评价主题挖掘及其倾向性识别

李 芳[1,3] 何婷婷[2,3] 宋 乐[2,3]
(华中师范大学国家数字化学习工程技术研究中心 武汉 430079)[1]
(华中师范大学计算机科学系 武汉 430079)[2]
(国家语言资源监测与研究中心网络媒体分中心 武汉 430079)[3]

摘 要 主要研究如何从在线评论文本中挖掘产品的评价主题,并对其倾向性进行分析。首先采用一种启发式规则和共现概率统计相结合的方法识别文本集合中的名词性短语,再运用LDA模型挖掘潜在的评价主题。然后利用多特征融合的方法计算句子的倾向性,进而根据特征词群统计出各主题的倾向性结果。最后通过对网络汽车评论文本语料的实验证实了该方法的有效性。

关键词 LDA,评价主题,倾向性识别
中图法分类号 TP391.3 **文献标识码** A

Opinion Topic Mining and Orientation Identification

LI Fang[1,3] HE Ting-ting[2,3] SONG Le[2,3]
(National Engineering Research Center for E-learning, Huazhong Normal University, Wuhan 430079, China)[1]
(Department of Computer Science, Huazhong Normal University, Wuhan 430079, China)[2]
(Network Media Branch, National Language Resources Monitoring and Research Center, Wuhan 430079, China)[3]

Abstract The paper mainly focused attention on how to mine opinion topic from online subjective text set, and identified the orientation of these opinion topics. It combined the heuristic rule and co-occurrence probability to identify area noun phrase, and adopted Latent Dirichlet Allocationn (LDA) to obtain local opinion topic. Then, it computed the sentence orientation with a method of multiple features fusion, and obtained the opinion topic orientation by adding up the number of each orientation of the sentences belonging to the specified topic. Experimental results show that the method can identify the topic and their orientation effectively.

Keywords LDA, Opinion topic, Orientation identification

1 引言

随着Internet的普及和发展,互联网商务化程度迅速提高,网络购物占整个互联网应用的33.8%[1],互联网中人们对事物的评价信息已成为驱动消费者购买、促进产品更新升级等的重要因素。然而,快速增长的在线评论网站和急速膨胀的评价信息使得我们面临着信息爆炸却知识匮乏的窘境。人们往往需要花费大量的时间浏览很多相关网站,在过滤掉大量无用信息之后,才能得到较全面的有价值的评价信息。

在这一背景下,意见挖掘技术成为了当前自然语言文本处理领域的研究热点,其主要目的就是挖掘并分析主观评论性文本中对人物、产品、传媒等事物的观点和看法。意见挖掘的主要任务[2]包括:(1)评价对象抽取;(2)倾向性分析;(3)意见持有者的识别;(4)陈述的选择。目前国内外的研究重点主要是评价对象抽取和倾向性分析。

评价对象抽取的任务是找出文本中的评价主体对象或主体属性,所采用的方法主要有两种:一是基于规则抽取的统计方法,这类方法首先采用搭配规则[3]或关联规则[4]来挖掘出候选特征术语,然后利用混合语言模型[3]或者共现概率等统计学方法最终确定描述评价对象的名词。另一种则是采用句法分析的方法,通过名词与极性词在句法结构上的修饰关系来确定评价对象[5,6]。

倾向性分析主要包括词语级、句子级、篇章级的倾向性识别。词语级的倾向性识别主要采用与种子词的共现概率统计[7]或基于现存语义知识库(如WordNet[8]、HowNet[9]等)两种分析策略,它是句子级和篇章级倾向性识别的基础。句子级和篇章级的倾向性识别多采用有指导的多特征融合的机器学习方法,比如朴素贝叶斯、最大熵和支持向量机等经典分

到稿日期:2011-07-28 返修日期:2011-11-05 本文受国家自然科学基金重大研究计划课题(90920005),国家自然科学基金(61003192),教育部哲学社会科学研究重大课题攻关项目(08JZD0032),教育部/国家外国专家局高等学校学科创新引智计划课题(B07042),湖北省自然科学基金计划项目(2009CDB145),武汉市晨光计划项目(201050231067),华中师范大学中央高校基本科研业务费项目(CCNU10A02009,CCNU10C01005)资助。

