文章编号:2096-7594(2022)01-0067-06 DOI:10.13501/j.cnki.42-1908/n.2022.03.011

基于多算法融合的文本抄袭检测的特征提取算法研究

陈 滔[1,2],张庆国[3*],何金波[1],周文竹[4]
(1.安徽农业大学 工学院,合肥 230036;
2.甘肃政法大学 民商经济法学院,兰州 720070;
3.安徽医科大学 临床医学院,合肥 230031;
4.安徽农业大学 资源与环境学院,合肥 230036)

摘要:特征提取是文本抄袭检测的重要环节,文本特征提取结果的质量将直接影响检测的可靠性.针对现有方法的不足,提出一种基于多算法融合的文本特征提取算法.该方法考虑到文本写作主题和写作风格对文本特征提取结果的影响,通过 LDA 主题模型、同义词林和 GloVe &TF-IDF 分别提取文本写作风格和文本主题的 3 个分特征向量,利用变分自编码器(VAE)进行混合和降维,提取出能够高度代表文本的融合特征向量.实验结果表明,该文本特征提取算法能够准确选择文本的特征集,解决了传统特征提取算法未考虑到文本写作风格和文本主题的缺点,检测的精确率达到了 97.93%,相较于其他算法有所提高.

关键词:特征提取;抄袭检测;多算法融合;写作风格;文本主题
中图分类号:TP391;TP301.6 **文献标志码**:A

Research on Feature Extraction Algorithm of Text Plagiarism Detection Based on Multi-algorithm Fusion

CHEN Tao[1,2],ZHANG Qingguo[3*],HE Jinbo[1],ZHOU Wenzhu[4]
(1.School of Engineering,Anhui Agricultural University,Hefei 230036,China;
2.School of Civil and Commercial Economic Law,Gansu University of Political Science and Law,Lanzhou 730070,China;
3.Clinical College of Anhui Medical University,Hefei 230031,China;
4.College of Resources and Environment,Anhui Agricultural University,Hefei 230036,China)

Abstract: As feature extraction is an important part of text plagiarism detection,the quality of text feature extraction results will directly affect the reliability of detection. Aiming at the shortcomings of existing methods,a text feature extraction algorithm based on multi algorithm fusion is proposed. Considering the influence of text writing topic and writing style on text feature extraction results,this method extracts three sub feature vectors of text writing style and text topic respectively through LDA topic model,synonym forest and GloVe & TF-IDF,and extracts the fusion feature vector that can highly represent the text by mixing and dimensionality reduction using variational self coder (VAE). The experimental results show that the text feature extraction algorithm can accurately select the text feature set,solve the shortcomings of the traditional feature extraction algorithm that does not consider the text writing style and text topic,and the detection precision reaches 97.93%,which is improved compared with other algorithms.

Keywords:feature extraction;plagiarism detection;multi algorithm fusion;writing style;text theme

2020 年以来,国家自然科学基金委员会通报近百起学术不端案例[1],其中不乏学术抄袭现象.为解决学术抄袭乱象,越来越多的检测算法被提出.抄袭检测算法在原理和技术上各不相同,但都需要先对文本进行特征提取再进行抄袭比对,而特征提取结果的优劣和提取特征指标都直接影响到文本抄袭检测结果的准确性与可靠性.龚科瑜等[2]利用 TF-IDF(Term Frequency-Inverse Document Frequency)算法对古籍文本内容进行特征提取,较好的提取到古籍文本中的内容特征.黄敏等[3]针对新闻类文本提出了一种 NewTF-IDF 算法,对 TF-IDF 算法做了多组合特征因子和离散度两个方面的改进,使得特征提取结果更加精确.金标等[4]提出

收稿日期:2021-11-17.
基金项目:安徽农业大学"优才计划"科研发展资助项目(xszz202006);安徽省学术和技术带头人及后备人选学术科研活动经费项目(2016H072).
第一作者简介:陈滔(1998-),男,工程师,主要从事图像处理和信息处理的研究;* **通信作者**:张庆国(1959-),男,博士,教授,主要从事应用数学的研究.

