基于网络评论的情感分类与观点抽取技术分析与研究

杨 丽
（湖北大学，湖北 武汉 430062）

摘要：互联网技术不断发展背景下，人们重要的获取信息的渠道之一即为网络，针对社会事件、公众人物等，人们意见和评论的发表也在各种网络渠道中进行，在海量评论资源中，通过分析与挖掘，将评论用户的情感倾向性识别出来，从而对用户消费习惯进一步的了解，具有十分重要的现实意义。基于此，文章基于网络评论，分析了情感分类与观点抽取技术。

关键词：网络评论 观点分类 观点抽取技术
中图分类号：TP391.1 文献标识码：A 文章编号：1673-1131（2017）12-0281-02

互联网时代中，网络不仅仅是人们获取信息的来源，更是人们发表观点、意见的主要平台。对于人们的网络评论信息，不同的使用者希望从中挖掘出自己所需的信息，如政府部门希望通过分析与挖掘网络评论，对网络舆情及时、准确的掌握。此种背景下，产生了情感分析技术，如观点抽取技术、文本情感分类技术。现阶段，网络评论信息获取的主要方式为信息自动采集技术，但随着爆炸式的增长评论信息，信息量越来越大，此种技术已经无法良好的进行收集与处理，必须要提出新型的收集与处理技术，以满足收集与处理需求，此种背景下，本文的研究具有十分重要的现实意义。

1 基于网络评论的情感分类

1.1 文本预处理

中文分词，也叫切词，作用是切分一长串汉字序列，使其变为多个单独的词条，之后以一定规范为依据，重组连续的字序列变为词序列。英语中，单词与单词间用空格隔开，因而计算机技术的识别工作较易进行，分词并不需要，但中文词与词之间不存在间隔，需要使用分析，便于计算机技术识别。对于中文断句来说，复杂程度及难度均比较高，歧义识别与新词识别为两大基本问题，较大程度的制约了中文分词的精度[1]。为此，在多层隐马尔可夫模型基础上，中科院计算机技术研究所提出了汉语分词系统ICTCLAS，ICTCLAS具备约500KB/s的单机分词速度，分词精度、词性标注精度分别可达到98.45%、94.63%，因此，本文进行中文分词时，即采用ICTCLAS。处理中文信息时，文本内容构成词汇中通常会包含名词、动词、形容词等多种，文本主要信息由名词等实词体现，有助于划分文本类别，而虚词等在文本类别倾向体现中并无作用，本文将其作为通用词，预处理文本过程中，需要过滤掉其中的通用词，实现方法为设计一个停用词表。

1.2 构建语义词典

第一，基础情感词典。分析文本倾向性时，基础工作为判断词语的极性，为便于构建，本文将情感词划分为褒义词和贬义词两类，褒义词表达正面情感倾向，贬义词表达负面情感倾向。对现有各类情感词典对比之后，本文构建基础情感词典

下不同的状态维修策略：
⑴变压器内部受潮绝缘下降的情形，要进行现场干燥处理。
⑵三相线圈直流电阻严重不平衡时，则要对变压器吊罩的缺陷部位进行故障处理。
⑶绝缘油中溶解的故障特征气体色谱分析异常时，对其进行跟踪分析，对于变压器的内部缺陷要进行吊检处理。
⑷变压器线圈变形测试异常的故障则要进厂大修。
⑸变压器重瓦斯保护动作试验异常时，要进行吊罩检查。
⑹变压器轻瓦斯频繁动作，采集的气体中含有故障特征气体时，则要根据其综合运行状态进行故障分析。

3.2 高压开关的状态维修策略分析

对于高压开关来说，要以六氟化硫气体泄漏量、六氟化硫气体湿度、导电回路直流电阻作为状态维修的依据 还要以真空灭弧室的耐压水平、真空度、导电回路直流电阻作为状态维修的依据；另外，其他的电气、化学、机械性能也是重要的状态维修的判定依据。

在选取高压开关的状态维修的策略中，主要把握以下原则：①操作系统、油开关本体应当依循现场检修的方式和策略。②六氟化硫开关本体的状态维修要采用置换的方式进行检修。③真空灭弧室只选取更换的检修方式。④对于开关本体的大修，要结合操作系统的大修进行处理。

