文章编号：2096-1472(2023)-01-19-05
DOI:10.19644/j.cnki.issn2096-1472.2023.001.005

基于Word2Vec的中文文本零水印算法

戴夏菁，徐谊程，王馨娅，佟德宇

(南京财经大学信息工程学院，江苏 南京 210023)
✉2415700426@qq.com; YichengXu421@163.com; 1520369099@qq.com; tdyforweb@163.com

摘　要：经典的文本鲁棒水印会修改文本内容或格式，从而降低文本的保真性和可用性，文章提出了一种基于Word2Vec的中文文本零水印算法，能够在不修改文本信息的前提下实现水印的生成和检测。首先对文本数据进行分词，统计词频并提取特征词，运用Word2Vec生成相应的特征词向量；然后采用SVD(奇异值分解)算法对其进行降维，并结合AES(高级加密标准)加密生成最终的零水印。水印检测时，通过对比SVD分解产生的特征值和特征向量判断版权归属。基于理论概述和实验结果综合分析，文章提出的零水印算法不需要对原始文本做任何修改，能够抵抗一定程度的增删、句型转换、同义词替换等攻击，具有一定的鲁棒性，切实有效地解决了文本的版权保护问题。

关键词：Word2Vec；SVD；零水印；中文文本；词向量
中图分类号：TP309.2　**文献标识码**：A

A Zero-Watermark Algorithm for Chinese Text based on Word2Vec

DAI Xiajing, XU Yicheng, WANG Xinya, TONG Deyu

(*School of Information Engineering*, *Nanjing University of Finance and Economics*, *Nanjing* 210023, *China*)
✉2415700426@qq.com; YichengXu421@163.com; 1520369099@qq.com; tdyforweb@163.com

Abstract: Classic text robust watermark can modify the content or the format, thereby reducing the fidelity and usability of the text. This paper proposes a Word2Vec-based zero-watermark algorithm for Chinese text, which ensures that watermark generation and detection make no modification to the original text. Firstly, by dividing the text into words, word frequency is counted and feature words are extracted; the corresponding feature word vector is generated by Word2Vec. Then, SVD (Singular Value Decomposition) algorithm is used to reduce its dimension, and the zero-watermark is finally generated by AES (Advanced Encryption Standard) encryption. In watermark detection, the copyright ownership is determined by comparing the eigenvalues and eigenvectors generated by SVD. Based on theoretical summary and comprehensive analysis of experimental results, the proposed zero-watermark algorithm does not need to make any modification to the original text, and can resist attacks such as addition and deletion, sentence pattern conversion and synonym substitution to a certain extent. It has certain robustness and effectively solves the problem of protecting the copyright of the text.

Keywords: Word2Vec; SVD; zero-watermark; Chinese text; word vector

1 引言(Introduction)

以中文文本为主要载体的信息在网络空间广泛传播，例如小说、论文、档案、博客等，但是此类文本信息极易被复制和修改，不仅侵害了文本原创作者的权益，更给网络空间安全带来了不可忽视的影响，亟须采取切实有效的技术手段解决文本的版权保护问题。

近年来，数字水印在文本的版权保护领域得到了广泛的研究和应用。经典的嵌入式水印的主要思想是对需要保护的文本进行小幅修改，将水印信息嵌入其中，同时保证嵌入的信息难以被肉眼察觉。目前，文本数字水印的算法主要分为三类：(1)基于文本内容的水印算法，例如肖湘蓉等提出基于英文字母的文字特性，用希腊字母进行替换的方法[1]；(2)基于字形结构，例如LIU等人提出了基于文本字形结构构造文本图像水印的方法[2]；(3)基于文本格式，例如将水印信息嵌入行间距信息中的改进方法[3]，以及将水印嵌入字间距中的嵌入方法[4]。但是，上述研究不可避免地需要对文本进行修改，势

基金项目：大学生创新创业训练计划项目(202110327082Y)；江苏省自然科学基金项目(BK20200839).