李 芳(1984—),女,博士生,主要研究方向为情感计算与自动文摘,E-mail:fang__lf@163.com;何婷婷(1964—),女,博士,教授,博士生导师,主要研究方向为自然语言处理、数据库与数据挖掘;宋 乐(1986—),男,硕士,主要研究方向为情感计算。

类方法[10]。

目前，意见挖掘的结果大致有两种：一种是给出文本的褒贬极性，但在同一篇文本中，作者可能对事物的多个特征有不同的态度，只给出整个文本的褒贬极性显得太过于笼统；另一种是给出句子的褒贬极性，具体评价事物的某一特征，但可能有多个句子都对这一特征做出了评价，且观点不一致，这样人们就无法直接得知关于某个特征的综合评价信息。评价主题的倾向性识别可以在一定程度上弥补这些不足。

通常，人们会从不同方面来评价一个事物(也称为评价主体)。那么，主体的评价信息中就包含了多个评价主题，这里的评价主题指人们到底是对主体对象哪些方面的特征进行评价。评价主题的倾向性识别对事物各方面的特征都能给出一个综合性的观点，更加方便用户浏览信息。

评价主题的倾向性识别首先需要从多个评论文本中抽取出评价主题，然后再分析各个主题的倾向性。由此，本文首先通过一种启发式规则和共现概率统计的方法识别出评论文本集合中的专业名词性短语，然后再用 LDA 模型对所有句子集合进行建模挖掘主题，从而得到各个主题的相关词群。在倾向性分析方面，本文考虑了句子的极性词、否定副词、词语搭配、标点符号、情感指示词 5 个特征，并将它们进行线性融合来判别句子的倾向性，再统计出包含主题特征词群的句子的褒贬倾向性数目，并以此作为该主题倾向性识别的结果。

2 评价主题抽取

2.1 LDA 模型简介

LDA 模型[11]是 Blei 等在 2003 年提出的一种概率生成模型。它假设集合中的每篇文本都是潜在主题的随机混合，每个主题又为词汇的概率分布，则语料中的每篇文本的生成过程如下：

I. 选择 $N\sim\text{Poisson}(\xi)$，N 代表文本的长度；

II. 选择 $\theta\sim\text{Dir}(\alpha)$，$\theta$ 是列向量，代表每个主题发生的概率；

III. 对于 N 个词语中的每一个词 w_n：

i)选择一个 topic $Z_n\sim\text{Multinomial}(\theta)$；

ii)根据 $p(w_n|z_n,\beta)$ 选择词语 w_n，β 记录各主题条件下生成词语的概率。

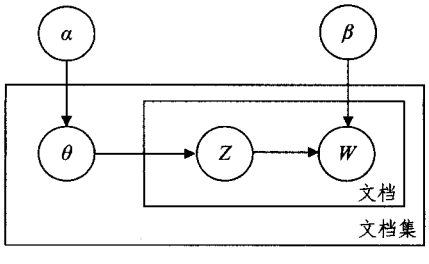

图 1 LDA 概率模型图

2.2 名词性短语识别

LDA 实质上是一种词袋模型(word-bag)，即对词语进行统计建模。然而评价主题的特征词在很多情况下都是专有名词或名词性短语，并非分词后的结果，例如“Cayenne/Toubo/的/舒适/性/功能/以及/行车/辅助/系统/都/非常/的/强大/”，其中“舒适性功能”和“行车辅助系统”都是评价主题词。为此在对文本集合进行 LDA 建模之前，需要识别出具有完整意义的名词性短语。

本文采用启发式规则和共现统计相结合的方法来识别名词性短语。首先对文本集合进行分词和词性标注，根据词性启发式规则抽取出名词性短语的候选词语组合。然后对这些候选词语组合进行共现概率计算，并通过阈值过滤掉部分噪音词语组合。

