

文章编号: 2096-1618(2018)03-0290-06

SAS:用于智能电网安全的移动内生大数据态势感知系统

姜文婷, 林少锐, 廖颖茜

(广东电网有限责任公司电力调度控制中心, 广东 广州 510000)

摘要:智能电网安全管理需要对用户行为进行感知和分析,通过对移动端用户的态势进行感知,从而了解智能电网部署的服务质量。实现移动端用户的态势感知需要对移动端内生大数据进行深入的分析,然而目前相关分析还不多见。构造了一种新的文本情感值计算方法,以一款常用APP为例分析了获取的11万条特定事件的新闻评论情感值,并对此进行统计分析。结果显示,在情感缓和的时间段发生新闻热点时对新闻的评论的情感值波动变化会急剧增加,情感值方差会急剧上升。而当新的新闻发生在另一新闻热点影响时间段内则评论的情感值波动变化趋于平稳乃至呈下降。这些结果对智能电网用户感知具有重要的应用价值,同时也可以用于舆情监控与预警、网络监管与建模。

关键词:移动APP内生数据;大数据分析;舆情态势感知;智能电网安全

中图分类号:TP393.08

文献标志码:A

doi:10.16836/j.cnki.jcuit.2018.03.012

0 引言

随着智能设备的不断增长,以及用户与智能终端的交互日益频繁,网络上越来越多的信息数据产生于移动设备,而且有取代来自网页数据的趋势。相较于传统的由网页端产生的用户交互感知数据,移动终端产生的信息在数量上更加庞大,交互上更加频繁和碎片化,而且信息内容上更加隐私化、个性化。因此,移动终端APP内生数据比起网页数据,更能反映真实的用户心理状态,更有分析的价值。然而,目前对这方面的研究仍然还不成熟。

智能电网的客户端管理需要对用户行为进行感知和分析,电力公司需要分析移动端用户的评价从而指导客户服务。同时,移动APP内生数据的建模与分析研究,通过分析移动APP在移动设备上产生的数据信息,可以为移动APP开发人员提供更好的完善APP的用户体验策略。对于网络数据监管人员,通过分析特定类别移动APP中特定种类信息的变化趋势,可以用于预测网络环境中用户行为的变化规律。

提出一个新的智能电网安全态势感知系统(smart grid security awareness system, SAS),系统可以监测和分析常用APP中有关特定事件评论信息的评论行为,并使用本文提出的一种新型情感分析技术对获取的评论信息进行情感值计算,同时使用一种新型分析方法来分析评论情感值变化规律与新闻信息之间的关系。

创新之处包括:首次针对移动应用内生大数据的分析开展研究;提出并使用新的评论情感值计算方法来分析评论情感值;研究并提出了针对离散频繁碎片化的评论数据进行态势感知的分析方法和相关算法;对移动APP内生用户评论数据的行为进行建模。

1 相关研究工作与关键技术

文本情感分析可以用于用户评论的分析与决策、舆情监控、信息预测等多个实用领域^[1],主要分为基于字典的情感分析技术与基于机器学习的情感分析技术。

杨经等^[2]使用基于SVM的分类方法对中文文本进行喜、怒、哀、乐四种粒度的情感分类,根据结果分析得知,其结果在COAE2009的测评任务中具有一定良好的效果。在刘志明等^[3]的研究中,使用3种不同的机器学习的方案包括SVM方案、NaiveBayes方案和N元语言模型,使用3种不同的特征选取的方案包括IG^[4](信息增益)、CHI统计方案、TF(文档频率)以及3种不同的特征权重计算方案:布尔型特征权重、词频型特征权重以及TF-IDF^[5]对来自微博的数据进行情感分类研究。依据研究结果表明,使用SVM、IG与TF-IDF作为研究方案得到的分类分析结果最好。但是该方案只能对特定领域的分类产生较好的结果,对于不同领域的文本分析需要建立不同的模型。