必影响文本的可用性和完整性，因此如何解决这一矛盾，是文本版权保护技术研究人员需要深入思考的问题。

零水印作为一种新兴的数字水印方法，它的原理是在不对原始载体信号进行任何更改的情况下，将版权信息与特征数据相结合，实现水印的生成并注册至可信的第三方机构，即可在发生版权纠纷时进行仲裁，实现版权信息的保护[5]。对于中文文本零水印的研究，龚礼春等人提出了通过训练文本实体识别模型得到实体特征的方法[6]；蒙应杰等提出了用对汉字进行矢量化处理的方法实现水印的嵌入[7]；张娜等将文本主题词在《同义词词林》中的编码与依据全文信息熵直方图获取的编码融合，加密后形成文本零水印[8]。

为进一步提高文本零水印的鲁棒性，特别是增强零水印对各种文本攻击方式的抵抗能力，本文基于自然语言处理技术中的Word2Vec模型，提出了一种适用于中文文本的零水印算法。

2 技术基础(Basic technologies)

2.1 Word2Vec

Word2Vec是一个将自然语言中的字词转化为计算机可理解的稠密向量的工具，它可以把文本内容处理简化为向量运算，并通过计算向量空间的相似度表示文本语义上的相似度[9]。Word2Vec的本质是一个浅层神经网络用于映射每个词至一个向量，这种映射生成的词向量可以很好地度量词与词之间的相似性。

Word2Vec算法多指用于计算的词向量的连续词袋(Continuous Bag-of-Word，CBOW)模型[10]和连续跳字(Skip-Gram)模型[11]，前者通过上下文预测当前词，后者通过当前词预测上下文。

以CBOW模型为例，输入层由One-hot编码的上下文$\{x_1,x_2,\cdots,x_c\}$组成，词汇表的大小为V，隐藏层为N维的向量，输出层为被One-hot编码的输出单词y。输入向量通过一个$V\times N$维的权重矩阵W连接到隐藏层；隐藏层通过一个$N\times V$的权重矩阵W'连接到输出层。CBOW的模型如图1所示。

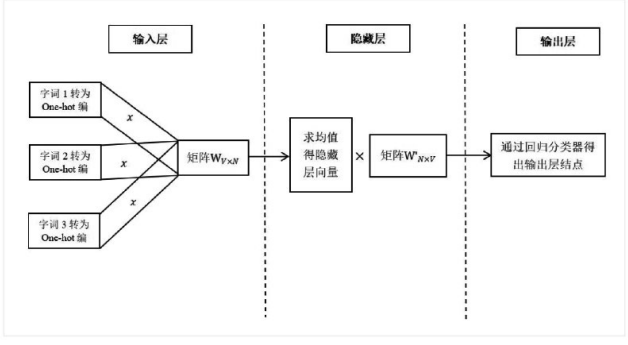

图1 CBOW模型

Fig.1 CBOW model

计算隐藏层h的输出，为输入向量的加权平均：

$$h=\frac{1}{C}W\cdot(\sum_{i=1}^{C}x_i) \tag{1}$$

计算在每个输出层每个节点的输入：

$$u_j={v_{wj}}'^{T}\cdot h \tag{2}$$

计算输出层的输出，输出y_j如下：

$$y_{c,j}=p(w_{y,j}|w_1,\cdots,w_c)=\frac{\exp(u_j)}{\sum_{j'=1}^{V}\exp(u'j)} \tag{3}$$

其中，W是原始的输入矩阵，${v_{wj}}'^{T}$是输出矩阵W'的第j列，u代表输出层的原始结果。

Skip-Gram模型则与之相反。

2.2 奇异值分解

奇异值分解(Singular Value Decomposition，SVD)是线性代数中对特征值分解的推广，适用于任何维度矩阵的分解；其本质是借助低维矩阵实现对原矩阵的特征提取。SVD的应用非常广泛，它不仅可用于降维算法中的特征提取[如在主成分分析法(Principal Component Analysis，PCA)中的应用[12]，还可被用于数据压缩[13]、自然语言处理[14-15]等领域[如在隐性语义索引(Latent Semantic Indexing，LSI)中的应用。

SVD变换有两个很好的性质：一是经过一些常见的处理，如滤波、压缩等操作，奇异值变化较小。二是奇异值包含丰富的内在属性，不易改变。这两个性质决定了SVD具有较强的安全性。

具体步骤如下：先将词向量A分别左乘和右乘它的转置A^T，并计算其特征向量和对应的特征值；接着取出$A*A^T$的特征向量组成U，A^T*A的特征向量组成V；最后求出非零的特征值的平方根，对应上述特征向量的位置，填入奇异值矩阵Σ。

设W是生成的$m\times n$词向量实数矩阵集合，w_i是其中的第i个矩阵，根据奇异值分解的原理：

$$W=U\Sigma V^T \tag{4}$$

可以将其转换为$m\times m$的单位正交阵U、$n\times n$的单位正交阵V及$m\times n$的奇异值矩阵Σ的乘积。对于奇异值矩阵Σ而言，它只有主对角线上的元素有值，其余元素均为0。位于主对角线上的元素被称为奇异值。即，Σ一般为以下形式：

$$\Sigma=\begin{bmatrix}\alpha_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \alpha_n\end{bmatrix} \tag{5}$$

