文章编号：2095-2163(2021)10-0134-05 中图分类号：TP309.2 文献标志码：A

基于 Word2Vec 的疫情虚假信息检测方法

齐浩翔，马莉媛，朱翌民
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：在面临突发大型公共事件时虚假信息的广泛传播将具有极大的破坏性。虚假信息的传播将严重干扰疫情的救治工作，针对以往传统分类模型存在特征稀疏，准确率不高等问题。提出了一种基于 Word2Vec 的疫情虚假信息检测方法。该方法使用 Word2Vec 模型训练词向量，解决了传统向量空间模型的特征稀疏问题，再引入 TFIDF 对词向量进行加权，最终将处理过后的数据输入到 SVM 模型。通过在国内新闻平台爬取的数据集上的实验验证，该方法较之传统方法，对虚假信息的检测在准确率上有 4%以上的提升。

关键词：疫情；Word2Vec；神经网络；SVM；文本分类

Word2Vec-based false information detection system of epidemic situation

QI Haoxiang, MA Liyuan, ZHU Yimin
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】The widespread dissemination of false information in the face of sudden large-scale public incidents will be extremely destructive. The dissemination of false information will seriously interfere with the treatment of the epidemic. In response of the problems of sparse features and low accuracy in traditional classification models in the past, this paper proposes a method for detecting false information about the epidemic based on Word2Vec. This method uses the Word2Vec model to train word vectors, then solves the feature sparse problem of the traditional vector space model, and introduces TFIDF to weight the word vectors, finally inputs the processed data into the SVM model. Through the experimental verification on the data set crawled by the domestic news platform, this method has more than 4% improvement in the accuracy of detecting false information compared with the traditional method.

【Key words】epidemic; Word2Vec; neural network; SVM; text classification

0 引 言

互联网的发展催生了微博、推特、贴吧、微信公众号等社交网络，新媒体时代人们获取信息更加便捷，但是随之而来的信息造假、虚假信息传播等问题也不容忽视。在信息传播的过程中，如若虚假信息传播过甚，并形成一定规模的网络舆情，就可能引发群体性事件，甚至造成不可估量的后果。在信息爆炸的大数据时代，人工识别虚假信息已然不能够满足当下需求，因此如何快速、精准地识别虚假信息是当前研究的热点之一。

识别虚假信息可以转换为文本分类问题。对于文本分类问题的研究，大多基于传统的向量空间模型。Salton 等人[1]提出的向量空间模型是现阶段使用最广泛的一种文本表示模型(VSM)，传统的 VSM 模型的主要问题是维度高或者文本表示向量稀疏。降维一般从 2 个方面进行，一是特征选择，二是特征提取，通过这两个方面的改进来提高文本分类的准确率。近年来，对于文本分类大多采用主题模型(LDA)，该模型由 Blei 等人提出，主要优点就是能够发现语料库中潜在的主题信息[2-3]；方东昊[4]利用 LDA 对微博进行分类，取得了不错的效果，但该方法需要大量的外部语料，复杂度相对较高；Kim[5]提出了一种利用卷积神经网络来处理句子分类问题的模型，该方法证明了深度学习相关技术在文本分类中具有很好的效果；Mikolov 等人[6-8]提出了 Word2Vec 模型用于计算文本中特征的词的分布式表示，该方法可以很好地表达句子中的语义信息，但却无法区分文本中词汇的重要程度。

针对目前虚假信息检测中存在的问题，本文提出了一种基于 Word2Vec 的虚假信息检测方法。该方法使用 Word2Vec 模型表示文本，针对文本间语义相似度难以很好度量的问题，进一步引入 TFIDF 模型计算 Word2Vec 词向量的权重，得到加权的 Word2Vec 模型，一个词在不同类别中分布得越不均匀，就应该赋予较高的权值。再将处理过后的数据

作者简介：齐浩翔(1996-)，男，硕士研究生，主要研究方向：人工智能、推荐系统；马莉媛(1995-)，女，硕士研究生，主要研究方向：人工智能、舆情分析、大数据；朱翌民(1995-)，男，硕士研究生，主要研究方向：人工智能、舆情分析、大数据。