对于基于语义词典的研究方案,现在大部分研究都是使用基于已有情感词典的方案。熊德兰等^[6]使用基于HowNet的词汇利用相似度算法与语义相似度

收稿日期:2018-03-30

基金项目:国家重点研发计划资助项目(2016YFB0901200);广东电网科技资助项目(036000KK52170002)

算法,语法相似度算法计算出一个句子的褒贬程度,计算结果表明该方案得到的结果与人工判断的结果相近。王振宇等^[7]基于 HowNet 和 PMI 语料库用于计算一个词语的情感极性,这种方案较之一般的方案提高了5%的准确程度。Dong 等^[8]介绍了 HowNet 相关知识,并详细说明了 HowNet 中的各种义原的意义。但是这些方案主要用于分析来自微博的数据,没有应用于分析移动 APP 内生数据,忽略了移动 APP 内生数据的特性与情绪性。

基于上述情感分析方法,使用基于字典的情感分析技术和统计分析技术处理来自移动 APP 中的数据,并使用一种新的计算评论情感值的方案以求获取更加准确的评论情感信息,基于这些数据对移动 APP 内生数据进行进一步的统计分析。

2 关键算法的提出

2.1 数据获取

实验以来自“今日头条”中有关萨德事件相关新闻的评论信息为例。直接获取这些新闻的评论信息是比较复杂的,因此需要首先获取有关的新闻信息。获取新闻信息的方法如下:在 Chrome 浏览器中打开“今日头条”有关萨德事件的新闻浏览页面,打开 Chrome 中的开发者模式。在浏览新闻的同时查看其 Network 部分,获取网页向“今日头条”新闻服务器的 Request URL。在获取了 Request URL 后利用爬虫向“今日头条”新闻服务器端发送大量的请求信息以求获取大量的有关萨德事件的新闻内容。通过解析返回的数据信息获取新闻在“今日头条”新闻服务器上的 groupid。

通过分析来自移动设备向“今日头条”评论服务器发送的数据包,可以判断出一个新闻的评论信息在评论服务器中存储的地址信息只与评论服务器地址以及该评论所属新闻的 groupid 相关。通过利用对应新闻的 groupid 构建面向评论服务器的请求地址,使用数据爬虫可以大量且迅速获取对应新闻下的所有评论信息。

共获取了相关萨德事件评论信息 287873 条,获取到的评论信息主要包括如下字段:createtime,userid,username,numofcomment,replycount,groupid,comment-text。

2.2 数据处理

2.2.1 清除无用数据

由于获取的数据并非全是有用数据且存在部分冗

余数据,因此要对获取的评论数据进行清理操作。具体方法如下:

(1)对获取的所有评论数据,检查是否由一位评论人员在一篇新闻中重复发出。若是,则删除该评论;否,则不作操作。

(2)对获取的所有评论数据,判断是否为广告类信息。通过构建一个小的广告关键字字库,检索所有评论信息,将所有含有关键字字库的评论信息删除即可。

(3)对获取的所有评论数据,判断评论内容是否为空。若为空则将该评论信息及该评论所对应的评论人员信息、点赞数目、回复数目等信息一应取全部删除。

通过上述方法,一共删除了无用信息共 13 万余条,其中大部分为重复信息。

2.2.2 结构化评论内容

基于用户在网络上发表文字的随意性,通过爬虫获取的评论内容在文本组织上是非结构化的,不严格满足汉语语法规范。

在分析了一部分的评论内容后发现,一个句子中(按照中英文的句号,分号,问号,感叹号等对评论进行分割所得到的各个部分)的不同意见句(按照一个句子中以中英文逗号与空格等进行分割所得到的各个部分)在文本组织上大体符合汉语文本组织规范。

基于上述描述,将一条非结构化评论内容转换为结构化意见句的步骤如下:

(1)将评论内容按照中英文的句号,分号,问号,感叹号等进行分割,得到若干句子;