了一种基于依存句法的文本抄袭检测算法,在依存句法分析的基础上,通过分析句子中词语间的关系以及合并短小词语建立句法框架,进而提取文本特征.李昌兵等[5]提出一种融合卡方统计和TF-IWF(Term Frequency-Inverse Word Frequency)算法的短文本分类方法,通过卡方统计对训练数据集提取特征词,然后利用TF-IWF算法对特征词赋予权重后进行文本分类检测,取得了良好的检测结果.

传统的文本特征提取方法在文本抄袭检测应用中已触及许多领域,但仍存在未考虑文本主题和文本作者写作风格导致检测结果不理想等问题.同时,国内外学者大多采用单一算法进行文本特征提取,但是不同的算法具有各自的优点和应用的局限性.本文采用NLP(Neuro-Linguistic Programming)与机器学习领域的相关算法,考虑文本写作主题和写作风格对文本特征提取结果的影响,自主设计了一套应用于抄袭检测的多算法文本特征提取流程.实验表明,本文所提出的文本特征提取算法能够准确地识别文本的特点,很大程度上提高了文本抄袭检测的精度与可靠度.

1 算法原理

1.1 LDA主题模型

文本的虚词、标点甚至俚语的使用对于提取作者的写作风格相当有效[6-7],并且使用虚词的特征能够有效避开文章主题对于作者写作风格提取的影响.LDA[8](Latent Dirichlet allocation)是一种文档主题的生成模型,其本质是3层贝叶斯概率模型,其中包含词(W)、主题(t)和文档(d)3层结构.本文创新性地将LDA算法应用在虚词的提取与分析中,其目的是提取作者虚词的使用以及其反映的写作风格,此时的3层结构为虚词(W)、写作风格(s)和文档(d),如图1.

图1中W代表实际可观测值,S代表写作风格隐含变量,方框代表吉布斯采样迭代次数,箭头表示各变量之间的概率依赖关系.图1中各变量所代表的含义见表1.

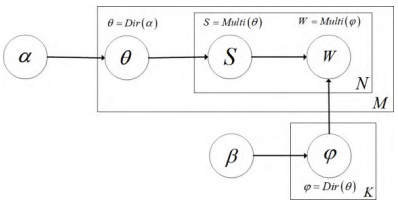

图1 LDA的贝叶斯网络图

Fig.1 Bayesian network diagram of LDA

表1 LDA主题模型中各变量含义

Tab.1 Meanings of variables in LDA subject model

变量	含义
α	文档-写作风格向量服从Dirichlet分布的超参
β	文档-虚词向量服从Dirichlet分布的超参
K	写作风格数量
θ	文档-写作风格分布
φ	文档-虚词分布
s	写作风格
W	虚词
M	文档总数
N	虚词总数

LDA主题模型的具体思想[9]如下:

设$D=\{d_1,d_1,\cdots,d_M\}$表示文档集,其中M表示文档数量.D隐含的写作风格数量为K,以$V=\{v_1,v_2,\cdots,v_N\}$表示文本中出现的虚词的集合,其中N表示虚词的总数.根据LDA主题模型做出以下定义.

定义1(文档-写作风格分布) 对D中任意文档d_i,生成所有的写作风格概率为$\theta_d=\{p_{s_1},p_{s_2},\cdots,p_{s_M}\}$,其中$p_{s_i}=N_{s_i}/N$表示第$i$个写作风格$s_i$的概率,$N_{s_i}$表示$d_i$中第$i$个写作风格$s_i$所含有的虚词的数量.

定义2(文档-虚词分布) 对任意写作风格s_i下对应词汇的概率为$\varphi_s=\{p_{w_1},p_{w_2},\cdots,p_{w_N}\}$,其中$p_{s_i}=N_{w_i}/N$表示$s_i$生成$V$中第$i$个虚词的概率,$N_{w_i}$表示分配写作风格$s_i$的$V$中第$i$个虚词的数量.