收稿日期：2021-03-30

输入到 SVM 模型，将数据分为 2 类，即：真实信息和虚假信息。最终通过与传统方法相比较可知，本文提出的方法能够有效地提高检测精度。

1　相关工作

检测虚假信息实际是一个映射的过程，将待检测的数据集 $D=\{d_1,d_2,\ldots,d_n\}$ 映射到预定的分类集 $C=\{c_1,c_2,\ldots,c_n\}$ 中，其中 n 表示数据集的数量，m 表示类别的数量，由于虚假信息检测结果只存在真或假，所以 $m=2$。本文的总体框架如图 1 所示。本文采用邻近匹配算法，利用首字母索引的词典，使同一首字母下的特征词按升序排列，该方法避免了每次增加新字就要重新在字典从头匹配的冗余操作。检测的步骤为：文本预处理、特征选择、文本表示、分类器训练。本文拟对此展开研究分述如下。

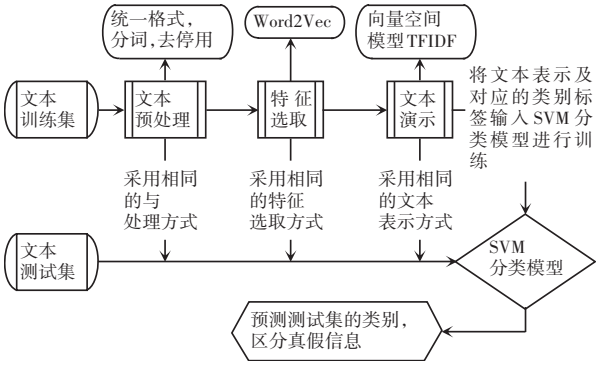

图 1　总体框架

Fig. 1　Overall framework

(1)文本预处理。对文本的语料库进行处理，主要包括去掉停用词、虚词、标点符号等，接着进行分词处理得到后续可以直接使用的数据。文本分类前去停用词能够有效改善文本的分类效果，通常是指一些使用频率极高、却没有实际意义的词，例如："的"、"了"、"我"、"吗"等，以及英文中的"the"、"a"、"of"等，还有一些数字、数学字符、标点字符[9]。

(2)特征选择。在对文本进行预处理后，得到的数据会有很高的维度且特征相对稀疏，这就需要选出最相关的特征，且降低数据的维度。

(3)文本表示。这个过程多是与过程(2)相结合，通过选取合适的文本表示模型将文本转化为数字表示，常用的方法是向量空间模型(VSM)。

(4)训练器分类。选取适合当前数据的分类算法，常用的分类器有 SVM，KNN 等。首先，用训练数据集进行训练；然后，再用测试集对样本数据进行测试；最后，根据选用的评价指标来对分类效果进行衡量。

2　本文算法

2.1　特征选取

数据经过预处理后进行特征提取。特征提取词袋模型是最早的以词为基本处理单元的文本向量化算法。该方法容易实现，但却存在很大的问题，当面对词典中包含大量单词的时候，必然会由于维数过多而导致数据稀疏，产生数目可观的无效位置，从而影响计算速度。由于解码后的数据也会面临词向量过于稀疏的问题，会产生很多无效位置和无标注数据。针对这个问题，本文提出使用 Word2Vec 模型来进行解决。研究中，利用神经网络从大量的无标注数据中快速地提取有用的信息，将其表达成向量的形式，还可以反映该词汇在上下文中的关系。Word2Vec 可以将离散的单独符号转化为低维度的连续值，也就是稠密向量，并且其中意思相近的词将被映射到向量空间中相近的位置。而使用神经网络可以灵活地对上下文进行建模。具体来说，输入层为 One-hot vector，隐藏层为线性单元，输出层使用的是 Softmax 回归。由于文本词语之间关联紧密，使用具有上下文情境的 Word2Vec 方法将词转化为向量表示，准确度更高。本文调用 *gensim* 函数训练 Word2Vec 模型，生成词向量矩阵[10]。Word2Vec 主要包括 2 种模型：Skip-gram 和 CBOW。