(2)对得到的句子,按照中英文的逗号以及空格进行分割得到若干意见句;

(3)完成对非结构化的评论内容转换为结构化的意见句。

2.2.3 构建情感分析字典

使用基于字典的情感分析技术对获取的评论内容计算其情感值。这里的字典主要包括以下几种:情感词词典,程度副词词典,否定词词典,停用词词典等。

情感词词典、程度副词词典基于知网(HowNet)^[9],通过统计收集出一部分常用的否定词形成否定词词典。使用基于现有(如哈工大,川大等)较成熟的停用词表组合成一张新的停用词表,该表是原多张停用词表的交集。

由于网络社交中新增网络用语速度快,因此原有的情感词词典、程度副词词典无法应对变化迅速的网络用语需要更新情感词词典与程度副词词典。

使用基于 Word2Vec^[10]的方案计算新增词汇的词

向量,并通过与原有各词典中词汇计算其相似度,得到新增词汇的类别与其值。算法如下:

- (1) 将所有的评论内容,新闻文章进行分割;
- (2) 将所有的意见句进行分词处理,并将结果写入一个文本文件中;
- (3) 使用 Word2Vec 方法对得到的文本文件进行训练,得到词向量模型 model;
- (4) 按序排列所获评论内容中所有的分词结果,将出现次小于 1 次的词删除,形成一个评论词典;
- (5) 搜所已有的情感词词典,否定词词典,停用词词典,程度副词词典,将评论词典中出现在这些词典中的词汇删除;
- (6) 对于评论词典中剩余的词汇,通过计算他们与情感词词典,停用词词典中词汇的相似性判断出这些词汇的类别。并得到这些词的值。

一般计算一个词的值的方法有多种^[11],计算新增词汇值情感值的方法如下:

- (1) 使用 model 模型,计算出该词与情感词词典或程度副词词典中某一词最大相似度 P ;
- (2) 该词的值 $W = P \times$ 词典中与该词拥有最大相似度的词的值;

通过该方法,可以得到未知类别的词汇的类别与值。

2.2.4 计算评论情感值

(1) 计算意见句情感值。对于一条评论,可以得到其所有意见句,首先计算其每一条意见句的情感值。每一条意见句都是一个完整的用于表达情感的最小结构。对其进行分词处理,之后按照顺序判断分词后的词是否属于停用词词典、否定词词典、程度副词词典以及情感词词典等。若该词属于停用词词典则将该词从原意见句中删除,剩下的词按照顺序存储。

计算一个意见句的算法如下:

- ① 设定 $W = 1$, $SCORE = 0$;
- ② 对意见句中每一个词进行判断;
- ③ 该词是否是否定词,若是则 $W = W^* - 1$,若不是则进行下一步;
- ④ 该词是否是程度词,若是则 $W = W^*$ 该词程度值,若不是则进行下一步;
- ⑤ 该词是否是情感值,若是则 $SCORE = SCORE + W^*$ 该词情感值, $W = 1$;
- ⑥ 该意见句是否处理完? 若是则该意见句得分为 $SCORE$,若不是则循环②~⑤步。

通过该方案即可计算出一个意见句的情感值得

分。其中使用到的情感词词典、程度副词词典、否定词词典、停用词词典由本文构成并提供。

(2) 计算句子情感值。通过上述步骤可以得到所有意见句的情感值得分。由于中文一般将重要的信息摆放在句子的结尾处^[12-13],通过分析评论中句子的情感倾向与该句子各个意见句情感之间的关系,可以判断出意见句的重要性与在句子中的位置相关,在句子中越靠后的意见句对句子整体情感影响越大。

