基于爬虫和 TFIDF-NB 算法的微博情感分析*

杨 戈[1,2]，杨麓涛[1]
(1.北京师范大学珠海分校 智能多媒体技术重点实验室，广东 珠海 519087；
2.北京大学深圳研究生院 深圳物联网智能感知技术工程实验室，广东 深圳 518055)

摘 要：针对微博网络舆情信息量大、无规则、随机变化的特点，提出 TFIDF-NB(Term Frequency Inverse Document Frequency-Naive Bayes)用于微博情感分析，设计与实现了一个基于 Scrapy 框架的微博评论爬虫，将某热点事件的若干条微博评论进行爬取并存进数据库，然后进行文本分割、LDA(Latent Dirichlet Allocation)主题聚类，最后使用 TFIDF-NB 算法进行情感分类。实验结果表明，TFIDF-NB 算法平均准确率高于线性支持向量机算法和 K 近邻算法，在精确率和召回率方面高于 K 近邻算法，具有较好的情感分类效果。

关键词：微博舆情；网络爬虫；情感分类
中图分类号：TN011；TP391.41　　**文献标识码**：A　　DOI：10.16157/j.issn.0258-7998.200748

中文引用格式：杨戈，杨麓涛. 基于爬虫和 TFIDF-NB 算法的微博情感分析[J].电子技术应用，2021，47(4)：59-62，66.
英文引用格式：Yang Ge，Yang Lutao. Sentiment analysis of Weibo based on TFIDF-NB algorithm[J]. Application of Electronic Technique，2021，47(4)：59-62，66.

Sentiment analysis of Weibo based on TFIDF-NB algorithm

Yang Ge[1,2]，Yang Lutao[1]
(1.Key Laboratory of Intelligent Multimedia Technology，Beijing Normal University(Zhuhai Campus)，Zhuhai 519087，China；
2.Engineering Lab on Intelligent Perception for Internet of Things(ELIP)，Shenzhen Graduate School，Peking University，Shenzhen 518055，China)

Abstract：In view of the large amount of public opinion information on Weibo，irregular and random changes，this paper proposes a Weibo sentiment analysis method based on TFIDF-NB (Term Frequency Inverse Document Frequency-Naive Bayes) algorithm. By coding a Weibo comment crawler based on the Scrapy framework，several Weibo comments on a hot event are crawled and stored in the database. Then text segmentation and LDA (Latent Dirichlet Allocation) topic clustering are performed. And finally the TFIDF-NB algorithm is used for sentiment classification. Experimental results show that the accuracy of the algorithm is higher than that of the standard linear Support Vector Machine algorithm and the K-Nearest Neighbor algorithm，and it is higher than the K-Nearest Neighbor algorithm in terms of accuracy and recall，and it has a better effect on sentiment classification.
Key words：Weibo public opinion；web crawler；sentiment classification

0 引言

网络舆情是指网络用户对社会各方面热点问题所发表的见解和建议的舆论，是社会舆情的一种体现，是公众对社会中各种热点事件和问题所表达的态度、想法、情绪等的集合。互联网的快速发展使得网络舆情的形成和传播速度不断提升，对社会的影响巨大。

文献[1]证明了网络舆情的发展具有混沌的特性，即表现为乱序、无规则、随机变化。在网络舆情传播的过程中，微博给网络舆情的形成、发酵和传播提供了一个强大的互联网平台，给其用户提供了一个向全世界分享信息、发表评论和表达诉求的平台，这些舆论内容在短时间内会大规模地扩散，甚至会影响事件的走向。

本文首先实现一个基于 Scrapy 框架的微博评论爬虫，将某热点事件的若干条微博评论进行爬取并存进数据库，然后进行文本分割和 LDA(Latent Dirichlet Allocation)主题聚类，最后采用 TFIDF-NB(Term Frequency Inverse Document Frequency-Navie Bayes)算法进行文本情感分类。

(1)爬虫

爬虫全称为网络爬虫，是一种可以对互联网上的信息进行自动化浏览的网络脚本或程序，可实现对海量互

* 基金项目：广东高校省级重大科研项目(2018KTSCX288，2019KZDXM015，2020ZDZX3058)；广东省学科建设专项(2013WYXM0122)；智能多媒体技术重点实验室(201762005)；北京师范大学珠海分校 2019 校级“质量工程”课程思政项目(201932)