根据定义1、2,可以得到虚词、写作风格、文档的联合分布为

$$P(\theta,S,W\mid\alpha,\beta)=P(\theta\mid\alpha)\prod_{n=1}^{N}P(s_n\mid\theta)P(w_n\mid s_n,\beta),\tag{1}$$

其中,θ是K维文档-写作风格分布,S是K维写作风格,W是N虚词和写作风格.要想得到每个虚词w的生成概率,就需要计算W的边缘分布,计算方法如下:

$$P(w\mid\alpha,\beta)=\int P(\theta\mid\alpha)\Big(\prod_{n=1}^{N}\sum_{s_n}P(s_n\mid\theta)P(w_n\mid s_n,\beta)\Big)\mathrm{d}\theta.\tag{2}$$

利用吉布斯采样对虚词进行抽样,再采用EM算法估计θ和φ,具体如下:

$$P(s_i=k\mid s_{-i},w_i,d_i)\propto\frac{C_{d_ik}^{DK}+\alpha}{\sum_{k=1}^{K}C_{d_ik}^{DK}+K\alpha}\times\frac{C_{w_ik}^{WK}+\beta}{\sum_{w=1}^{W}C_{w_ik}^{WK}+W\beta},\tag{3}$$

$$\theta_{m,f}=\frac{n_m^{(f)}+\alpha_f}{\sum_{k=1}^{K}n_m^{(f)}+K\alpha_f},\tag{4}$$

$$\varphi_{f,n}=\frac{n_f^{(n)}+\beta_n}{\sum_{w=1}^{W}n_f^{(n)}+V\beta_n}.\tag{5}$$

其中,s_{-1}表示 w_i 和 d_i 分配给非写作风格的概率,直到 θ 和 φ 趋于稳定时,停止采样.

1.2　同义词林

使用哈尔滨工业大学社会计算与信息检索研究中心研制的《大词林》[10]对文本的用词偏好特征进行提取,《大词林》收录了 75 万核心实体和核心实体对应的 1.8 万细粒度概念词表.

对输入的文本集合进行切词,将词列表看作词袋,统计各个同义词类内部的概率分布信息[11].对于任意的输入词袋,每一个同义词类内部各个词语 w 的概率的 p_w 总和为 1.选取词林中词语来代表等价类,并将信息记录在同义词字典(简称词字典)中.根据已经建立的各个同义等价类的字典(称为类字典),用词字典的词条对应一个词的方法,类字典的词条就会对应一个等价类.对于类字典中的类词条 C,在维护等价类内成员的总出现频率计数器 cou_C,还为类内任一个词语 w 维护一个计数器 cou_w.cou_C 与 cou_w 的关系应该满足:

$$\sum_{w\in C}cou_w=cou_C.\tag{6}$$

由于既统计了等价类内部一个词语 w 的频率,又统计了整个等价类 C 的频率,实际上统计了等价类内部的概率分布,每一个词语在一个同义等价类中被选用的概率为

$$p_w=\frac{cou_w}{cou_C}.\tag{7}$$

当分析完整个词袋时,词袋中出现的所有同义等价类内所有出现的词语被选用的概率分布,而该分布体现了文本的同义词选用特征.于是,两个词袋 l_1 和 l_2 间各个词被选用的概率的差别为

$$D_{w,C}=|p_{w,C,l_1}-p_{w,C,l_2}|,D_{w,C}\in[0,1].\tag{8}$$

通过比较不同作者撰写的文本 W_1 词袋集、W_2 中的词袋,将文章中抓取的每一个词语对平均差异的贡献进行累加,就得到每一个类的总贡献,依序找到最能区分不同作者的等价类并进行等价类排序.利用主成分分析法[12]对向量进行降维,就得到文本的同义词林维度的向量.

1.3　GloVe 和 TF-IDF 算法

在文本主题特征向量维度提取中,采用 GloVe 算法[13](Global Vectors for Word Representation)进行词向量训练,然后提取 1 篇文章内 TF-IDF 值最高的词汇,将这些词汇对应的词向量以正比其 TF-IDF 值[14-15]的权重进行加权,最终通过词向量叠加得到了文本主题向量维度.