通过如下计算方法获得一个句子的情感值:

$$SCORE_{\text{sentence}} = \sum_{i=1}^L \frac{i}{L} \times SCORE_{\text{opinionsentence}}$$

i 代表意见句在整个句子中的位置, L 代表该句子的长度,也即该句所包含的意见句的数量。通过该算法可以获得每一个句子的情感值

(3) 计算评论情感值

分析段落中不同位置的句子对段落情感值的影响,得到计算段落情感值的方法:

$$weight = \begin{cases} \frac{L-i}{L} & i < \frac{L}{4} \\ \frac{i}{L} & i \geq \frac{L}{4} \end{cases}$$

$$SCORE_{\text{paragraph}} = \sum_{i=1}^L weight \times SCORE_{\text{sentence}}$$

i 代表这个句子在段落中的位置,从 1 开始。 L 代表段落的长度,通常表示该段落句子的数目。

通过上述方法可计算出每一条评论的情感值信息,共计算得到了 113657 条评论的情感值。

3 系统评价与模型分析

3.1 评论数量长时间下的分布

分析评论行为在较长时间段的特征,统计了 2,3 月份每天评论人数变化趋势,如图 1 和图 2 所示。

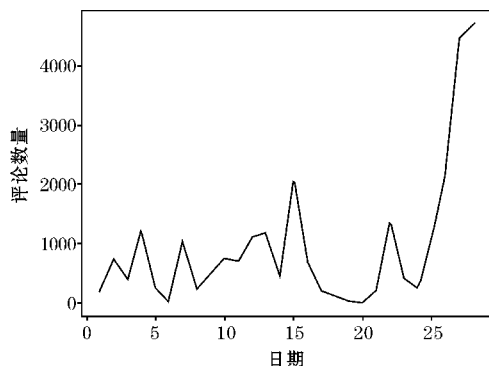


图1 2月份评论人数与时间的关系

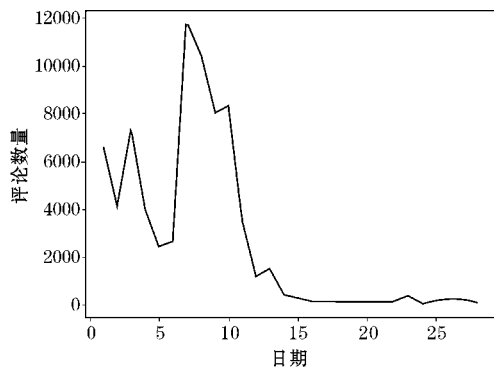


图2 3月份评论人数与时间之间的关系

通过观察2月份与3月份人们对于萨德事件的评论数量波动对比可以发现两个较大的波峰,分别是2月27,28日和3月7日。通过回顾当时的情况可以了解到2017年2月27日乐天董事局决定为部署萨德提供部署用地,2月28日下午乐天中国官网即进入瘫痪状态。2月27日和28日乐天中国官网流量较平时骤然提升至平时的10至25倍。而3月7日上午韩国国防部发布消息,萨德的部分装备将于3月6日通过运输机运抵驻韩美军基地。

从这两个时间可以看出,移动端上的评论人数数量和新闻热点信息的爆发有着一定的关联性,当新闻热点出现,人们在移动端上的评论数量会呈现上升的趋势。由于移动互联网的发展,人们可以及时了解到任何其关注信息的最新动向。通过人数分布规律可以得出:移动端中新闻的评论人数总是随着新闻热点的产生而增加,随着新闻热度的下降而减少。

3.2 评论数量一天之内的分布

分析评论数量24 h的分布规律,如图3所示。

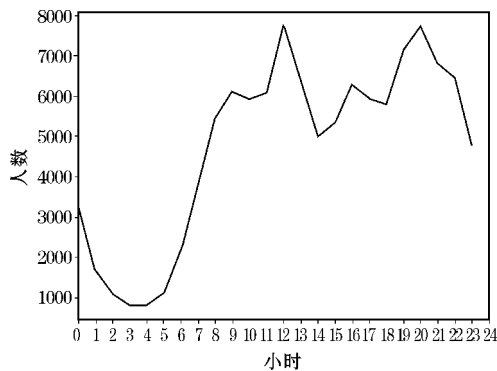


图3 一天之内评论人数与时间的关系

由图3所示,评论数量在一天24 h内的分布规律是一个非常明显的波动曲线。在当天21时至第二天4时,评论数量随着时间的流逝而逐渐降低。可能是因为人们在这段时间内逐步的进入到休息时间从而减

