尼相机的评论:“很漂亮的外观,就是快递太坑爹了”。其中{外观,快递}为评价对象;{漂亮,坑爹}为评价词。查询领域本体发现“快递”是无关的对象,加入待修剪集合。图 3 为该句子的语法树结构,“快递”与“漂亮”之间的句法路径为 NN→NP→IP→VP→IP→NP→DNP→ADJP→JJ,与“坑爹”之间的句法路径为 NN→NP→IP→VP→VV。

搜索句法路径库发现库中包含第二条句法路径,不包含第一条句法路径,可知与“快递”相对应的评价词是“坑爹”。在语法树中找到这两个节点的共同父节点“IP”。剪去以该节点 IP 为根节点,并含有“快递”和“坑爹”的子树,此时该节点下已无子树,剪除该节点。

句子 2 关于苹果手机的评论:“三星是很漂亮,但上档次的还是苹果”,句子中{苹果,三星}为评价对象,{上档次,漂亮}是评价词,查询本体发现“三星”是无关的评价对象,加入待修剪集合。图 4 为该句子的语法树结构,“三星”与“漂亮”之间的句法路径为 NR→NP→IP→ VP→VA,与“上档次”之间的句法路径为 NR→NP→IP→NP→CP→IP→VP→VV。

搜索句法路径库发现库中包含第一个句法路径,而不包含第二个句法路径,可判断出与“三星”相对应的评价词是“漂亮”。在语法树中找到这两个节点的共同父节点“IP”。剪去以该节点 IP 为根节点,并含有“三星”和“漂亮”的子树,此时该节点下已无子树,剪除该节点。

经过语法树剪枝后,结合 CRF 识别出的评价对象(产品的各种属性)和评价词,利用文献[16]中提到的邻近法就可以直接抽取出评价单元。

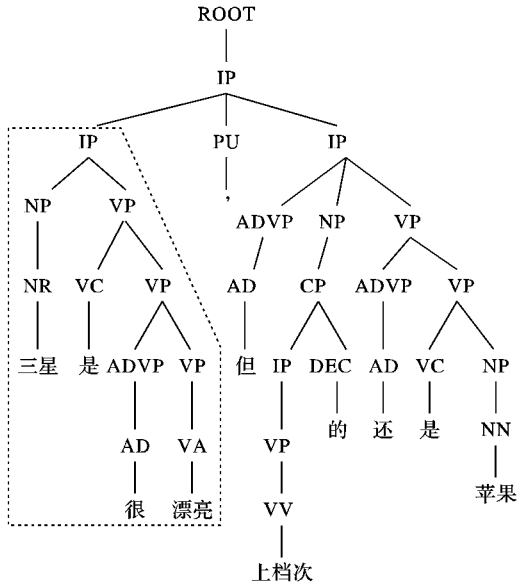

图 4　句子 2 的语法树剪枝示例

3　可视化产品报告

可视化产品报告是指根据生成的评价单元,抽取出一个评价对象的所有评价词。形成一份针对评价对象的可视化产品报告,如表 4 关于苹果手机的可视化产品报告。

表 4　可视化产品报告

产品	属性	正面评价词	负面评价词
苹果手机	外观	好看,漂亮,大气,…	难看,一般,普通,…
	信号	强,好,给力,…	弱,差劲,坑爹,…
	…	…	…

4　实验结果及分析

4.1　数据收集及预处理

本次实验的数据选择京东商城上“苹果手机”和“索尼相机”两个产品领域的评论。具体数据集信息如表 5。

表 5　实验数据统计信息

产品	正面评论数	负面评论数
苹果手机	9357	7661
索尼相机	9891	5785

首先对评论文本中的无效网址及字符作预处理,为使分词精度更高,更易识别罕见评价对象和评价词,利用自然语言处理与信息检索平台(Natural Language Processing and Information Retrieval sharing platform, NLPIR)汉语分词系统进行新词发现。语料半自主标注前,分别从两类评论数据中随机抽取了 5% 的正面评论文本和负面评论文本,信息如表 6,并对这些评论进行分词,手动标注类别。通过基于 MapReduce 的并行化 Tri-training 算法对未标注语料进行标注,再进行人工验证,进一步保证语料标注的准确性。最后将两个领域的语料,分别进行六折交叉验证,即将所有语料分为 6 份:5 份作为训练集,1 份作为测试集进行循环实验。