2.2.1 基于启发式规则的候选名词性短语识别

常用的分词系统无法准确地识别名词性短语，但从分词的结果(见表 1)可以发现，被分割开的专有名词短语在词性标注上体现出一定的规律：这些短语基本上由动词＋名词、名词＋名词、形容词＋名词和数量词＋名词组合而成。由此提出了一个启发性规则：如果名词前出现一个或多个动词、名词、形容词、数量词以及其他英文符号，就将它们组合起来形成候选名词性短语。比如表 1 的第 3 句话中，候选名词性短语有“踩住刹车踏板”、“住刹车踏板”、“刹车踏板”和“启动发动机”。可以看出，候选名词性短语中有不少噪音，因此需要进一步过滤筛选。

表 1 词性标注结果

1. Cayenne/nx Toubo/nx 一直/d 是/v 以/p 马力/n 强劲/ad 著称/v，/w 而/c 这次/r 改装/v 主要/d 针对/p 动力/n 系统/n。/w
2. Cayenne/nx Toubo/nx 的/u 舒适/a 性/ng 功能/n 以及/c 行车/vn 辅助/vn 系统/n 都/d 非常/d 的/u 强大/a 。/\w
3. 踩/v 住/v 刹车/n 踏板/n，/w 启动/v 发动机/n，/w 一切/r 非常/d 平稳/a。/w
4. Cayenne/nx Toubo /nx 的/u 操/n 控/vg 系统/n 显然/ad 就是/d 要/v 让/v 驾驶/v 者/k 很/d 容易/ad 察觉/v 过/u 弯/a 极限/n 所在/n 。/w
5. 最/d 大/a 达到/v 1770/m 升/v 的/u 载/n 物/ng 空间/n，/w 能够/v 满足/v 任何/r 出/v 行/ng 的/u 需要/n 。/w

2.2.2 基于共现概率的短语过滤

基于共现概率的候选名词性短语过滤基于这样一个假设：如果几个词语可以构成一个专有名词性短语，那么这些词语只要在某篇文章中同时出现，它们基本上都是连续出现的。比如上节中的候选名词性短语“刹车踏板”，在汽车评论语料中，只要文章中同时出现“刹车”和“踏板”，很可能这两个词语就是连续出现的。而“住”、“刹车”、“踏板”连续出现的概率很低。为此，候选名词性短语 d 的概率统计公式为：

$$Score(d)=p(w_1,\cdots,w_n)=\frac{ContiDF(w_1,\cdots,w_n)}{CocurDF(w_1,\cdots,w_n)+1}$$

式中，$w_1,\cdots,w_n$ 为候选名词性短语分词后切分成的多个词语，$ContiDF$ 为词语 $w_1,\cdots,w_n$ 在文本集合中连续出现的文档频率，$CocurDF$ 为词语 $w_1,\cdots,w_n$ 在文本集合中同时出现的文档频率。

2.3 评价主题抽取

为了使 LDA 模型得到的主题能够较为准确地描述主体的某些属性，将文本集合中所有句子作为 LDA 的输入，即把每个句子看成是 LDA 的一篇输入文档。通过 LDA 模型将会得到句子与主题、主题与词语两个重要的概率分布，这里的主题即为评价主题。

在主题与词语的概率分布矩阵 φ 中，$\varphi_{i,j}$ 表示词语 i 属于主题 j 的概率。对于一个词语 i 在矩阵中的行向量，仅保留其中概率值最大的元素，将其他元素值赋值为 0，即假设一个词语只属于一个主题。然后从每个主题 j 中抽取出概率较大的前 N 个名词或名词性短语作为该主题的特征词群。

3 评价主题倾向性分析

主题的倾向性是通过统计包含该主题词群的句子的褒贬

倾向性数量得到的。首先利用一种多特征融合的方法识别出句子的倾向性，然后统计出包含主题特征词群的褒贬倾向性句子的数量，并以此作为各个主题倾向性识别结果。由此可以看出，句子倾向性识别是这一阶段的重点，直接决定了评价主题倾向性识别的效果。