少社交类活动。4-8时,评论的数量逐步上升,人们睡醒早起之后习惯拿起移动设备观看当天最新的新闻事件,并对一些感兴趣或者有争议的事件发表自己的意见或建议,可以看出移动互联网的发展已经开始逐渐影响与改变人们接收新事物与信息的方式。9-11时评论数量比较平稳,可能是因为此间大部分人员都在上班工作,无暇进行评论,而有时间评论的人员会继续评论。11-12时以及下午14-17时与19-21时的评论数量上升阶段代表用户这段时间观看新闻的兴趣逐步增加。而12-14时与17-18时评论数量下降可能是因为午休时间不会过度关注手机而晚餐时间对于新闻热点的关注度没有其他时间那么高。由此可以看出,人们的评论热情与其生活时间分布息息相关。通过上述分析,发现舆论传播集中的时段,便于有效地控制舆论传播效力。

3.3 点赞与回复数量的关系

点赞与回复数量之间的关系研究有利于了解评论用户的行为特征。统计了相关11万条评论数据的点赞数目以及其回复数目之间的点状图,如图4所示。

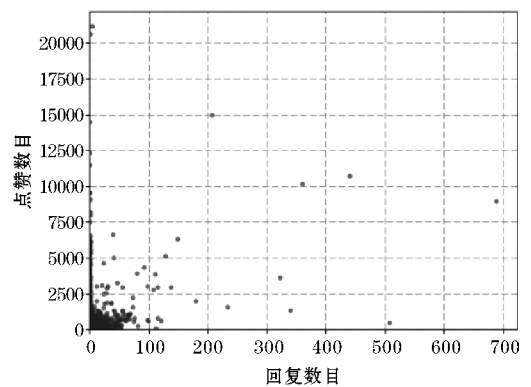


图4 点赞数目与回复数目的关系

图4中,每个点都代表一个评论,横坐标代表评论回复的数量,纵坐标代表评论的点赞数量。统计可以判断出绝大部分评论的点赞与回复数目都集中在点赞数目小于2000,且回复数目小于50这一区域内。即对于大部分的评论而言,点赞行为与回复行为满足一定区域内的分布(点赞数目小于2000且回复数目小于50)。而超出该区域的点赞与回复行为之间则呈现一种离散分布的现象,也可以认为对于拥有特别多点赞数量的评论而言,该评论的回复数目是不可预知的。因此,评论的点赞行为与回复行为在小范围区域内(点赞数目小于2000且回复数目小于50)是可以预知与分析的,而超过一定区域则不可预知,呈现一种离散分布状态。