4.2　句法路径库的构建

根据 2.2 节介绍的句法路径库构造方法,用 Stanford Parser 对两个领域的文本进行语法分析,并对评价对象和评价词之间的句法结构进行统计,根据文献[18]中提到的阈值 m 设定的方法,得到 $m=5$,故得到表 7。

表 6　随机采样数据统计信息

产品领域	正面评论			负面评论		
	D1	*D2*	*D3*	*D1*	*D2*	*D3*
苹果手机	468	468	468	383	383	383
索尼相机	495	495	495	289	289	289

表 7　出现最频繁的五种句法路径

序号	句法路径	出现频率
1	NN→NP→IP→VP→VA	14617
2	NN→NP→NN	12035
3	NN→NP→VP→VV	11084
4	NN→NP→ADJP→JJ	10562
5	NN→NP→IP→VP→VV	9051

4.3　情感要素抽取结果

本文采用 CRF 模型作为情感要素识别的工具,采用哈尔滨工业大学的语言技术平台(Language Technology Platform, LTP)获得词、词性及依存句法特征,采用领域本体知识获得本体域特征,采用 HowNet 的中文评价词典获得评价信息特征。用 Stanford Parser 进行语法分析,获得语法树。用 5 种情感要素识别方法对不同领域的评论语料进行识别效果的比较,评价指标为精准率 P(Precision)、召回率 R(Recall)和精准率和召回率的调和平均值 F(F-measure),所得结果如表 8。观察表 8 可以看出:第一种方法,两个领域内的精准率都达到 80% 左右,说明依存句法特征发挥了一定的作用,可以捕获词与词之间的依赖关系,但是召回率不是很高,低至 59.1%。第二种方法将 CRF 与本体域相结合,一定程度上提高了评价对象的识别率,但是正负面评价词的识别效果很不理想。分析发现,领域本体特征能很好捕捉评价对象信息,而且第一种方法在正负面评价词的识别性能上要优于第二种方法,故综合两种方法,评价对象与正负面评价词的整体识别效果都相对理想,达到 78% 左右。为进一步提高情感要素的识别率,第四种方法引入评价信息特征,正负面评价词的识别效果有很大提升,这是由于引入的特征在捕捉感情词时更加灵敏,识别出了很多被遗漏的评价词,但是对评价对象的识别影响不大,因为评价对象一般是名词或名词短语,难以用情感特征去捕捉。在这些特征的基础上,第五种方法对语法树进行了剪枝,去掉与主体无关的评价对象,从结果中可以看出,手机领域和相机领域评价对象的识别效果进一步得到提升,正负面评价词的精准率和召回率也相应提升,说明语法树剪枝法一定程度上去掉了无关评价的干扰,使得评价对象和评价词的特征更具鲜明性,识别效果更加理想。