很多研究认为，句子的倾向性与句中极性词语的倾向性在多数情况下是保持一致的。然而真实文本中的句子倾向性更为复杂隐蔽。除极性词之外，句子中的标点符号、否定副词、情感指示词、词语搭配等均有可能对句子的倾向性产生重要影响。这里将重点讨论这5类特征，并将它们进行线性融合来衡量句子的倾向性。

(1)极性词：极性词是形成句子倾向性的核心成分，具有稳定的感情色彩，比如“美丽”、“丑陋”等。通过极性词表来判别句子中是否出现极性词，若句子中的第 i 个极性词为褒义，则令 $PL_i=1$；若为贬义，则令 $PL_i=-1$。

(2)否定副词：否定性副词的出现往往会导致句子倾向性的逆转。如果句子中含有否定词，则令 $DEN_i=-1$，否则 $DEN_i=1$。其中 i 表示该否定词与第 i 个极性词的距离最近。

(3)搭配规则：有些词语的感情色彩在一定的规则下会发生变化，比如“他是个实在的人”中的“实在”是形容词，其极性为褒义。而“这事实在干得不怎么样”中的“实在”是副词，极性则为中性。还有一些词和某些动词搭配后会使整个句子体现出一定的褒贬性。比如“这部电影拍得很有意思”中的“意思”和“有”。如果句子中出现这类词且符合搭配规则，则视其为极性词，若为褒义，则令 $PL_i=1$，否则 $PL_i=-1$。

(4)标点符号：一般包含感叹号、问号和省略号的句子具有倾向性的可能性最大，称这3类为情感标点符号。判断句子中是否出现情感标点符号，如果有，则令 $SG=1$，否则 $SG=0$。

(5)情感指示词：在汉语句子中，有很多谓语动词直接暗示了句子的主观性，称这类词语为情感指示词。比如“我认为卡宴的车型还是比较不错的”，其中谓语动词“认为”就预示作者在表达观点。如果句子中出现情感指示词，则令 $INS=1$，否则 $INS=0$。

在得到句子的每个特征之后，通过线性组合的方式给出句子倾向性值 pos 的权重，再利用阈值范围得到句子的褒贬倾向性。其中句子倾向性值的具体计算公式如下：

$$pos=0.8\times\frac{\sum_{i}^{n}PL_i\times DEN_i}{n}+(0.1\times SG+0.1\times INS)\times \operatorname{sgn}\left(\sum_{i}^{n}PL_i\times DEN_i\right)$$

4 实验结果及分析

为了验证本文方法在网络真实评价语料中的有效性，从新车评网[1)]的原创车评板块和腾讯汽车[2)]的车友点评板块中下载了371篇关于保时捷、卡宴的评论性文本(共计2482个句子)作为实验数据集。

4.1 评价主题抽取效果分析

首先对语料进行分词，然后采用启发式规则和共现概率统计相结合的方法对分词结果进行修正，其中共现概率的阈值设置为0.58。共识别出62个关于汽车领域的专业名词性短语，部分结果如表2所列。

表2 专业名词性短语及其共现概率示例表

名词性短语	概率	名词性短语	概率
新 Cayenne	0.97	V8 双涡轮增压发动机	0.84
全景天窗	0.94	刹车踏板	0.84
四驱系统	0.93	保时捷跑车	0.84
Cayenne S	0.93	上代 Cayenne	0.83
增压发动机	0.91	作为跑车世家	0.79
空气悬挂	0.90	不同车型	0.64
双涡轮增压发动机	0.89	全方位智能	0.61
辅助系统	0.89	大块头	0.61
城市 SUV	0.89	动态底盘	0.61
自动变速箱	0.88	得到保障	0.61
松开刹车	0.87	主动式全驱系统	0.61
进气格栅	0.87	不费力气	0.59
混合动力版本	0.86	8档前速自动变速箱	0.58
合金轮圈	0.86	入门款	0.58
燃油直喷技术	0.84	不同车型	0.64