4 结论

综上所述，在我国电网规模日益扩大的背景这下，传统的检修模式和方法显现出迟滞性，为此，应当根据变电设备的实际运行状态，进行基于状态评价和风险评估的状态维修，全面考虑变电设备的安全可靠性、环境、成本等要素，进行状态评价和风险评估和计算，并以变电设备的主变压器和高压开关为重点，确定适宜的状态维修策略和方法，以更好地实现对变电设备的更换、改造、试验和检修处理，较好地避免变电设备状态维修的盲目性。

在我国的电力行业不断扩张的背景下，变电设备的维修工作是极为重要的内容和组成部分，原有的变电设备的周期检修和故障检修存在一定的缺陷和不足，对于电网的安全稳定运行有极大的威胁和干扰，导致变电设备运行维护成本的浪费。为此，要转变传统的检修方式，要建立基于设备状态的状态维修理念和模式，全面而充分地考虑变电设备的实际运行状态，关注变电设备运行中的各个相关要素，从而制定变电设备的维修策略，确保电力系统的安全稳定运行。

参考文献：

[1] 陈庆前.电力系统安全风险评估与应急体系研究[D].华中科技大学,2012.
[2] 郭磊.考虑输电设备可靠性的电网风险分析研究[D].浙江大学,2012.

作者简介：谢蓓敏(1971-) 女，高级工程师，研究方向为自动化控制技术。

时选用的词库为《中文情感词汇本体库》(大连理工大学提供)，之后，再将词库在NTUSD及“情感分析用词语集”辅助下进一步的扩充。完成后，还需对情感词所具备的情感倾向强度做出区分，对于中文情感词，情感强度分为5档，分别为1,3,5,7,9，如果属于褒义词，对应的情感极性值分别为+1，+2，+3，+4，+5，若为贬义词，对应的情感极性值分别为-1，-2，-3，-4，-5。

第二，用户词典。对某个领域文本做出研究时，能够显示该领域用途的部分词语为名词性短语，如酒店领域中的门卡、高级大床房等，对于这类词语，ICTCLAS的切分并不能正确开展，这就需要进行用户词典的构建，以能对这类词语做出识别[2]。实际上，ICTCLAS自身即有办法能够解决该问题，即新词识别方法，将全部评论文本中出现的新词全部统计出来，之后根据词频个数，由高到低的排列这些新词，以3作为阈值，词频如在3以上，人工识别后向用户词典中添加，并将词性标注出来。第三，用户情感词典。形容词和动词为主要集中用户情感的词语，为使用户情感词典进一步扩充，只将这两种词性选取进来，利用ICTCLAS分词结果，把所有评论文本中的形容词及动词统计出来，先对形容词或动词是否在基础情感词典中出现做出判断，未出现时进行添加，再将其极性值计算出来。

第四，否定词词典。否定词可将句子极性改变，而这正是其对句子的主要影响。通常，在评论文本中，情感词的修饰前缀即为否定词，如“他非常不开心，没有获得胜利”，此句子中，“开心”、“胜利”均属于情感词，仅对这两个词统计时，该句子属于正面情感，但由于“不”、“没有”的前缀锈蚀，句子属于负面情感，因此，必须要进行否定词典的构建。

1.3 情感分类

文本分类时，可采用的方法包含两种，一种为基于语义，另一种为基于机器学习，二者成熟度均相对较高，但二者各自均存在一定的缺点，因此，本文结合着两种方法，进行二阶段情感分类，以将这两种分类方法的优点均发挥出来，弥补各自存在的不足。二阶段情感分类中，先进行一阶段的基于语义的分类，将所有文本的情感得分计算出来，挖掘出存在明确情感倾向的文本，再进行二阶段基于机器学习的分类，此步骤中，必须要筛选出情感词，但需要去除可影响情感分类但不具备情感倾向的词语，其具体流程见图1。