4.4　评价单元抽取结果

现有提取评价单元的方法仅限于邻近法,故识别情感要素后,对传统邻近法与剪枝后的邻近法抽取评价单元的性能作了比较,结果如表 9。

表 8 情感要素抽取结果 %

序号	情感要素识别方法	情感要素	苹果手机评论文本			索尼相机评论文本		
			P	*R*	*F*	*P*	*R*	*F*
1	基于依存句法树结构的 CRF 模型(词 + 词性 + 依存句法特征)[7]	评价对象	80.6	62.9	70.7	78.4	59.1	67.4
		正面评价词	79.2	68.7	73.6	83.3	76.4	79.7
		负面评价词	76.5	80.9	78.6	78.6	75.2	76.9
2	融合领域本体的 CRF 模型(词 + 词性 + 领域本体特征)[9]	评价对象	85.3	75.7	80.2	83.5	70.1	76.2
		正面评价词	73.3	64.7	68.7	70.2	65.9	68.0
		负面评价词	79.8	57.2	66.6	71.5	69.6	70.5
3	融合依存句法和领域本体的 CRF 模型(词 + 词性 + 依存句法 + 领域本体特征)	评价对象	86.2	73.9	79.6	84.7	78.3	81.4
		正面评价词	80.1	72.8	76.3	72.3	80.6	76.2
		负面评价词	76.3	79.2	77.7	82.1	77.4	79.7
4	融合评价信息的 CRF 模型(词 + 词性 + 依存句法 + 领域本体 + 评价信息特征)	评价对象	88.4	79.2	83.5	87.2	71.3	78.5
		正面评价词	85.6	80.5	83.0	80.7	84.9	82.7
		负面评价词	78.4	85.2	81.7	82.9	79.8	81.3
5	基于 CRF 和语法树剪枝(词 + 词性 + 依存句法 + 领域本体 + 评价信息特征 + 剪枝)	评价对象	86.4	89.1	87.7	89.5	87.5	88.5
		正面评价词	87.6	90.9	89.2	91.1	87.4	89.2
		负面评价词	92.1	88.4	90.2	89.2	93.7	91.4

表 9 评价单元抽取结果 %

评价单元抽取方法	苹果手机			索尼相机		
	P	*R*	*F*	*P*	*R*	*F*
传统邻近法	71.2	67.4	69.2	75.9	66.7	71.0
邻近法 + 语法树剪枝	85.4	87.6	86.5	89.1	90.2	89.6

分析结果,可明显发现邻近法抽取的评价单元精准率和召回率都很低,主要是因为邻近法获取评价单元的过程中比较注重经验,而且是在规定窗口内以评价词为中心寻找匹配的评价对象,窗口的大小限制了评价单元的抽取效果,窗口太小可能找不到合适的评价对象,窗口太大则会找到多个评价对象。为此语法树剪枝后,去掉句子中的无关评价,可以在大窗口范围内寻找匹配对象,使得评价单元精准率和召回率都提高了很多;最后将两个领域内提取出的评价单元,形成第 3 章所述的可视化产品报告。

5 结语

现有的细粒度情感分析方法大多忽略了多个评价对象的句子中,无关评价对情感要素和评价单元提取的影响。本文提出 CRF 模型和语法树剪枝相结合的方法来去掉无关评价的干扰。在准备实验语料的过程中,采用基于 MapReduce 的并行化 Tri-training 的方法进行语料的半自主标注,手工标注一小部分语料,利用半监督的学习方法实现未标注语料的标注,节省人力和时间资源,最后进行人工验证。然后融合多种可以捕捉语义信息和情感信息的特征,利用 CRF 抽取出情感要素,再通过构建领域本体和句法路径库,对识别出的评价对象和评价词进行筛选,抽取出正确的评价对象和评价词,从而抽取出正确的评价单元,生成可靠的可视化产品报告。

特征的选取对细粒度情感分析有着关键性的作用,一个有价值的特征可以捕获到更多语义信息,也有助于情感要素的识别。语法树剪枝方法在处理一些较为复杂的句子,如比较句、转折句等时,效果不是很理想,还存在一定的局限性,因此,在今后的工作中将尝试提取更多有价值的特征,并进一步研究如何处理较为复杂的句子。此外,语料的半自主标注中,标注的准确性需要人工验证,在保证准确性的同时降低了效率,这也是今后需要改进的地方。

参考文献:

[1] DAS S, CHEN M. Yahoo! for amazon: extracting market sentiment from stock message boards [C]// Proceedings of the 2001 Asia Pacific Finance Association Annual Conference. Bangkok: [s. n.], 2001: 35 – 43.

[2] TURNEY P. Thumbs up or thumbs down? sentiment orientation applied to unsupervised classification of reviews [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2002: 417 – 424.