3.4 评论得分与点赞(或回复)间的关系

点赞人数与评论得分情况的散点图以及回复人数与评论得分情况如图 5 和图 6 所示。

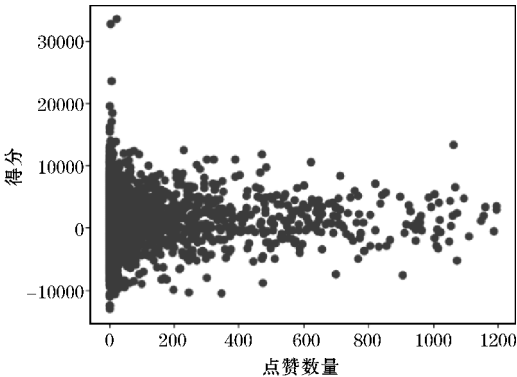


图 5 评论得分与点赞数量之间的关系

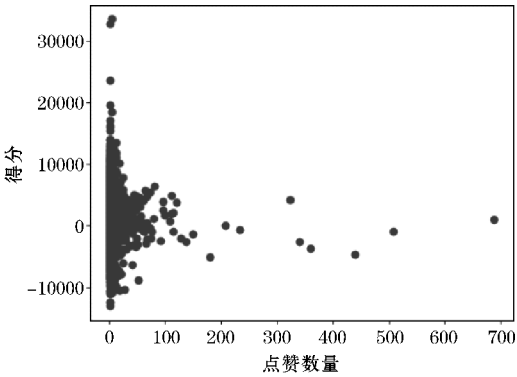


图 6 评论得分与回复数量之间的关系

由于评论得分过小在图片中不易显示,因此在图 5 中将评论得分扩大 1000 倍,在图 6 中将得分状况扩大 100 倍,可以发现回复数量与点赞数量的增加,评论得分基本趋于 0,即获得较高点赞数目与回复数目的评论一般而言是比较中立的。且对于回复数目与点赞数目较低的评论而言,其得分状况分布广阔,可以认为比较极端或者无意义的评论(有时无意义的评论会被判断为中立性评论)很少有人会去点赞与回复,即点赞与回复行为需要成本,并非所有的评论都会有点赞行为的出现。

对比图 5、图 6 可以看出点赞行为要比回复行为更加活跃,人们有时更热衷于点赞某些评论而不是回复某些评论。由此可以得出虽然点赞行为与回复行为都需要一定成本,但是回复行为所花费的成本要比点赞行为的更高。因此对于具有大量回复行为的评论信息应该受到人们的大量关注。

3.5 评分结果分析

对于获取的评论,为了分析其情感值计算结果的准确性,随机选择了 100 条评论信息,得到使用方案对

这些评论信息的打分结果。在此假设,对于分数大于-0.3 并且小于 1 的评论信息属于中立信息,分数大于 1 的评论内容的情感倾向为正向,情感值小于-0.3 的评论的情感倾向为负面评论。通过人工检测判断这些评论信息的情感倾向。两次判断得到的结果如表 1 所示。

表 1 使用计算机与人工分别判断 100 条评论情感倾向的结果

	正向	负向	中立
计算机识别	34	48	18
人工识别	33	50	17

通过表 1 可知,使用文内方法所计算的一段评论的结果大致上是合理的。可通过计算上述结果的正确率 P ,召回率 R 和 F 值来检测所提出计算情感值方案的正确性。

正确率 $P=A/B$,召回率 $R=A/C$, F 值 $F=2PR/(P+R)$ ^[14-15]。其中 A 为人工标注和计算机标注都为正向(负向或中立)的评论数量, B 为计算机标注为正向(负向或中立)的评论数量, C 为人工标注为正向(负向或中立)的评论数量。

经统计后得到结果如表 2 所示。

表 2 评论结果的正确率、召回率和 F 值/%

句子极性	P	R	F
正向	28/34=82.35	28/33=84.85	83.58
负向	45/48=93.75	45/50=90	91.84
中立	14/18=77.78	14/17=82.35	80

通过表 2 可以发现计算评论情感值的方案是合理的,该方法对于负面评论有着比正向和中性倾向的评论更高的检测效率,对于探寻负面评论的准确度有着更好的效果。

4 结论

移动应用内生数据具有重要的分析价值,但目前较少有人分析。针对这一问题,分析了一批来自特定 APP 中有关特定事件的评论内容,就该评论内容以及其评论相关的点赞数目、回复数目、评论发表时间等进行统计分析,对评论的情感值计算提出了一种新的计算方案。从实验结果可得出如下结论:评论数量在时间上的变化与新闻的产生相关;一天之内评论数量的变化与人们的作息工作时间相关;评论的点赞回复数目在小范围内存在关系而在大范围尺度内关系不明显;当热点发生在已经缓和的时间段时,评论的情感值