Skip-Gram 模型设计如图 2 所示。图 2 中，Skip-Gram 模型主要通过使用目标词汇来预测当前语境下的上下文词汇，简而言之就是输入为特定的词向量，输出为该词向量对应的上下文，而 CBOW 模型则与 Skip-Gram 模型相反，就是通过上下文信息，预测目标词汇出现的概率。Skip-Gram 模型适用于数据集较多的情况，而 CBOW 在小型数据库中有着更好的表现。本文采用 Skip-Gram 模型用于特征提取。

将模型输出层和隐藏层的权值表示为一个 $V\times N$ 的矩阵 $\boldsymbol{W}$，$\boldsymbol{W}$ 中的每一行是一个 N 维的向量，词典 V 中第 i 个特征词 w_i 在 $\boldsymbol{W}$ 中相应的表示为 v_{w_i}，假设输入层的输入 $x\in R^v$[11]，其中 $x_k=1$，$x_{k'}=0$，$x\neq x'$，则隐含层可以表示为：

$$\boldsymbol{h}=\boldsymbol{W}_{k'}^{\mathrm{T}}=\boldsymbol{v}_{w_t}^{\mathrm{T}} \tag{1}$$

输出层共有 C 个 V 维向量，$w_{c,j}$ 表示第 c 个向量的第 j 个特征词，$u_{c,j}$ 表示隐含层到输出层第 c 个向量到第 j 个单元的线性和，$y_{c,j}$ 表示 *softmax* 处理后的概率值。则目标函数定义为：

$$p(w_{c,j}\mid w_k)=y_{c,j}=\frac{\exp(u_{c,j})}{\sum_{j=1}^{V}\exp(u_{c,j})} \tag{2}$$

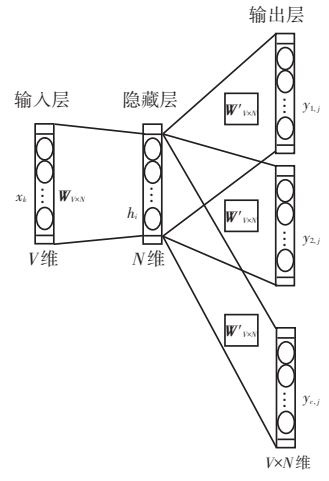

图 2 Skip-Gram 模型

Fig. 2 Skip-Gram model

由于隐含层到输出层共享相同的权值 $\boldsymbol{W}$，可以推得：

$$\boldsymbol{u}_{c,j} = \boldsymbol{v}'^{\mathrm{T}}_{w_j} \cdot \boldsymbol{h}, c = 1,2,\cdots,C \tag{3}$$

其中，$\boldsymbol{v}'_{w_j}$ 为特征词 w_j 的输出向量。

此时，目标函数更新为：

$$E = -\log p(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_c \mid \boldsymbol{w}_k) = -\log\prod_{c=1}^{C}\frac{\exp(u_{c,j})}{\sum_{j'=1}^{v}\exp(u_{j'})} = -\sum_{c=1}^{C}u_{c,j} + C\cdot\log\sum_{j'=1}^{v}\exp(u_{j'}) \tag{4}$$

假设语料词典 vocab 和文档 $\boldsymbol{d}_i = \{\boldsymbol{w}_1\ \boldsymbol{w}_2,\cdots,\boldsymbol{w}_i\}$ 文本经过 Word2Vec 模型训练语料后，得到单词词向量，将文档 $\boldsymbol{d}_i$ 的向量表示为：

$$R(\boldsymbol{d}_i) = \sum_t Word2Vec(\boldsymbol{t})\ ,\ \boldsymbol{t} \in \boldsymbol{d}_i \tag{5}$$

其中，$Word2Vec(\boldsymbol{t})$ 表示词汇 $\boldsymbol{t}$ 的 $Word2Vec$ 词向量[12]。

再引入 TFIDF 模型来计算词向量的权重，由于 TFIDF 模型本身不具备反映词向量分布情况的能力，所以将其与 Word2Vec 模型融合，得到一种 TFIDF 加权的 Word2Vec 模型，将加权过后的词向量累加得到新的文档向量表示：

$$\begin{cases} weightR(\boldsymbol{d}_i) = \sum_t Word2Vec(\boldsymbol{t}) \times \boldsymbol{w}_t \\ where\ \boldsymbol{w}_t = tfidf_t \end{cases} \tag{6}$$