从表2可以看出，此方法可以识别出一些分词系统根本不可能得到的专业名词短语，特别是较长的专业术语，如“动态底盘控制系统”、“燃油直喷技术”等。将上述实验结果的62个名词短语与人工挑选出的57个名词短语进行比较，准确率达到76.6%，召回率达到70.9%，*F-Measure* 值为0.736，说明了该方法对专业名词短语识别的有效性。但结果中也存在一些噪音短语，如“作为跑车世家”、“松开刹车”，这也是启发式规则需要改进的地方。

之后，采用 Gibbs 采样算法[12]对文本集合中的所有句子进行 LDA 建模，LDA 模型的参数设置为 $T=5$，$\alpha=0.1$，$\beta=0.05$，迭代次数为1000次。表3列出了自动抽取的评价主题及部分相关词群。

表3 评价主题及相关词群

评价主题	部分相关名词性词群
外形 & 总体感观	保时捷，Cayenne，造型，卡宴，保时捷跑车，底盘，视觉，车身，全景天窗，玻璃，排气管，车顶，车外，流行趋势，车头，车灯，车型，风格，观感，气息，影响……
油耗 & 动力性能	发动机，燃油直喷技术，双涡轮增压发动机，自动变速箱，汽油版，Cayenne S，机械增压发动机，驱动系统，越野模式，千米，发动机盖，速度，油耗，功率，排量，……
重量 & 空间大小	后排座椅，载物空间，车内，加长，尾厢，前排座椅，储物格，重量，实用性，坐姿，毫米，内距，高度，乘坐舒适性，尺寸，大小，空间，宽敞程度，容量……
内饰 & 车载配置	卡宴 S Hybrid，塑料材质，标配，方向盘，刹车踏板，音响，选装件，英寸，合金轮圈，档案座，入门款，配备，PTV，中控台，组合，感觉，驾驶者……
价格 & 其他	价格，版本，保时捷，老款，上代 Cayenne，进气格栅，城市 SUV，宝马，Cayenne Turbo……

可以看出，利用 LDA 模型能够比较客观地挖掘出文本集合中的评价主题，表3中的主题词群基本上都是相关的，可以反映各个主题到底评论的是汽车哪些方面的特征。另外，LDA 模型会将有紧密联系的多个话题聚集成同一个主题，这也在一定程度上反映了人们讨论问题的一些习惯。如第2个主题的相关词群既有“双涡轮增压发动机”等术语，又涉及到“排量”、“油耗”，这是因为人们一般在谈论汽车发动机的同时，都会习惯性地提及汽车的油耗问题。

1) http://www.xincheping.com/

2) http://auto.qq.com/

表中第5个主题不是很明显，主题词群比较杂乱。这是因为LDA模型比较依赖词语分布情况，它把文本集合中讨论较少，甚至只出现一次的内容都归结成一个主题，这可能与主题数目设置较少有关，但这样可以避免因主题数过多而导致其他主题发散。如何确定效果最好的主题数目是我们需要进一步研究的问题。

4.2 主题倾向性识别效果分析

首先对文本集合中的每个句子进行倾向性识别，倾向性得分的阈值设定为0.2和−0.2，即 *pos* 值大于0.2为褒义，小于−0.2为贬义，两值之间为中性。表4列举了部分句子的倾向性得分以及句中的评价主题词群。

表4 句子倾向性识别结果示例

句子	倾向性得分	评价主题相关词群
我觉得保时捷Cayenne凭借其张力十足的外形和纯正的跑车血统，赢得了众多车迷的青睐。	0.9（褒义）	保时捷 Cayenne
前门壁板的储物格可以充当杯架，也可以容易折叠雨伞等杂物，但后门壁板的载物空间就有点缩水了，实用性不如前门的好。	−0.8（贬义）	储物格 载物空间
原装的4.5升V8双涡轮增压发动机，峰值转矩可以达到450马力，在城市SUV阵营当中已经是数一数二了。	0.8（褒义）	双涡轮增压 发动机
最让人不解的是，居然不提供电动调节方向盘，而是十万元车都能提供的手动调节方向盘。	0（中性）	方向盘