[3] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? sentiment classification using machine learning techniques [C] // Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2002: 79 – 86.

[4] HU M, LIU B. Mining and summarizing customer reviews [C]// KDD'04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168 – 177.

[5] POPESCU A M, ETZIONI O. Extracting product features and opinions from reviews [C]// Proceedings of the 2005 Human Language Technology Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2005: 339 – 346.

[6] JIN W, HO H H. A novel lexicalized HMM-based learning framework for Web opinion mining [C]// Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 465 – 472.

[7] ZHANG Y. Finer grained opinion analysis on product reviews [D]. Harbin: Harbin Institute of Technology, 2013: 21 – 27. (张玥. 面向产品评价的细粒度情感分析技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2013: 21 – 27.)

[8] LI F, HAN C, HUANG M, *et al*. Structure-aware review mining and summarization [C]// Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg: ACL, 2010: 653 – 661.

[9] DING S, JIANG T. Comment target extraction based on conditional random field & domain ontology [C]// Proceedings of the 2010 International Conference on Asian Language Processing. Piscataway: IEEE, 2010: 189 – 192.

(下转第 3505 页)

[6] CHEN Z, LI Z. Collaborative filtering recommendation algorithm based on user characteristics and item attributes [J]. Journal of Computer Applications, 2011, 31(7): 1748 – 1750. (陈志敏, 李志强. 基于用户特征和项目属性的协同过滤推荐算法[J]. 计算机应用, 2011, 31(7): 1748 – 1750.)

[7] SCHAFER J B, FRANOWISKI D, HERLOCKER J, *et al*. Collaborative filtering recommender systems [C]// The Adaptive Web, LNCS 4321. Berlin: Springer, 2007: 291 – 324.

[8] LI G, ZHANG Z, LIU F, *et al*. Nonlinear combinatorial collaborative filtering recommendation algorithm [J]. Journal of Computer Applications, 2011, 31(11): 3063 – 3067. (李国, 张智斌, 刘芳先, 等. 非线性组合的协同过滤推荐算法[J]. 计算机应用, 2011, 31(11): 3063 – 3067.)

[9] SHI K, LIU J. Effects of directive second-order similarity on collaborative filtering recommender [J]. Journal of University of Shanghai for Science and Technology, 2014, 36(1): 31 – 33. (石珂瑞, 刘建国. 二阶有向相似性对协同过滤算法的影响[J]. 上海理工大学学报, 2014, 36(1): 31 – 33.)

[10] ZHANG Z. Social tagging systems: structure, dynamic and function [J]. Journal of University of Shanghai for Science and Technology, 2011, 33(5): 445 – 451. (张子柯. 社会化标签系统的结构, 演化和功能[J]. 上海理工大学学报, 2011, 33(5): 445 – 451.)

[11] PAZZANI M J. A framework for collaborative, content-based and demographic filtering [J]. Artificial Intelligence Review, 1999, 13(5/6): 393 – 408.

[12] KONSTAN J A, MILLER B N, MALTZ D, *et al*. GroupLens: applying collaborative filtering to USENET news [J]. Communications of the ACM, 1997, 40(3): 77 – 87.

[13] MASLOY S, ZHANG Y. Extracting hidden information from knowledge networks [J]. Physical Review Letters, 2001, 87(24): 248701.

[14] ZHOU T, KUSCSIL Z, LIU J, *et al*. Solving the apparent diversity-accuracy dilemma of recommender systems [J]. Proceedings of the National Academy of Sciences, 2010, 107(10): 4511 – 4515.

[15] ZHANG Y-C, BLATTNER M, YU Y-K. Heat conduction process on community networks as a recommendation model [J]. Physical Review Letters, 2007, 99(15): 154301.

[16] HUANG Z, CHUANG W, CHEN H. A graph model for e-commerce recommender systems [J]. Journal of the American Society for Information Science and Technology, 2004, 55(3): 259 – 274.