联网信息进行浏览、爬取等操作，并将抓取到的信息存储于本地中。

网络爬虫可以分为4种[2]：通用网络爬虫[3]、主题网络爬虫[4]、增量式网络爬虫[5]、深层网络爬虫[6-7]。

(2)情感分类

情感分析是指识别文本中潜在的想法、情感和态度的方法[8]。情感分类是情感分析的核心内容，情感分类的作用是识别文本数据中的观点，对情感的积极或消极情绪进行分类[9]。

目前情感分类主要有两种方法，一种是基于词典的方法[10-13]，另一种是基于机器学习的方法[14-16]。

1 微博情感分析

基于Scrapy框架的微博评论爬虫(Weibo Commit Crawler based on Scrapy)开始运行，检测微博内容存储数据库是否存在，如果数据库不存在，程序会先新建一个新的数据库。创建数据库结束后，检测存储数据表是否存在，如果数据表不存在，程序会新建一个新的数据表。当程序确认数据库和数据表均已存在后，开始依次构造每一页的评论的请求，并依次遍历每一页的评论文本数据，当遍历页面达到程序设定要求时，即进行数据的清洗，否则仍继续遍历。数据清洗结束后，将数据存储进数据库中，并关闭数据库，爬虫程序至此运行结束。

网络爬虫运行的基本流程主要分为以下步骤：(1)网页获取；(2)网页解析；(3)数据存储。网络爬虫的原理是进行浏览器的模拟发送HTTP请求，爬虫程序通HTTP请求向网页Web服务器发送请求，获取服务器端的响应后对网页进行下载，接着完成爬虫的爬取工作。网页解析是对网页进行去噪的操作，现今的网页大多数是以HTML的格式存在，网页去噪就是对网页中所需要的内容进行提取，网络爬虫在对网页内容进行提取时，需要分析网页HTML结果，从而对有用信息进行提取。

1.1 文本分割

微博评论爬虫所爬取的文本数据以语句为主，为了更好地进行分析，需对微博评论文本进行分割，将评论从语句转变成为词语，即分词：将语句文本切割成以词语为单位。本文的文本分割技术使用了开源的一个Python第三方库：jieba。jieba库在如今的中文分词领域中是较好的第三方库，为了实现更好的分割效果，本文使用分词精确模式。精确模式是将文本以最准确的方式进行切分，适用于情感分析等精度要求较高的分析。为了提升分割准确率，本文先进行停用词(Stop Words)的设置。停用词设置是一个信息检索过程中常用的手段，可使用户节约硬件和时间成本，提升分割准确率，在自然处理语言文本的过程中会过滤某些无用字词。为了使某些较为新或者复杂的词被误分割，因此本文在分割前设置了停用词，通过使用腾讯、搜狗、盘古等输入法公司的词库建成停用词集，以此提升精确度。

1.2 LDA主题聚类

LDA是一个三层贝叶斯主题的无监督机器学习算法，其目的是发现文本中隐藏的主题信息，在无标注的情况下在文本中发现隐性的语义维度，从而整理出文本的主题。隐性语义实质上就是利用文本中词语的共同特征来发现文本的主题分类结构。

LDA模型生成文档方式如下：(1)按先验概率$P(d_i)$选择一篇文档d_i；(2)从狄利克雷分布α中取样生成文档d_i的主题分布θ_i，即主题的分布θ_i由超参数α的狄利克雷分布生成；(3)从主题的多项式分布θ_i取样生成文档d_i第j个词的主题$Z_{i,j}$；(4)从狄利克雷分布生成β取样生成主题$Z_{i,j}$对应的词语分布$\phi_{Z_{i,j}}$，即词语的分布$\phi_{Z_{i,j}}$由超参数β的狄利克雷分布生成；(5)从词语的多项式分布$\phi_{Z_{i,j}}$中采样最终生成词语$\omega_{i,j}$。

1.3 基于TFIDF-NB的情感分类算法

1.3.1 TF-IDF算法

TF-IDF(Term Frequency - Inverse Document Frequency)是一种用于信息检索、文本挖掘的算法。TF-IDF算法作为一种统计学的检索算法，可以用于评估一个字或一个词在一份文本或者语料集中的重要性。这个算法由两部分组成，分别是词频和逆向文件概率，一个词在一篇文章的重要性与出现频率成正相关关系，但也会随着它在语料库中出现的频率成负相关关系。其核心思想是：若一个词语在一份文本中出现的频率很高，但在其他文章或者语料集中出现的频率较低，则该词语拥有较好的代表性，适合作为主题词。