2.2 文本分类

经过数据的预处理、特征选择后，最终得到文本的向量表示。本文选择支持向量机(SVM)模型进行分类。SVM 是一种成熟的机器学习中的算法。多用于解决复杂的非线性分类问题，在线性不可分的情况下，SVM 通过某种事先选择的非线性映射(核函数)将输入变量映射到一个高维的特征空间，将其变为高维空间线性可分，在这个高维空间中构造最优分类超平面，因此 SVM 也可称为大间距分类器：把正负样本以最大的距离分开[13]。当训练数据线性不可分时，对每个数据样本引入一个松弛变量 $\zeta \geqslant 0$ 和一个惩罚参数 $c \geqslant 0$ 后得到以下公式：

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2 + c\sum_{i=0}^{n}\zeta_i \tag{7}$$

$$\text{s.t. } y_i(\omega\cdot x_i + b) \geqslant 1 - \zeta_i, i = 1,2,3,\cdots,n$$
$$\zeta_i \geqslant 0,\ i = 1,2,3,\cdots,n \tag{8}$$

根据对偶性算法得到公式：

$$\min_{\omega,b}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_i\,\partial_j\,y_i\,y_j K(x_i\cdot x_j) - \sum_{i=1}^{n}\partial_i \tag{9}$$

$$\text{s.t. } \sum_{i=1}^{n}\partial_i\,y_i = 0$$
$$0 \leqslant \partial_i \leqslant c, i = 1,2,3,\cdots,n \tag{10}$$

再利用核函数通过数据样本和相似性函数来定义新的特征变量，将原空间的数据映射到新的空间，从而训练复杂的非线性边界，本质上是将其转化为了线性问题。其中 $K(x_i,x_j)$ 为核函数，本文选择高斯核函数[10]。参数 c 权衡模型准确性和复杂性，c 值越小，会有所下降。参数 ζ 用于调整模型复杂度。ζ 值越小，高斯分布越窄，模型复杂度越低；ζ 值越高，高斯分布越宽，模型复杂度越大。经过多次参数调整，确定 $c = 2, \zeta = 0.1$ 时分类效果最优。

3 实验

3.1 准备工作

本文实验在 Windows10 操作系统下进行。为了验证本文所提出方法(The proposed Model)的有效性，本文选择传统的 TFIDF 模型、LDA 主题模型和基于深度学习的 Word2Vec 模型进行对比试验。

在数据集的选取基础上，本文从国内腾讯新闻平台针对此次疫情设立的辟谣板块上抓取了10 073 条疫情相关样本数据，以上疫情相关数据均已表明该条数据是否属实。

在文本预处理上，本文在去掉文档的标点符号后，提取与正文相关内容，采用 THULAC[14] 分词工具，对正文进行分词，该分词工具具有识别能力强、准确率高、分词速度快等优点。分词后将得到的数据输出到一个文件中供实验模型训练，通过预处理将语料库数据转化成相关模型可以直接处理的数据。

词向量的维度方面，本文分别采用 $S(Size) = [100,200,300,400,500]$ 种维度进行对比实验。最终数据将被分为 2 种主题，即：真或假。

3.2　评价指标

分类的结果一般从分类器的准确度和速度两个方面来评判。运算速度主要由算法的时间复杂度和空间复杂度决定，准确度的衡量标准为准确率（*Precision*）、召回率（*Recall*）和 F_1 值。这里需用到的数学公式可写为：

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R} \tag{13}$$

其中，*TP*（True Positives）表示在标签中为正，实际也被分为正类；*TN*（True Negatives）表示在标签中为负，实际也被分为负类；*FP*（False Positives）表示在标签中为负，实际被分为正类；*FN*（False Positives）表示在标签中为正，实际被分为负类。