[17] ZHOU T, REN J, MEDO M, *et al*. Bipartite network projection and personal recommendation [J]. Physical Review E, 2007, 76(4): 046115.

[18] LIU J, WANG B-, GUO Q. Improved collaborative filtering algorithm via information transformation [J]. International Journal of Modern Physics C, 2009, 20(2): 285 – 293.

[19] SHANG M-S, JIN C-H, ZHOU T, *et al*. Collaborative filtering based on multi-channel diffusion [J]. Physica A: Statistical Mechanics and Its Applications, 2009, 388(23): 4867 – 4871.

[20] MARSDEN P V, FRIENKIN N E. Network studies of social influence [J]. Sociological Methods & Research, 1993, 22(1): 127 – 151.

[21] LIU J, HU Z, GUO Q. Effect of the social influence on topological properties of user-object bipartite networks [J]. The European Physical Journal B, 2013, 86(11): 1 – 11.

[22] LATAPY M, MAGNIEN C, VECCHIO N D. Basic notions for the analysis of large two-mode networks [J]. Social Networks, 2008, 30(1): 31 – 48.

[23] WU Y, ZHANG P, DI Z, *et al*. Study on bipartite networks [J]. Complex Systems and Complexity Science, 2010, 7(1): 1 – 12. (吴亚晶, 张鹏, 狄增如, 等. 二分网络研究[J]. 复杂系统与复杂性科学, 2010, 7(1): 1 – 12.)

[24] ZHANG Y, NI J, GUO Q, *et al*. Empirical analysis of diversity of online user interests [J]. Application Research of Computers, 2014, 31(11): 3250 – 3252. (张一璐, 倪静, 郭强, 等. 在线用户兴趣多样性的实证研究[J]. 计算机应用研究, 2014, 31(11): 3250 – 3252.)

(上接第 3486 页)

[10] ZHU J, WANG H, MAO J. Sentiment classification using genetic algorithm and conditional random fields [C]// Proceedings of the 2nd IEEE International Conference on Information Management and Engineering. Piscataway: IEEE, 2010: 193 – 196.

[11] ZHANG K, XIE Y, YANG Y, *et al*. Incorporating conditional random fields and active learning to improve sentiment identification [J]. Neural Networks, 2014, 58(5): 60 – 67.

[12] YANG L, LIU G, LIU Q, *et al*. Analyzing sequence data based on conditional random fields with co-training [C]// Proceedings of the 2012 8th International Conference on Computational Intelligence and Security. Piscataway: IEEE, 2012: 94 – 98.

[13] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001: 282 – 289.

[14] ZHOU Z. The cooperative training model of semi-supervised [M]// Machine Learning and Application. Beijing: Tsinghua University Press, 2007: 259 – 275. (周志华. 半监督学习中的协同训练风范[M]// 机器学习及其应用. 北京: 清华大学出版社, 2007: 259 – 275.)

[15] WANG R, JU J, LI S, *et al*. Feature engineering for CRFs based opinion target extraction [J]. Journal of Chinese Information Processing, 2012, 26(2): 56 – 61. (王荣洋, 鞠久鹏, 李寿山, 等. 基于 CRFs 的评价对象抽取特征研究[J]. 中文信息学报, 2012, 26(2): 56 – 61.)

[16] LIU K, XU L H, ZHAO J. Opinion target extraction using word-based translation model [C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012: 1346 – 1356.

[17] EFSTRATIOS K, CHRISTOS B, THEOLOGOS D. Ontology-based sentiment analysis of twitter posts [J]. Expert System with Applications, 2013, 40(10): 4065 – 4074.

[18] ZHAO Y, QIN B, CHE W, *et al*. Appraisal expression recognition based on syntactic path[J]. Journal of Software, 2011, 22(5): 887 – 898. (赵妍妍, 秦兵, 车万翔, 等. 基于句法路径的情感评价单元识别[J]. 软件学报, 2011, 22(5): 887 – 898.)