词频(Term Frequency)：表示词语在某一文本中的出现频率，如式(1)所示，其中$T_{i,j}$代表词语i在文件d_j中的词频，n代表词语，$n_{i,j}$代表词语i在文件d_j中出现的次数，分母表示文件d_j所有词语的总数。

$$T_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

逆向文件频率(Inverse Document Frequency)：某一词语的逆向文件频率，可以通过总文件数目除以包含该词语文件的数目，再把商的结果取对数即可得到，如式(2)所示，其中I_i代表词语i的IDF，$|D|$是语料集中的总文件数目，$|\{j:n_i\in d_j\}|$是含有该词语的文件数目。当该词语不存在于语料集中时，会导致分母为0，从数学理论上来说是不可以进行对数运算的，故一般情况下使用$|\{j:n_i\in d_j\}|+1$。

$$I_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|+1} \tag{2}$$

所以一个词语用TF-IDF算法表示为：

$$\text{TF-IDF}=T_{i,j}\cdot I_i \tag{3}$$

本文通过TF-IDF算法对微博爬取数据进行一个关键词提取，由于微博评论内容通常包括有一些与主题无关的文本，如用户名、特殊符号如“@”、表情或一些无实

际意义的文本，通过该算法可以实现对这些非重要数据进行过滤和排除，优化数据。

1.3.2 朴素贝叶斯算法

朴素贝叶斯分类算法的原理是，先通过已经给定的训练集，假设特征词之间相互独立，对从输入到输出的联合概率分布进行机器学习，在输出模型的基础上，输入 X 从而求出使其后验概率最大的输出 Y。

设有一样本数据集表示为 $D=\{d_1,d_2,\cdots,d_n\}$，对应样本数据的特征属性集表示为 $X=\{x_1,x_2,\cdots,x_d\}$，类变量表示为 $Y=\{y_1,y_2,\cdots,y_m\}$，即 D 可以分为 y_m 类别。其中假设 $x_1,x_2,\cdots,x_d$ 相互独立且随机，则 Y 的先验概率 $P_{\text{poior}}=P(Y)$，Y 的后验概率 $P_{\text{post}}=P(Y|X)$，由朴素贝叶斯算法可求得，后验概率可以由先验概率 $P_{\text{poior}}=P(Y)$、证据 $P(X)$、类条件概率 $P(Y|X)$ 计算出，如式(4)所示：

$$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)} \tag{4}$$

朴素贝叶斯算法在基于各特征之间相互独立的假设前提下，在给定类别为 y 的情况下，可推出式(5)：

$$P(X|Y=y)=\prod_{i=1}^{d}p(x_i|Y=y) \tag{5}$$

联立式(4)、式(5)可得后验概率，得式(6)：

$$P_{\text{post}}=P(Y|X)=\frac{P(Y)\prod_{i=1}^{d}p(x_i|Y)}{P(X)} \tag{6}$$

由于 $P(X)$ 大小不变，故在比较后验概率时只比较式(6)的分子部分即可。因此可以得到一个样本数据属于类别 y_i 的朴素贝叶斯计算，如式(7)所示：

$$P(y_i|x_1,x_2,\cdots,x_d)=\frac{P(y_i)\prod_{j=1}^{d}p(x_j|y_i)}{\prod_{j=1}^{d}p(x_j)} \tag{7}$$

本算法设定，当 $P(y_i|x_1,x_2,\cdots,x_d)$ 的值小于 0.4 时，表示为消极情绪；当值为大于等于 0.4、小于 0.7 时，表示为中立情绪；当值大于等于 0.7 时，表示为积极情绪。

2 实验结果

2.1 实验环境

本文的实验机器为 Apple MacBook Air，处理器为 1.6 GHz 双核 Intel Core i5，RAM 为8 GB，操作系统为 macOS Catalina 10.15.2，IDE 为 PyCharm 2019.1.3，Python 版本为 Python3。

本文以新浪微博平台一未上映超级英雄电影《上气》引发争议为例，其人物刻画在微博引发了巨大争议，支持和反对声此起彼伏。本文选取此事件为例，通过利用基于 Scrapy 框架技术编写实现的微博评论主题爬虫，爬取了新浪微博平台 10 条微博新闻媒体及自媒体博主的微博下共计 3 500 余条评论作为实验数据集，以进行后续的实验。