3.3　实验结果

各方法的准确率、召回率和 F_1 值结果如图 3～图 5 所示。各方法在维度为 400 的情况下，基本都取得了最优效果。由图 3、图 4 可以看出，LDA 主题模型相比传统的 TFIDF 模型在各个维度的准确率上均有 5%以上的提升，而与 Word2Vec 模型相比，LDA 主题模型在准确率、召回率和 F_1 上都有一定的差距，这也说明了，深度学习在文本分类中的有效性。另外，本文提出的方法对比效果最好的基线方法 Word2Vec 在各维度均有一定的优势，在 $K = 400$ 的情况下，本文提出的方法在准确率、召回率和 F_1 上分别有着 9.46%、5.47%和 5.62%的提升。这也切实说明了缺乏语义信息的 TFIDF 模型与 Word2Vec 模型结合后，2 种模型相辅相成，能够很好地提升文本表达的效果。从而证明本文提出的方法在对疫情虚假信息鉴别的准确度上有一定的优势。

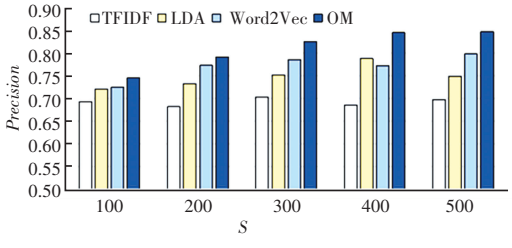

图 3　各方法的准确率

Fig. 3　The accuracy of each method

除此之外，本文选取了传统的 K 邻近（KNN）和决策树（DTC）[15] 两种分类算法作为分类器的对比试验。由上文可知在维度为 400 的情况下，各方法相对取得最好的效果，因此选择 $S = 400$ 时，各方法在不同分类器下的 F_1 进行对比。

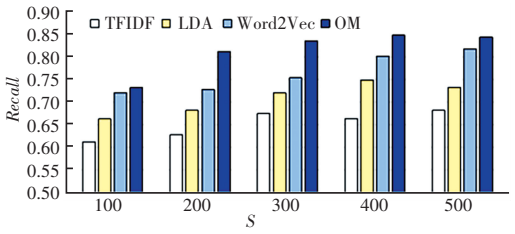

图 4　各方法的召回率

Fig. 4　The recall of each method

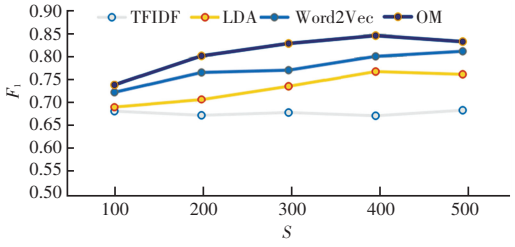

图 5　各方法的 F_1

Fig. 5　F_1 of each method

各分类方法在不同分类器上的表现见表 1。由表 1 可知，传统的 TFIDF 模型在 3 种分类器上表现的差异并不明显，这与其缺乏语音信息相关；其次可以看出，各模型在 SVM 分类器上的总体表现要优于其他 2 种分类器，这是由于本文的分类主题较少，对于疫情的相关信息，只能区分其真假，因此分类主题为二，而 SVM 模型在面对主题少的数据时具有很好的分类效果；研究同时还发现，本文提出的方法在 SVM 随着训练数据集的增加 SVM 分类器上得到的准确率。召回率和 F_1 比 KNN 分类器下的效果有 5.11%、6.81%、4.19%的改进。与 DTC 分类器相比，分别有 12.07%、16.97%、12.50%的提升，这是因为 DTC 模型更适用于分类主题多的数据，在面对分类主题少的数据时，其表现较为普通。

表 1　各分类方法在不同分类器上的表现

Tab. 1　The performance of each classification method on different classifiers

分类器	分类方法	*Presion*	*Recall*	F_1
DTC	TFIDF	0.659	0.664	0.667
	LDA	0.664	0.631	0.716
	Word2Vec	0.702	0.698	0.736
	OM	0.754	0.725	0.752
KNN	TFIDF	0.671	0.632	0.653
	LDA	0.743	0.667	0.759
	Word2Vec	0.761	0.773	0.795
	OM	0.804	0.794	0.812
SVM	TFIDF	0.686	0.664	0.672
	LDA	0.788	0.749	0.768
	Word2Vec	0.772	0.802	0.801
	OM	0.845	0.848	0.846