GloVe 算法模型为

$$J=\sum_{i,j}^{N}f(X_{i,j})\ (v_i{}^Tv_j+b_i+b_j-\ln(X_{i,j}))^2,\tag{9}$$

其中,X 为共现矩阵,其元素 $X_{i,j}$表示单词 i 和 j 共同出现在一个窗口的次数,其中窗口大小一般为 5~10.v_i 和 v_j 是单词 i 和 j 的词向量,b_i 和 b_j 是偏差项,N 是共现矩阵 $N\times N$ 的维度,f 是权重函数.权重函数满足[16]:

$$f(x)=\begin{cases}(x/x_{\max})^{\alpha},x<x_{\max},\\1,x\geqslant x_{\max}.\end{cases}\tag{10}$$

选取 $\alpha=3/4$,$x_{\max}=100$ 作为参数值.

TF-IDF 的计算公式为

$$TF\text{-}IDF_{m,i}=\frac{n_{m,i}}{\sum_k n_{m,k}}\cdot\log\frac{M}{1+\sum_l n_{l,i\neq 0}},\tag{11}$$

其中,$n_{m,i}$为词汇 w_i 在文档 m 中出现的次数,M 为文档总数,得到 TF-IDF 值.

1.4　变分自编码器 VAE

变分自编码器(Variation Auto-Encoder,VAE)是 2014 年提出的一种基于变分贝叶斯推断的生成式结构

模型[17-20].对于训练集数据集$\{x_1,x_2,\cdots,x_n\}$,假设$x_i\sim N(\mu_i,\sigma_i^2)$,则VAE的边缘对数似然函数可以由每个数据点的边缘对数似然值求和所确定,即$\log p(x_1,\cdots,x_n)=\sum_{i=1}^{n}\log p(x_i)$引入近似后验分布$q_\theta(z|x)$拟合真实后验分布$q(z|x)$,则单个数据点的边缘对数似然值计算方法如下:

$$\log p(x_i)=D_{KL}(q_\theta(z|x)\,||\,q(z|x))+l_{VAE}(\theta;x_i), \tag{12}$$

其中,$D_{KL}(q_\theta(z|x)\,||\,q(z|x))>0$表示近似后验分布与真实后验分布之间的KL散度.$l_{VAE}(\theta;x_i)$表示该数据点的边缘对数似然的变分下界.通过不断优化$l_{VAE}(\theta;x_i)$的值就可以使近似后验分布$q_\theta(z|x)$不断逼近真实后验分布$q(z|x)$.

本文中主要使用VAE来进行各方面信息的混合和降维得到文本的综合特征向量.

1.5 相似度计算

利用Jaccard系数[21-23]来衡量文本的相似度,在得到文本特征后,统计两个文本中相同的特征向量与两个特征向量,相同的特征向量与特征向量的比即为相似度值.其计算方式为

$$SIM(a,b)=\frac{|f(a)\cap f(b)|}{|f(a)\cup f(b)|}, \tag{13}$$

其中,$f(a)$和$f(b)$分别表示文本A和待对比文本B的特征向量.本文取$SIM(a,b)\geqslant 0.3$时判定文本A存在剽窃文本B的现象.

2 多算法文本特征提取方法

多算法文本特征提取方法如图2所示,具体提取流程如下.

1) 利用Jieba工具对文本进行预处理,将虚词、实词分别存储,从而方便后续对文本特征进行提取.

2) 利用分词结果,分别从虚词、实词两个方向进行分析.对于虚词,采用LDA主题模型对虚词进行特征提取,得到文本写作风格其中一个方面的向量维度,即虚词使用维度向量.对于文本中的实词,将其细化为两个部分,分开处理.由于在汉语言中存在大量的同义词,认为不同作者使用同义词时有所差异,通过同义词林对于相同作者的文本使用的同义词进行统计,通过基本的降维,得到写作风格的另一方面的向量维度,即用词偏好向量;针对主题的分析,使用GloVe算法提取词向量,然后使用TF-IDF赋值加权计算得到代表写作主题的向量,即主题向量.