2.2 运行结果

2.2.1 数据统计可视化

通过对切割下来的文本进行数据的可视化与统计，可以快速地确认微博评论文本中出现的高频词汇，从而初步确定微博舆情的主题和方向。本文所采用的可视化方法为词云和高频词语柱状图。

高频词语柱状图即以出现频率最高的词语作为计算参考值，取值为 1。然后词频相对较低的词语按词频从大到小依次排列，以柱状图的形式给人以直观的对词频的了解。

通过文本分割技术将微博评论爬虫所爬取的评论数据将所有语句分割成约 15 000 个词语，通过词云和高频词语柱状图进行数据可视化统计，以初步确定舆情主题和关键词信息。高频词语柱状图结果如图 1 所示。

2.2.2 LDA 主题聚类

本文通过 LDA 主题生成模型，对文本分割后的15 000

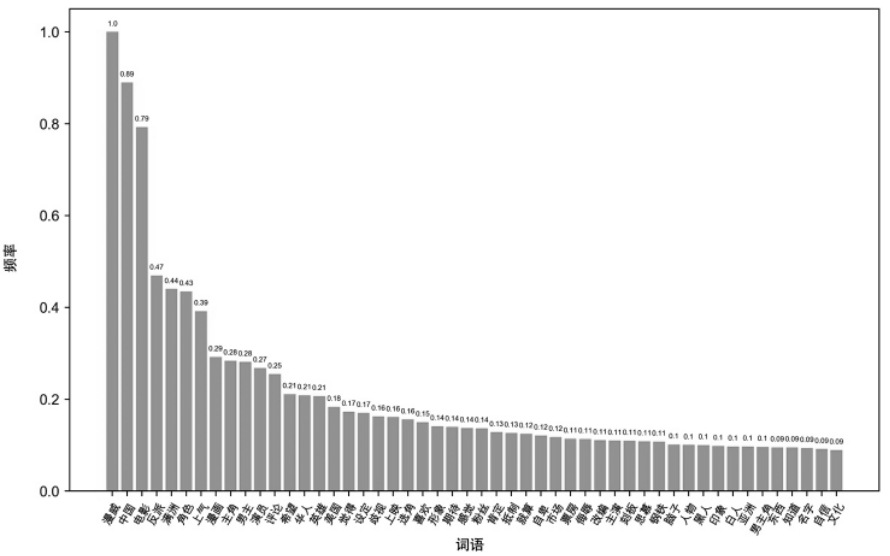

图 1 高频词语柱状图结果

余词语进行主题聚类，在无标注的情况下在文本中发现隐性的语义维度，从而整理出文本的关键词和主题。隐性语义实质上就是利用文本中词语的共同特征来发现文本的主题分类结构。并通过人工分析和总结出主题信息，用所爬取的微博实际评论作为主题代表案例。最终LDA主题聚类结果整理如表1所示。

表1 LDA主题聚类结果

主题	关键词	评论案例
主题1:电影涉嫌侮辱和歧视，应该反对和抵制	满大人，傅满洲，设定，抵制，侮辱	“我觉得身为一个中国人，应该要拒绝这部电影任何涉及到傅满洲的影视作品都应该受到中国人的抵制。”
主题2：对于中国演员LCW出演反派角色表示不满和反对	LCW，反派，演员，选角，晚节不保	“LCW是疯了吗！这可是傅满洲啊啊啊！当初听说要拍上气就觉得满满的恶意某些人能不能别洗了什么满大人不是傅满洲脑子是个好东西。”
主题3:对男女主角选角表示不满	男主，女主，丑，显老，亚洲脸，香蕉人	“不看！男主女主选的什么玩意？美国人你们什么审美？好歹主角选个像李小龙的也行啊？？？男主还没我长得帅呢！信不信，中国票房细碎！从我做起…这部不看…”

2.2.3 情感分类

本文将情感极性分析结果分为以下3类：积极的、消极的、中立的，其中本话题下中立情感词占49.0%，积极情感词占20.6%，消极情感词站30.4%，可以看出消极负面类情绪词汇在本话题舆情中占主体地位。

2.3 TFIDF-NB算法对比

2.3.1 模型训练

本文通过收集如腾讯、盘古、搜狗等互联网输入法公司的词库，制作了正、负极性词典共计9 000余词汇，作为TFIDF-NB算法的训练样本，通过前期训练以达到更高的准确率。文本情感分类领域中常使用准确率(Accuracy)、精确率(Precise)、召回率(Recall)、F_1度量值(F1 measure)作为评价一个模型或者算法性能的关键指标[17]。