通常，将奇异值按照由大到小的顺序排列在主对角线上，奇异值越大，它所代表的信息越多。因而，当选取前面若干个最大的奇异值时，往往能够在很大程度上还原数据本身，这就实现了对矩阵的降维压缩。

2.3 余弦相似度

余弦相似度用向量空间中的两个向量夹角的余弦值作为衡量两个个体间差异大小的度量，其值越接近1，就说明夹角角度越接近0°，也就是两个向量越相似[16]。余弦相似度通常用

于文本挖掘中的文件比较[17]。

设两个属性向量分别为A和B，A_i和B_i分别表示向量A和B的各分量，向量A和B的余弦相似性θ由点积和向量长度给出，公式如下所示：

$$similarity = \cos\theta = \frac{A\cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{6}$$

余弦相似度通常用于正空间，因此给出的值为−1—1。−1意味着两个向量指向的方向正好截然相反，1表示它们的指向是完全相同的，0通常表示它们之间是独立的，而介于0和1的数值则表示中间程度的相似性或相异性。因此，余弦相似度可用于水印相似度的检验。

3　算法原理与步骤(The principle and procedure of the proposed algorithm)

3.1　算法思路

数字水印的安全性是指未经授权的用户无法进行检测与解码、嵌入和删除水印等操作。数字水印的安全性依赖于密钥，针对安全性的攻击目的也是为了获得水印系统密钥。本文通过对中文文本的特征词的词向量矩阵进行加密，从而生成零水印。

本文在构造零水印时，首先对中文文本进行预处理，主要包括删除标点符号和根据词性对文本进行分词处理，以方便词频的统计及特征的提取。然后借助词频统计算法，对所选文本中所有的词汇进行词频统计，选择中频词作为备选词，以充分体现文本的特征，保证备选词的典型性、代表性。本文采用连续词袋模型进行Word2Vec的训练。在训练时，借助随机梯度下降法和反向传播误差算法(Error Back Propagation)提高神经网络训练的效率，不断迭代生成权重矩阵，直至完成训练。借助Word2Vec生成备选词词向量后，为了减小特征数据量，采用SVD对词向量矩阵进行降维处理，生成SVD特征矩阵。为了保证零水印的权威性，避免未授权人员的水印检测，引入密钥对SVD的特征值采用AES加密，最终生成难以被破解的零水印信息，可提交至第三方机构进行版权注册，水印生成流程如图2所示。

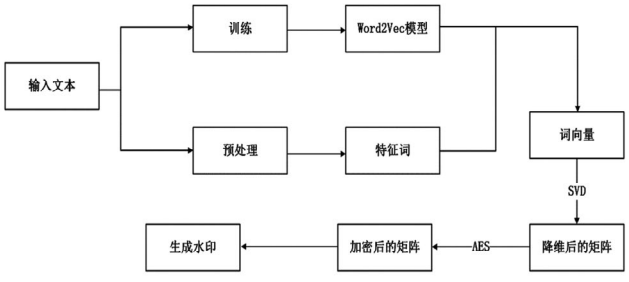

图2 水印生成流程图

Fig.2 Watermark generation flow chart

3.2　算法步骤

Step1：Word2Vec模型的训练。

输入的文本记为X，文本X的容量大小记为V，隐藏层的大小为N，隐藏层的向量记为n，权重矩阵记为M，M矩阵中一个与输入层相关的单词的N维向量记为V_M。具体的训练步骤如下。

(1)对文本X进行One−hot编码，One−hot编码中每个单词的向量表示为$\{x_1, x_2, x_3, \cdots, x_v\}$，其中只有一个节点$x_k$为1，其他节点为0，得到编码后的文本$X'$。

(2)根据公式计算得到输入层与输出层之间的权重矩阵M，M_{vn}表示第v个输入节点到第n个隐藏层节点边的权值，公式如下：

$$M = V \times N = \begin{pmatrix} M_{11} & \cdots & M_{1n} \\ \vdots & \ddots & \vdots \\ M_{v1} & \cdots & M_{vn} \end{pmatrix} \tag{7}$$