波动变化会急剧增加;当热点发生在另一热点影响时间之内则评论的情感值波动变化趋于平稳乃至呈下降趋势。该系统的分析结果对智能电网用户感知具有重要的应用,同时也具有一定的通用性,例如也可以用于舆情监控与预警,网络监管与建模等公共领域。

参考文献:

- [1] 赵妍妍,秦兵,刘挺.文本情感分析[J].软件学报,2010,21(8):1834-1848.
- [2] 杨经,林世平.基于SVM的文本词句情感分析[J].计算机应用与软件,2011,28(9):225-228.
- [3] 刘志明,刘鲁.基于机器学习的中文微博情感分类实证研究[J].计算机工程与应用,2012,48(1):1-4.
- [4] 石慧,贾代平,苗培.基于词频信息的改进信息增益文本特征选择算法[J].计算机应用,2014,34(11):3279-3282.
- [5] Wu Ho Chung, Luk, Robert Wing Pong, Wong, Kam Fai, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. ACM Transactions on Information Systems, 2008, 26(3):1-37.
- [6] 熊德兰,程菊明,田胜利.基于HowNet的句子褒贬倾向性研究[J].计算机工程与应用,2008,44(22):143-145.
- [7] 王振宇,吴泽衡,胡方涛.基于HowNet和PMI的词语情感极性计算[J].计算机工程,2012,38(15):187-189.
- [8] Dong Z, Dong Q. HowNet-a hybrid language and knowledge resource [C]. International Conference on Natural Language Processing and Knowledge Engineering Proceedings, 2003.
- [9] 杨小平,马奇凤,余力,等.评论簇在网络舆论中的情感倾向代表性研究[J].现代图书情报技术,2016,32(7):51-59.
- [10] Xue B, Fu C, Zhan S. A Study on Sentiment Computing and Classification of SinaWeibo with Word2vec [C]. IEEE International Congress on Big Data, 2014:358-363.
- [11] 赵文清,侯小可,沙海虹.语义规则在微博热点话题情感分析中的应用[J].智能系统学报,2014,9(1):121-125.
- [12] 王文,王树锋,李洪华.基于文本语义和表情倾向的微博情感分析方法[J].南京理工大学学报(自然科学版),2014,38(6):733-738.
- [13] 王雍凯,毛存礼,余正涛,等.基于图的新闻事件主题句抽取方法[J].南京理工大学学报(自然科学版),2016,40(4):438-443.
- [14] 陈忆金,曹树金,陈桂鸿.网络舆情意见挖掘:用户评论情感倾向分析研究[J].图书情报知识,2013(6):90-96.
- [15] 周胜臣,瞿文婷,石英子,等.中文微博情感分析研究综述[J].计算机应用与软件,2013,30(3):161-164.

SAS: a System Via Situation Awareness on Endogenous Big Data for Smart Grid Security

JIANG Wen-ting, LIN Shao-rui, LIAO Ying-qian

(Guangdong Power Grid Corporation, Guangzhou 510000, China)

Abstract: In the security management of Smart grid, user behaviors should be sensed and analyzed, which is always achieved by situation awareness on client and helps for improving QoS. To achieve the situational awareness of mobile end users, it is necessary to carry out in-depth analysis on the endogenous big data of mobile terminal, at present, however, relevant analysis is rare. In this paper, a new method for calculating the emotional value of text is constructed, and a common APP is used as an example to analyze the emotional value of the news commentary on the 110000 specific events obtained, and the statistical analysis is carried out. The results show that the emotional value fluctuation of the news will increase sharply when the news hot spot occurs during the period of emotional relaxation, and the variance of emotional value will increase sharply. However, when the new news happened in another news hot spot, the emotional value fluctuation of the comments tended to be stable and even declined. These system and research results provide valuable information for networking situation awareness on popular opinions and network event supervision or modeling.

Keywords: mobile APP data; big data analysis; situation awareness; smart grid security