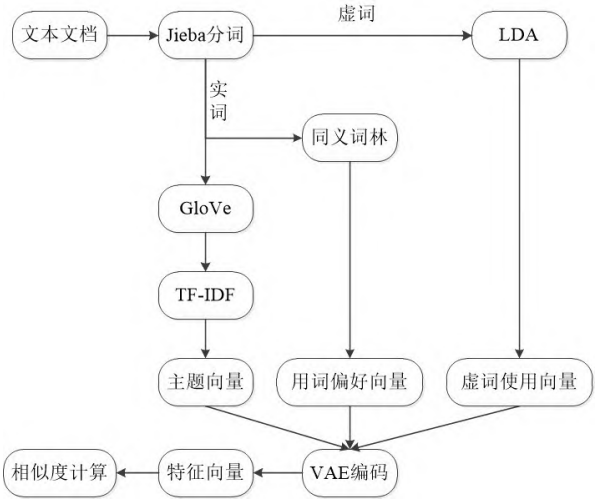

图2 特征提取流程

Fig.2 Feature extraction process

3) 用变分自编码器VAE将3个向量进行降维和混合,得到了能够代表文本的综合特征向量.

4) 进行相似度计算,相似程度给出是否存在抄袭的判断.

3 实验方案及结果分析

3.1 实验数据集

使用python爬取各大新闻平台的新闻文本,搜集了近3年来发布的13 029篇新闻,分为历史、军事、文化、读书等8个类别,所有的新闻文本均以txt文本下载保存.通过借助人工检测和计算机辅助手段,随机选取2 000篇新闻文本作为训练集,剩余11 029篇新闻文本作为验证集.

3.2 评价指标

从可行性与有效性两个方面对抄袭检测算法的准确性进行验证,选取精确率、召回率和F_1作为评价指标[24-25].对于实验所用数据集,实际存在抄袭行为的新闻文本量用N_p表示;经过相似度计算后检测到的存在抄袭行为的新闻文本量为N,其中检测结果判定正确的新闻文本量为N_t,检测结果判定错误的新闻文本量为N_f.

① 精确率P:正确检测出抄袭行为的新闻文本量N_t与检测出来存在抄袭行为的新闻文本量N的比值,

即 $P=N_t/N$;② 召回率 R:正确检测出抄袭行为的新闻文本量 N_t 与实际存在抄袭行为的新闻文本量 N_p 的比值,即 $R=N_t/N_p$;③ F_1 值:$F_1=2PR/(P+R)$.

3.3　实验结果与分析

使用 LDA、GloVe&TF-IDF 和同义词林 3 种算法结合生成文本的特征向量,而后基于此特征进行相似度分析.为了对比特征提取结果的有效性,将本文提出的多算法文本特征提取方法与其他常用于文本生成向量的算法在抄袭检测结果的精确度进行了对比,包括只使用基于虚词的 LDA、GloVe&TF-IDF、同义词林统计算法,以及使用基于实词的传统 LDA 算法、skip-gram&TF-IDF 算法、BOW 算法和 doc2vec 算法[21].实验结果如表 2 所示.

表 2　不同特征算法下的抄袭检测结果对比

Tab.2　Comparison of plagiarism detection results under different feature algorithms

实验方法	P/%	R/%	F_1/%
多算法文本特征提取方法	97.93	89.01	92.38
基于实词的 LDA	96.12	63.51	76.65
基于虚词的 LDA	94.66	49.86	65.48
GloVe&TF-IDF	93.41	57.66	69.49
同义词林	92.77	59.78	66.93
skip-gram&TF-IDF	94.47	77.98	82.73
BOW	89.50	69.73	71.99
doc2vec	97.40	83.01	91.32

从表 2 可以看出,本文所提出多算法文本特征提取方法的 F_1 值最高,比 doc2vec 算法高出 1.06%,这是因为虽然 doc2vec 性能很好,但是没有突出作者写作风格的影响.同时从表 2 可以看出本文所提出的算法的精确率和召回率均为最高,分别达到了 97.93% 和 89.01%,具有良好的检测效果,说明了本文所提出方法的高度可行性.为了更加直观清晰地展示改进的特征提取算法对分类器分类的指标提升效果,本文根据表 2 中的数据绘制了不同分类算法的精确率、召回率和 F_1 对比图,如图 3 所示.