(3)h_i表示第i个隐藏层节点(单词的词向量)，M矩阵的每一行代表一个与输入层相关单词的N维向量，记为V_M，h_i的计算公式如下：

$$h_i = V_M^T \times X' = M\cdot(x_1 + x_2 + \cdots + x_n) \tag{8}$$

(4)依次计算N个单词的词向量，然后计算隐藏层的h向量，将输入的上下文单词的向量累加后求均值。

$$h = V^T \times X' = \frac{1}{N}\cdot(h_1 + h_2 + \cdots + h_N) \tag{9}$$

(5)隐藏层的输出h向量即训练好的词向量。

Step2：文本预处理。

考虑到选取的中文文本较长，本文采用了中文分词库——jieba分词库(可参考官网https://pypi.org/project/jieba/)，根据已有的中文词库，推算单个汉字之间相互关联的概率，计算出结果后将概率大的汉字组成词语，记为分词后的结果，存放在词典D中。

Step3：词频统计。

分词结束后，遍历词典D中的所有字词，并记录每组字词出现的次数。为了保障所选取的特征词在具有文本代表性的同时，排除虚词、人称代词等带来的影响，本方法将所有字词按照出现的次数从大到小进行排序，并选取出现次数占总次数的40%—60%的中频词作为该文本的特征词，记为A数组。

Step4：词向量生成。

将提取出的特征词A输入训练好的Word2Vec模型，得到由隐藏层输出的词向量记为W。

Step5：奇异值分解。

将生成的W向量空间进行降维压缩，本算法选取4作为奇异值矩阵Σ的维数，生成低维矩阵：

$$\Sigma_n = Diag(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \tag{10}$$

式(10)作为AES加密的原始数据。重复上述操作，得到矩阵集合：

$$\Sigma = \{\Sigma_1, \Sigma_2, \Sigma_3, \cdots, \Sigma_n\} \tag{11}$$

Step6：运用AES生成水印。

本文将SVD降维后得到的矩阵Σ作为水印嵌入载体，采用AES加密算法，与密钥key相结合，得到加密后的密文T：

$$CipherText\ T = AES(\Sigma, Key) \tag{12}$$

Step7：将生成的零水印T和版权所有者信息R提交至可信第三方机构或版权管理部门。

3.3 水印检测

当产生版权纠纷时，可以提取注册的零水印进行水印检测，从而确定版权归属，具体的检测流程如图3所示。

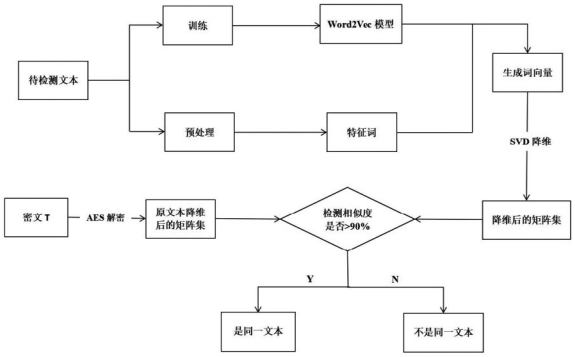

图3 水印检测流程图

Fig.3 Watermark detection flow chart

Step1：根据公式(11)对注册的零水印T进行AES解密，对解密后的零水印进行SVD处理后，得到矩阵集Σ'。

Step2：将预处理后的待检测文本输入Word2Vec模型进行训练，得到特征词的词向量W^*。

Step3：将词向量W^*进行SVD处理后，得到Σ^*。

Step4：根据公式(6)将Σ'和Σ^*进行相似度检测，若相似度高于阈值θ，则说明为同一文本或认定为文本抄袭。

4 实验与结果分析(Experimental results and analysis)

本文构造的零水印不会更改原始数据载体信号，而是通过数据本身的特征构造水印密文，故可以有效抵挡破坏数据载体的简单攻击，如压缩、噪声、剪切、旋转等。

本实验构造零水印的第一个步骤是根据数据源的词频特性，选择中频词作为数据的特征词，这就意味着水印构造对于数据源内容的删减、替换等攻击具有高度的敏感性，对攻击行为的辨别能力很强；此外，特征选择时是以词语为最小单位，故水印在面对句式转换攻击时，仍能保持较高的稳定性和鲁棒性。

本实验基于Windows 10操作系统、Anaconda环境，编程语言为Python，以名著《钢铁是怎样炼成的》为例，进行零水印的生成与检测。在换用《童年》《一千零一夜》《红与黑》等长篇书籍进行多次试验后，本文将阈值定为0.8。