2.3.2 对比结果

通过使用本文提出的TFIDF-NB算法对所爬取的微博内容进行情感分类，并与其他情感分类模型算法模型进行比较。经多次实验结果对比后取平均值，结果如表2所示。

表2 算法结果比较

算法	准确率/%	精确率/%	召回率/%	F_1度量值/%
TFIDF-NB	81	72	73	72
标准线性支持向量机[18]	75	75	74	74
K近邻[10]	68	68	68	69

3 结论

本文通过实现一个基于Scrapy框架的微博评论爬虫，对热点话题微博评论进行爬取存储，并通过文本分割、数据可视化统计、LDA主题聚类对舆情进行分析，可以有效地对舆情进行分析和预判。此外，本文提出的基于机器学习的TFIDF-NB情感分类算法经过模型训练后，可以对微博评论情感进行有效的分类，其准确率高于传统的支持向量机和K近邻算法，在精确率和召回率方面高于K近邻算法，具有较好的情感分类效果。

参考文献

[1] 魏德志，陈福集，郑小雪.基于混沌理论和改进径向基函数神经网络的网络舆情预测方法[J].物理学报，2015，64(11)：52-59.
[2] 潘晓英，陈柳，余慧敏，等.主题爬虫技术研究综述[J].计算机应用研究，2020，37(4)：961-965，972.
[3] 方美玉，郑小林，陈德人，等.商品评论聚焦爬虫算法设计与实现[J].吉林大学学报(工学版)，2012，42(S1)：377-381.
[4] 张莉婧，曾庆涛，李业丽，等.面向图书主题的爬虫算法研究[J].计算机科学，2017，44(S2)：460-463，469.
[5] 孟涛，王继民，闫宏飞.网页变化与增量搜集技术[J].软件学报，2006(5)：1051-1067.
[6] 郑冬冬，赵朋朋，崔志明.Deep Web爬虫研究与设计[J].清华大学学报(自然科学版)，2005(S1)：1896-1902.
[7] 侯东阳，武昊，王军锋，等.基于深层网络爬虫的Web地图服务发现方法[J].地理与地理信息科学，2015，31(5)：10-13，19.
[8] MEDHAT W，HASSAN A，KORASHY H.Sentiment analysis algorithms and applications：a survey[J].Ain Shams Engineering Journal，2014，5(4)：1093-1113.
[9] 赵妍妍，秦兵，刘挺.文本情感分析[J].软件学报，2010，21(8)：1834-1848.
[10] 李然，林政，林海伦，等.文本情绪分析综述[J].计算机研究与发展，2018，55(1)：30-52.
[11] 张林，钱冠群，樊卫国，等.轻型评论的情感分析研究[J].软件学报，2014，25(12)：2790-2807.
[12] 刘德喜，聂建云，万常选，等.基于分类的微博新情感词抽取方法和特征分析[J].计算机学报，2018，41(7)：1574-1597.
[13] 栗雨晴，礼欣，韩煦，等.基于双语词典的微博多类情感分析方法[J].电子学报，2016，44(9)：2068-2073.
[14] 陈龙，管子玉，何金红，等.情感分类研究进展[J].计算机研究与发展，2017，54(6)：1150-1170.
[15] 刘思叶，田原，冯雨宁，等.游客微博主题情感分析方法比较研究[J].北京大学学报(自然科学版)，2018，54(4)：687-692.
[16] 孙念，李玉强，刘爱华，等.基于松散条件下协同学习的中文微博情感分析[J].浙江大学学报(工学版)，2018，52(8)：1452-1460.
[17] 段吉东，刘双荣，马坤，等.基于集成学习的文本

(下转第66页)