通过大量的实验得出本文提出的方法相较于传统的方法,对虚假信息具有较高的识别率,具体来说在准确率、召回率和 F_1 上均有 5%以上的提升,因而具有一定的实用价值。

4 结束语

针对当下虚假信息检测时常出现的识别度低的问题,本文提出了一种基于 Word2Vec 的虚假信息检测方法。该方法利用 Word2Vec 模型引入传统向量空间模型不具备的语义特征,同时解决以往向量空间模型特征稀疏的问题。再针对 Word2Vec 模型无法很好地度量文本间的语义相似度的问题,利用 TFIDF 模型对 Word2Vec 模型进行加权融合,最后再利用 SVM 模型优越的二分类能力,以此来区分真假信息。通过相关实验得出本文方法对虚假信息辨别有着极高的准确率,具有良好的性能。

在接下来的工作中,因为 TFIDF 模型具有很强的表示文本能力,因而还需要做进一步的深入研究。所以考虑利用基于词向量距离的文本分类方法,例如用 EMD 距离度量方式来平衡词与词之间的相似度。后续也将继续探索效果更精准的疫情虚假信息的检测方法。

参考文献

[1] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975,18: 613-620.

[2] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Reseach, 2003, 3(4-5): 993-1022.

[3] BINGHAM E, MANNILA H. Random projection in dimensionality reduction: Applications to image and text data[C]// Proceedings of The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California: ACM, 2001:245-250.

[4] 方东昊. 基于 LDA 的微博短文本分类技术的研究与实现[D]. 沈阳:东北大学,2011.

[5] KIM Y. Convolutional Neural Networks for sentence classification [J]. arXiv preprint arXiv:1408.5882,2014.

[6] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. Lake Tahoe CA:NIPS,2013:3111-3119.

[7] MIKOLOV T, YIH W, ZWEIG G. Hltnaac L. Linguistic regularities in continuous space word representations [C]// Proceedings of NAACL-HLT 2013. Atlanta, Georgia: ACL, 2013: 746-751.

[8] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Proceedings of the International Conference on Learning Representations (ICLR 2013). Scottsdale, AZ:dblp,2013:1-12.

[9] PATEL B, SHAH D. Significance of stop word elimination in meta search engine[C]//IEEE International Conference on Intelligent Systems and Signal Processing. Vallabh Vidyanagar, India: IEEE, 2013:52-55.

[10] 邓君,孙绍丹,王阮,等. 基于 Word2Vec 和 SVM 的微博舆情情感演化分析[J]. 情报理论与实践,2020,43(8):112-119.

[11] 朱磊. 基于 Word2Vec 词向量的文本分类研究[D]. 重庆:西南大学,2017.

[12] 张谦,高章敏,刘嘉勇. 基于 Word2Vec 的微博短文本分类研究[J]. 信息网络安全,2017(1):57-62.

[13] 陈武,梁刚,杨进. 一种改进的 SVM 算法在入侵检测中的应用[J]. 计算机安全,2013(6):2-7.

[14] 孙茂松,陈新雄,张开旭,等. THULAC:一个高效的中文词法分析工具包[Z]. 北京:清华大学,2016.

[15] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter [C]//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India: ACM, 2011: 675-684.

(上接第 133 页)

参考文献

[1] 孙红,王瑞琪,付东翔,等. 新工科的兴起与智能复合人才培养的研究[J]. 计算机教育,2019(10):27-30.

[2] 中华人民共和国教育部. 教育部关于印发《高等学校人工智能创新行动计划》的通知[EB/OL]. [2018-04-03]. http://www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html.

[3] 陈志勇,叶桦畅,张笑钦. 计算机类专业的课程思政:核心元素、基本原则与实施策略[J]. 中国大学教学,2021(4):34-38,65.