4.1 文本删减攻击

该实验通过随机删除特定比例的句子造成删减攻击，文本删减率确定在0.10—0.45。将经过不同的文本删减率删减的每篇文章选取的特征备选词与原文本的特征词进行比对，计算二者间的相似度。将不同文章在不同的文本删减率下与原文本的相似度的平均值作为分析依据，得到的结果如图4所示。

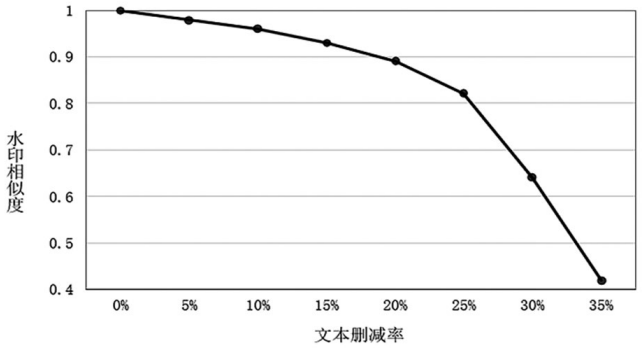

图4 不同文本删减率的水印相似度情况

Fig.4 Similarity under different pruning rates

可见，当文本删减率增大时，水印相似度降低，并且删减率越高，水印相似度下降幅度越大。当文本删减率大于25%时，水印相似度小于阈值0.8，此时该水印被判定为不相似。从实验结果来看，本文算法在应对删减攻击时具有较强的稳定性。

4.2 句型转换攻击

将实验文本进行不同程度的句型转换，然后将转换后的文本水印与原文本水印进行对比，计算它们的相似度。句式转换主要分为以下五种：①将陈述句改成“把字句”和“被字句”；②“把字句”和“被字句”的互改；③肯定句和否定句的互改；④把直接叙述改为间接叙述；⑤不同语气句子的句式改变。以句子“被抛弃后，请不要哭泣”为例，将其改为“别人把你抛弃后，请不要哭泣”，句型虽然从“把字句”变为“被字句”，但是其句意未变。本次实验选取0%—35%的文本句子，将这些句子进行句型转换，句型转换攻击下水印相似度如图5所示。

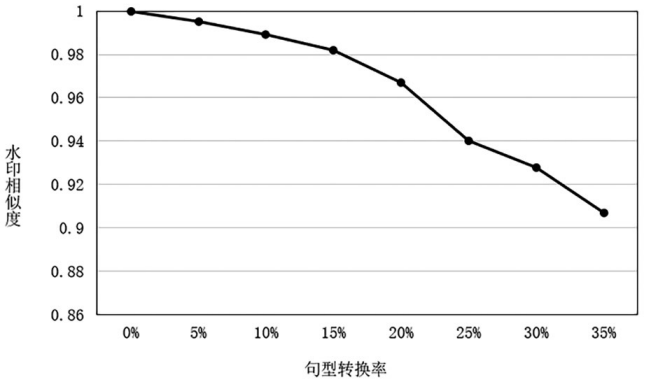

图5 句型转换攻击下水印相似度结果

Fig.5 Watermark similarity results under sentence pattern conversion attack

从图5可以看出，在不同程度的句型转换攻击下，转换后

文本与原文本的水印相似度都在0.9以上，这是因为本文算法注重提取词频特征，而句型转换对词频的影响较小，因而算法对于句型转换攻击具有较强的鲁棒性。

4.3 同义词替换攻击

从文本中随机选取10%的字词，对其进行同义词替换，用于模拟同义词替换攻击。选取被攻击后文本中的部分特征词，将其转为词向量并对矩阵进行降维处理，最后将生成的新矩阵与原矩阵进行相似度比对，结果如图6所示。

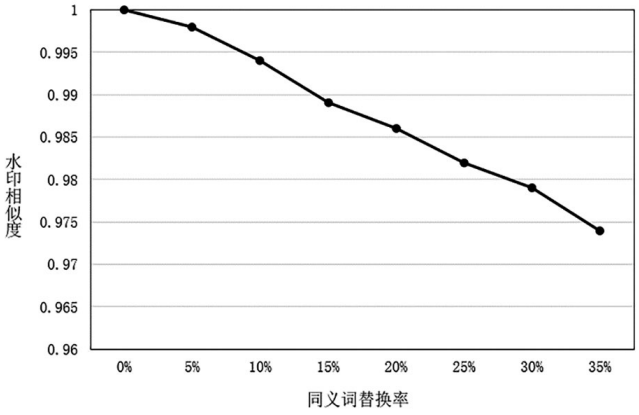

图6 不同同义词替换率下的水印相似度情况

Fig.6 Watermark similarity under different synonym replacement rates

从图6可以看出，在替换的同义词较少的情况下，新文本和原文本的相似度接近1；若将字词替换的比例扩大为35%，再将新矩阵和原矩阵进行比对发现，相似度仍高于0.97。由此可见，此算法稳定性较高，能抵抗大范围的同义词攻击，这是因为该算法选取的加密文本为中频词，所以能够避免一定程度上的同义词替换对提取特征的影响。