从表4可以看出，多特征融合的方法可以有效地识别大部分句子的倾向性。但是对于不含有任何极性词的句子，此方法还不能识别出来，比如表4中的第4个句子。对于这些情感比较含蓄的句子只能通过准确的语义分析才能够识别，这是自然语言处理的难点。

实验语料的2482个句子的倾向性识别结果中褒义句的正确率为76.2%，召回率为51.7%，而贬义句的正确率为55.4%，召回率为38.2%，褒义句的识别效果也远远高于贬义句。这主要是因为人们批评事物往往习惯于运用隐晦含蓄的表达方式，很少使用感情强烈的负面极性词；另外，反讽、黑色幽默等负面评价方式也是很多人的语言习惯，这就给自动的倾向性识别造成了困难。

得到每个句子的倾向性后，根据特征词群统计出评价主题的倾向性，结果如图2所示。主题的倾向性结果可以直观地给出汽车各方面特征的综合评价信息。

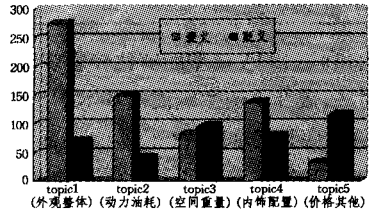

图2 保时捷、卡宴5个评价主题的倾向性统计

结束语 本文应用LDA模型挖掘评论文本集合中潜在的评价主题，并利用多特征融合的方法识别评价主题的褒贬倾向性，给用户提供更直接、更方便的评价信息。通过对网络汽车评论文本语料的实验证实了该方法的有效性。

文中LDA的主题数目是根据语料情况人工设定的，存在一定的局限性。下一步的工作将研究如何将聚类稳定性算法应用到LDA主题模型中，以自动确定主题的数目。句子的倾向性识别只应用了一些特征信息。为了取得更好的效果，需要在统计方法的基础上融入浅层句法分析，从句法结构方面来分析句子的倾向性。

参 考 文 献

[1] 中国互联网络中心(CNNIC). 第26次中国互联网络发展状况统计报告[R]. 北京，2010

[2] 姚天昉，程希文，徐飞玉，等. 文本意见挖掘综述[J]. 中文信息学报，2008，22(3)：71-80

[3] Yi J, Nasukawa T, Bunescu R, et al. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques[C]// Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003). Melbourne, Florida: IEEE Press, 2003: 427-434

[4] Hu M, Liu B. Mining and summarizing customer reviews[C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining (SIGKDD-2004). New York: ACM Press, 2004: 168-177

[5] 娄德成，姚天昉. 汉语句子语义极性分析和观点抽取方法的研究[J]. 计算机应用，2006，26(11)：2622-2925

[6] 王倩，何婷婷，闻彬，等. 基于依存关系的中文情感要素抽取技术研究[C]// 中国计算机语言学研究前沿进展(2007-2009). 北京：清华大学出版社，2009：624-629

[7] Turney, Petter. Thumbs up or thumbs down Semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of 40th Annual meeting of the association for Computational Linguistics (ACL-2002). Philadelphia, Pennsylvania, USA, 2002: 417-424

[8] Kamps J, Marx M, Mokken R J, et al. Using WordNet to measure semantic orientation of adjectives[C]// the 4th International Conference on Language Resources and Evaluation (LREC-2004). Lisbon, 2004: 1115-1118

[9] 朱嫣岚，闵锦，周雅倩，等. 基于HowNet的词汇语义倾向计算[J]. 中文信息学报，2006，20(1)：14-20

[10] Pang B, Lee L, Vaithyanathan S. Thumbs up senti-ment classification using machine learning techniques[C]// the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). Philadelphia, USA, 2002: 79-86

[11] Blei D, Ng A, Jordan M. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022

[12] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences, 2004, 101(1): 5228-5235