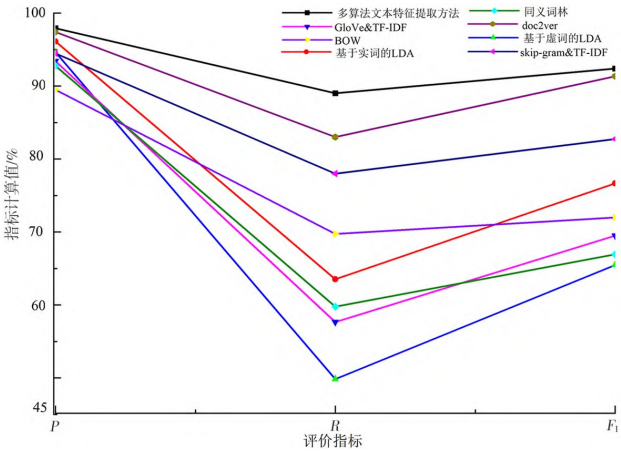

图 3　不同分类算法的精确率、召回率和 F_1 对比图

Fig.3　Comparison of precision, recall and F_1 of different classification algorithms

由图 3 可知,通过折线图的趋势可以快速直观地看出本文所提出的算法在 3 项指标上的数值均领先于其他算法,同时可以发现 doc2vec 算法对文本抄袭检测的精确率较高,能够有效对文本进行抄袭检测,但是由于算法没有考虑到文本写作风格和文本主题的影响,导致正确检测出新闻抄袭文本数量相较于本文提出的算法低.本文提出的多算法文本特征提取方法对文本抄袭检测具有高度的有效性与可靠性.

通过上述实验结果分析可以看出,提出的多算法组合对文本特征提取效果有以下优点:① 考虑文本主题与文本风格对文本特征提取结果的影响,设计了一套完整的多算法流程组合对文本进行提取,并将其应用于文本抄袭检测技术中,在减少文本特征数量的基础上增加了抄袭检测的准确率;② 通过构建合理的组合算法流程成功提取出了具有代表性的文本特征,实验结果较为理想.通过上述实验可知,提出的多算法流程法能够准确提取文本特征,降低了特征数量,在查准率和查全率上都有一定的提高.

4　结语

通过分析国内外学者对文本特征提取的缺陷,提出了一套多算法结合的文本特征提取流程,在文本主题和文本写作风格两个维度,将文本从实词和虚词两个部分进行分析.对虚词部分使用 LDA 主题模型获得了一部分文本写作风格其中一个方面的向量维度;针对实词部分,使用同义词林算法得到文本写作风格另一方面的向量维度,再对实词使用 GloVe&TF-IDF 算法得到关于文本主题的向量维度.最后,使用变分自编码器(VAE)对上述向量进行混合和降维,提取出能够高度代表文本的特征向量.实验结果表明,考虑文本主题和文本写作风格的多算法融合特征提取算法能够准确提取文本特征,提高了抄袭检测精度.

参考文献:

[1]　国家自然科学基金委员会.2021 年查处的不端行为案件处理决定(第三批次)[EB/OL].(2021-10-22)[2021-11-16].