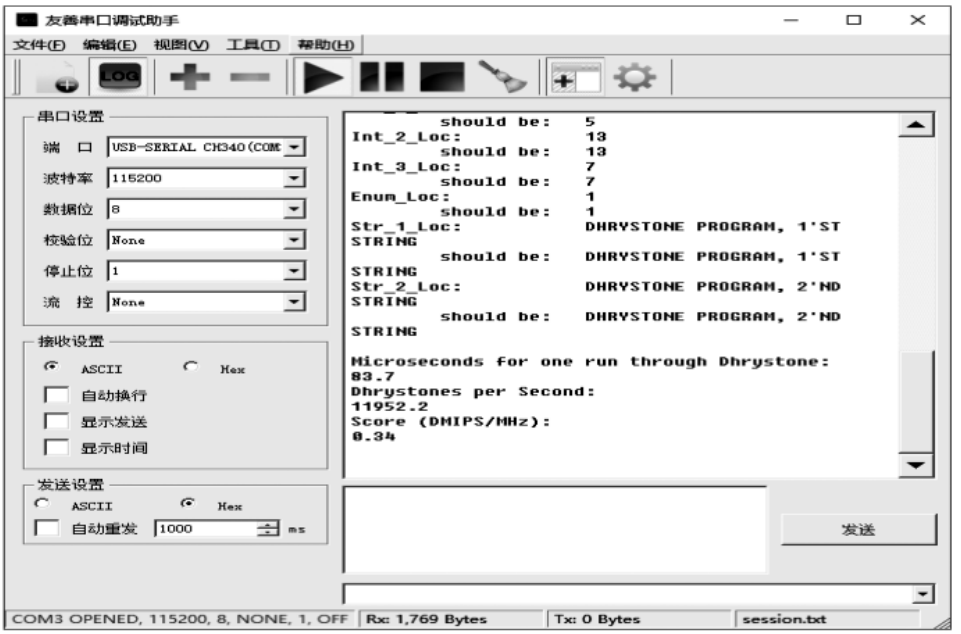

图 7 benchmark 跑分结果

Utilization Post-Synthesis | Post-Implementation

Graph | Table

Resource	Utilization	Available	Utilization %
LUT	24059	133800	17.98
LUTRAM	22	46200	0.05
FF	9320	267600	3.48
BRAM	284.50	365	77.95
DSP	10	740	1.35
IO	248	285	87.02
BUFG	2	32	6.25

图 8 资源使用情况

419-422.

[3] 李安新.BP 神经网络研究与硬件实现[D].青岛:山东科技大学,2010.

[4] 陈亮.基于嵌入式 CPU 的数据加解密子系统的设计研究[D].杭州:浙江大学,2013.

[5] 刘政林,张振华,陈飞,等.基于 AHB-Lite 总线的祖冲之密码算法 IP 核研究[J].微电子学与计算机,2015,32(8):88-92.

[6] 刘培龙.基于 FPGA 的神经网络硬件实现的研究与设计[D].成都:电子科技大学,2012.

[7] 杨程.基于 FPGA 的人工神经网络的研究与实现[D].西安:西安电子科技大学,2016.

[8] 李昂,王沁,李占才,等.基于 FPGA 的神经网络硬件实现方法[J].北京科技大学学报,2007(1):90-95.

[9] 钱玉多.基于 FPGA 的神经网络硬件实现研究[D].武汉:华中科技大学,2012.

[10] 杨银涛,汪海波,张志,等.基于 FPGA 的人工神经网络实现方法的研究[J].现代电子技术,2009,32(18):170-174.

[11] 洪启飞.面向深度学习的 FPGA 硬件加速平台的研究[D].成都:电子科技大学,2018.

[12] 王蒙,常胜,王豪.一种自适应训练的 BP 神经网络 FPGA 设计[J].现代电子技术,2016(15):115-118.

[13] 余子健,马德,严晓浪,等.基于 FPGA 的卷积神经网络加速器[D].杭州:浙江大学,2017.

[14] 李永红.基于 OVM 的 SoC 验证平台的设计与实现[D].西安:西安电子科技大学,2013.

[15] 张强.UVM 实战[M].北京:机械工业出版社,2014.

[16] 程翼胜.SoC 芯片 FPGA 原型的软硬件协同验证[J].单片机与嵌入式系统应用,2017,17(11):7-10,13.

(收稿日期:2020-09-26)

作者简介:

徐文亮(1995-),男,硕士研究生,主要研究方向:数字 IC 设计与验证。

(上接第 62 页)

情感分类方法[J].济南大学学报(自然科学版),2019,33(6):483-488.

[18] 刘铭,咎红英,原慧斌.基于 SVM 与 RNN 的文本情感关键句判定与抽取[J].山东大学学报(理学版),2014,49(11):68-73.

(收稿日期:2020-07-14)

作者简介:

杨戈(1974-),通信作者,男,博士,副教授,主要研究方向:计算机视觉、视觉跟踪技术,E-mail:yangge@pkusz.edu.cn。

杨麓涛(1998-),男,本科,主要研究方向:计算机视觉。

版权声明