研究人员结合上述三种实验发现，本文算法在句型转换及同义词替换的攻击下能较好地保留原有的水印信息，具有很强的抵抗性，同时对文本适量删减也有一定的抵抗能力。

5 结论(Conclusion)

数字水印的安全性不同于密码，它比密码更复杂，主要表现为鲁棒性与安全性并不是独立的，它们是相互影响的。

本文基于自然语言处理领域的Word2Vec技术提出了一种零水印算法，首先根据文本的中频词生成词向量，然后利用SVD算法和AES加密进行词向量的降维与加密，从而生成能够充分体现文本特征的零水印。实验结果表明，本文提出的零水印算法，能够抵抗一定程度的增减、句型转换、同义词替换等攻击，水印具有较强的鲁棒性，能够有效解决文本版权的鉴定和保护问题。

参考文献(References)

[1] 肖湘蓉,孙星明.基于内容的英文文本数字水印算法设计与实现[J].计算机工程,2005,31(22):29-31.

[2] LIU Y X, GUO W, QI W F. Researches on text image watermarking scheme based on the structure of character glyph[J]. Applied Mechanics and Materials, 2015, 731:163-168.

[3] 惠路华.基于WORD文档的数字水印算法研究与实现[D].南京:南京理工大学,2008.

[4] 谭瑛.基于字符偏移的文本数字水印算法研究[J].微型电脑应用,2014,30(02):20-22.

[5] 温泉,孙锬锋,王树勋.零水印的概念与应用[J].电子学报,2003,031(002):214-216.

[6] 龚礼春,姚晔,唐观根,等.基于命名实体识别的医疗文本零水印方案[J].密码学报,2020,7(05):643-654.

[7] 蒙应杰,司蕾,是垚.基于矢量图形的中文文本零水印算法[C]//樊建平.中国电子学会通信学分会会议论文集.北京:科学出版社(Science Press),2009:33-37.

[8] 张娜,张琨,张先国,等.基于主题词与信息熵编码的文本零水印算法[J].计算机与数字工程,2021,49(08):1612-1618.

[9] 程盼,徐弼军.基于word2vec和logistic回归的中文专利文本分类研究[J].浙江科技学院学报,2021,33(06):454-460.

[10] 王辉,潘俊辉,王浩畅,等.基于改进的CBOW与BI-LSTM-ATT的文本分类研究[J].计算机与数字工程,2021,49(07):1372-1376.

[11] 黄鹤,荆晓远,董西伟,等.基于Skip-gram的CNNs文本邮件分类模型[J].计算机技术与发展,2019,29(06):143-147.

[12] 聂振国.基于奇异值分解的信号处理关键技术研究[D].广州:华南理工大学,2016.

[13] 张玉瑶,程学林,尹天鹤.基于深度学习和矩阵分解的推荐算法[J].计算机技术与发展,2021,31(07):21-27.

[14] 王怡,盖杰,武港山,等.基于潜在语义分析的中文文本层次分类技术[J].计算机应用研究,2004,(08):151-154,165.

[15] 李秋伶,郑静.基于隐马尔可夫模型的文本情感分析[J].杭州电子科技大学学报(自然科学版),2020,40(06):50-55.

[16] 陈大力,沈岩涛,谢槟竹,等.基于余弦相似度模型的最佳教练遴选算法[J].东北大学学报(自然科学版),2014,35(12):1697-1700.

[17] 张振亚,王进,程红梅,等.基于余弦相似度的文本空间索引方法研究[J].计算机科学,2005,(09):160-163.

作者简介：

戴夏菁(2001-)，女，本科生.研究领域：人工智能，零水印.

徐谊程(2002-)，女，本科生.研究领域：人工智能，零水印.

王馨娅(2002-)，女，本科生.研究领域：人工智能，零水印.

佟德宇(1989-)，男，博士，讲师.研究领域：信息安全，数字水印.