https://www.nsfc.gov.cn/publish/portal0/tab434/info81957.htm.
[2] 龚科瑜,张一驰.基于TF-IDF的古籍文本内容特征提取方法[J].电子技术与软件工程,2019(17):130-131.
[3] 黄敏,闫思贤.基于NewTF-IDF的新闻文本特征提取算法研究[J].湖北民族大学学报(自然科学版),2021,39(2):187-192.
[4] 金标,赵萌萌,吴国华.一种用于文本抄袭检测的特征提取算法[J].计算机应用研究,2018,35(9):2781-2784.
[5] 李昌兵,段祺俊,纪聪辉,等.融合卡方统计和TF-IWF算法的特征提取和短文本分类方法[J].重庆理工大学学报(自然科学),2021,35(5):135-140.
[6] ROCHA A,SCHEIRER W J,FORSTALL C W,et al.Authorship attribution for social media forensics[J].IEEE Transactions on Information Forensics and Security,2017,12(1):5-33.
[7] OVERDORF R,GREENSTADT R.Blogs,twitter feeds,and reddit comments:Cross-domain authorship attribution[J].Proceedings on Privacy Enhancing Technologies,2016(3):155-171.
[8] 席笑文,郭颖,宋欣娜,等.基于word2vec与LDA主题模型的技术相似性可视化研究[J].情报学报,2021,40(9):974-983.
[9] 罗颖,陈伟,张超.基于LDA模型的疾病患者网络社区发现方法[J].九江学院学报(自然科学版),2021,36(3):59-62.
[10] 哈工大社会计算与信息检索研究中心.《大词林》数据库[EB/OL].(2020-03-30)[2021-11-16].http://101.200.120.155/browser.
[11] 赵逢毅,钟晓芳.基于字典释义关联方法的同义词概念撷取:以《同义词词林(扩展版)》为例[J].中文计算语言学期刊,2013,18(2):35-55.
[12] 秦传波,冯宝,谌瑶.基于主成分分析法和极限学习机的尿沉渣图像识别算法研究[J].现代电子技术,2019,42(11):45-49.
[13] PENNINGTON J,SOCHER R,MANNING C.Glove:global vectors for word representation[C]//Conference on Empirical Methods in Natural Language Processing,Doha:Qatar,2014:1532-1543.
[14] 路健,范增民,刘彩娜.基于TF-IDF算法的供应链信息定向挖掘模型[J].计算机仿真,2021,38(7):153-156.
[15] 张云婷,叶麟,方滨兴,等.基于词频-逆文档频率和法律本体的相似案例检索算法[J].智能计算机与应用,2021,11(5):229-235.
[16] 方炯焜,陈平华,廖文雄.结合GloVe和GRU的文本分类模型[J].计算机工程与应用,2020,56(20):98-103.
[17] KINGMA D P,WELLING M.Auto-encoding variational bayes[EB/OL].(2013-11-20)[2021-11-16].https://arxiv.org/abs/1312.6114.
[18] 王德文,魏波涛.基于孪生变分自编码器的小样本图像分类方法[J].智能系统学报,2021,16(2):254-262.
[19] 张蕾,钱峰,赵姝,等.利用变分自编码器进行网络表示学习[J].计算机科学与探索,2019,13(10):1733-1744.
[20] 贾修一,张文舟,李伟湋,等.基于变分自编码器的异构缺陷预测特征表示方法[J].软件学报,2021,32(7):2204-2218.
[21] 董仕,马怀祥.基于改进Jaccard系数的证据间相似性度量方法[J].石家庄铁道大学学报(自然科学版),2021,34(2):66-71.
[22] 周艳平,李金鹏.一种基于词向量及位置编码的Jaccard相似度算法[J].青岛科技大学学报(自然科学版),2020,41(6):93-98.
[23] LE Q V,MIKOLOV T.Distributed representations of sentences and documents[J].Eprint Arxiv,2014(4):1188-1196.
[24] 祖月芳,凌海风,吕永顺.基于NLP技术的装备故障文本匹配算法研究[J].兵器装备工程学报,2021,42(11):204-208.
[25] 黄奇文,李丽颖,沈富可,等.基于集成特征选择的网络异常流量检测[J].华东师范大学学报(自然科学版),2021(6):100-111.

责任编辑:高 林

(上接第18页)

[23] 卜贵军.大泥炭藓(*Sphagnum palustre*)生理特性及其腐殖化指标研究[D].武汉:湖北大学,2019.
[24] OHNO T.Fluorescence inner-filter correction for determining the humification index of dissolved organic matter[J].Environmental Science and Technology,2002,36:742-746.
[25] 肖隆庚,陈文松,陈国丰,等.中国南海CDOM三维荧光光谱特征研究[J].环境科学学报,2014,34(1):160-167.
[26] 何文远,杨海真,顾国维.腐殖质生物活性机理研究进展[J].腐植酸,2007(3):11-16.

责任编辑:高 山