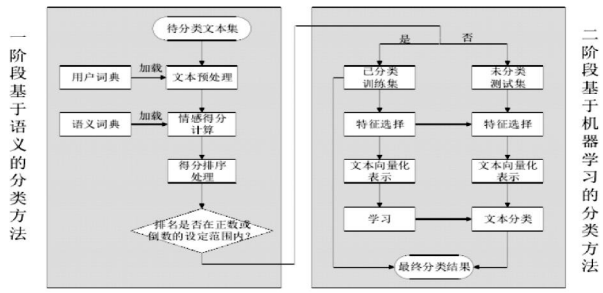

图1 二阶段情感分类

2 基于网络评论的观点抽取技术

对于一个评论文本，为将其中观点属性特征的情感倾向获取，还需抽取中其中可代表观点的情感词，本节以酒店领域的网络评论文本为基础，分析了具体的观点抽取技术。

2.1 观点情感极性算法

潜在用户在查看酒店评论时，不仅希望了解酒店整体的好与坏，更希望从中获取大量的细节信息。观点属性提取方法基础上，观点属性特征词典即可建立之后，之后利于基于语义的情感词分类方法，于一个观点属性或几个观点属性上浓缩整个句子的情感极性值。

依据规定，可采取以下观点情感极值算法：首先，应用ICTCLAS对评论语句进行分词，将分词结果匹配属性特征词典，如果成功，表明属性特征词出现在该语句中，若不成功则表示未描述属性对象。属性特征词找到后，前向与后向窗口均要设置，用于修饰属性特征词情感词的搜索，出现后，情感计算该短语结构(中心为情感词)，获得极性值后，向该属性特征词传递，此结果即为观点抽取情感[3]。在一个评论句子中，观点抽取结果可能会获得多个，但要注意，词语类别为环境卫生时，属性特征词会出现“血迹”、“噪声”这类的，因其自身也属于情感词，所以搜素窗口并不会搜索到，而是将其直接作为情感词。各个评论语句中，属性特征词类别相同情况下，将其极性值累加后，各类别属性极性值即可获得，最后，提取出所有文本的观点属性，累加极性值，求和，各属性类别极性值累加值与其属性词出现次数相除后，属性类别最终得分获得。

2.2 观点抽取流程

观点抽取利用基于语义的方法进行，因此其流程相似于基于语义的情感分类流程，这两个流程的区别为前者需判断一个文本中观点属性出现与否，如果出现，要情感计算观点属性，由于计算仅涉及短语结果，所以不用再特殊处理感叹句与反问句。具体说来，观点抽取流程主要通过4步进行：第一，文本预处理，用户词典加载后，过滤掉文本中的不规则符号，将繁体字转换为简体字，再利用ICTCLAS进行分词，最后将停用词去除；第二，识别观点属性特征词，观点属性特征词典加载，对文本做出判断，如果属性特征词出现在其中，继续进行下面的操作，如果属性特征词并未出现，该文本处理完成，进行下一篇文本的处理；第三，计算观点情感极性，语义词典加载，按照基于语义的短语情感极性值算法，情感打分提取出的观点；第四，观点抽取结果获取[4]。

3 结论

本文中，利用基于语义方法及基于机器学习方法研究了网络评论文本中的情感分类及观点抽取方法，能够较为准确的划分某个领域网络评论中的情感词分类，提取出文本中的观点，应用价值较高，但由于此方面的研究还比较少，本文还存在一定不足，需进行更为深入的研究。

参考文献：

[1] 陈巧红,孙超红,贾宇波.文本数据观点挖掘技术综述[J].工业控制计算机,2017,30(2):94-95+102.

[2] 朱圣代.评价对象、短语、搭配关系抽取及倾向性判断[J].电脑知识与技术,2013,9(9):2044-2045+2047.

[3] 曾伏秋,罗毅辉,杨刚,等.网络论坛教学评论的自动情感分析方法——以湖南商学院枫华论坛为例[J].湖南商学院学报,2013,20(1):106-110.

[4] 张文文,王挺.不规范文本的无监督观点句抽取[J].计算机与数字工程,2013,41(1):64-68.

基金项目：杨丽(跨领域的中文网络评论情感倾向分析与研究)编号：(530/165301301003)

作者简介：杨丽(1985-)，女，山西潞城人，博士，讲师，研究方向：大数据、智能方法等。