基于 HybridDL 模型的文本相似度检测方法*

肖 晗[1],毛雪松[1],朱泽德[2]
(1.武汉科技大学 信息科学与工程学院,湖北 武汉 430081;
2.中科院合肥技术创新工程院,安徽 合肥 230031)

摘 要:为了提高文本相似度检测算法的准确度,提出一种结合潜在狄利克雷分布(Latent Dirichlet Allocation,LDA)与 Doc2Vec 模型的文本相似度检测方法,并把该算法得到的模型命名为 HybridDL 模型。该算法通过 Doc2Vec 对文档训练得到文档向量,再利用 LDA 模型得到文档主题与各个主题下特征词出现的概率,对文档中各主题及特征词计算概率加权和,映射到 Doc2Vec 文档向量中。实验结果表明,新算法模型比传统的 Doc2Vec 模型对相似文本的判断更加敏感,在文本相似度检测上具有更高的准确度。

关键词:Doc2Vec;潜在狄利克雷分布;文本表示;文本相似度
中图分类号:TN957.52;TP391.1 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.191257

中文引用格式:肖晗,毛雪松,朱泽德. 基于 HybridDL 模型的文本相似度检测方法[J].电子技术应用,2020,46(6):28-31,35.
英文引用格式:Xiao Han,Mao Xuesong,Zhu Zede. Text similarity detection method based on HybridDL model[J]. Application of Electronic Technique,2020,46(6):28-31,35.

Text similarity detection method based on HybridDL model

Xiao Han[1],Mao Xuesong[1],Zhu Zede[2]
(1.School of Information Science and Engineering,Wuhan University of Science and Technology,Wuhan 430081,China;
2.Institute of Technology Innovation,Hefei Institutes of Physical Science,Chinese Academy of Sciences,Hefei 230031,China)

Abstract:In order to improve the accuracy of text similarity detection algorithm, this paper proposes a text similarity detection method combining latent Dirichlet Allocation(LDA) and Doc2Vec model, and names the model obtained by the algorithm HybridDL model. This algorithm obtains the document vector through Doc2Vec training of the document, and then obtains the probability of the occurrence of the document topic and the feature words under each topic with the LDA model, calculates the probability weighted sum of each topic and feature words in the document, and maps them to the Doc2Vec document vector. Experimental results show that the new algorithm model is more sensitive to the judgment of similar text than the traditional Doc2Vec model, and has higher accuracy in the detection of text similarity.

Key words:Doc2Vec;latent Dirichlet allocation;text representation;text similarity

0 引言

在当下这个信息时代,互联网已经成为人们生活中不可或缺的一部分,在机器计算能力大幅度提高的同时,获得的数据也呈爆炸式增长。文本数据作为数据中的重要组成部分,量大且关键。因此,从大量的文本数据中高效地提取出满足人们需要的信息成为了当下的热门话题。在自然语言处理领域,通过计算机处理文本数据时,由于语言的多样性,相同的词语在不同的句子或者语境中表达出来的意思可能会存在差异,导致计算机无法直接并准确地获取文本特征[1]。所以,如何从语料中学习到好的文本表示,如何提升文本表示模型的性能,对于后续自然语言处理的相关研究,如机器翻译、文本分类[2]、情感分析[3]、问答系统、文本检索等,具有十分深远的意义。

建立文本表示模型包括对词、主题、语句、文档等各个级别任务进行建模。对于词级别的文本表示模型,通常使用被称作词向量的数学表示方法来处理。词向量顾名思义是一种通过向量来表示句子中词语的方法,向量中的每一维都在实数范围内进行取值操作[4]。

词向量最早由 BENGIO Y、DUCHARME R、VINCENT P 等人提出[5],其传统做法是 One-hot 表示方法,即将不同词用相对应的维度很高的向量来表示,其中,向量的维度对应字典大小,在各个词的向量中只存在一个位置为 1,其余位置为 0。该表示模型十分简洁,便于理解,但是由于数据稀疏会造成维数灾难,并且该模型没有考虑词与词之间的关联性,准确度不高。

近年来,被称作词的分布式表示的向量表示得到了

* 基金项目:国家自然科学基金(61806187)

较为广泛的应用，理论思路是通过训练，将句子中的各词语映射到 N 维向量空间中。该方法在 One-hot 的基础上，联系了前后文的语义信息，使语义相近词语所映射得到的词向量比较接近，而 One-hot 法得到的是毫无关联的词向量。即可以通过词与词在空间中的距离计算词与词在语义上的相关性，距离越小则语义越相关，距离越大则越无关。2013 年，MIKOLOV T 等人提出利用神经网络模型来训练分布式词向量[6]，所得到的模型被称作 Word2Vec，该模型可以通过前后文的词汇预测中心词或者通过中心词来预测前后文的词汇。它相当于一个里程碑，现在也被广泛使用。Doc2Vec 是 Word2Vec 的扩展，于 2014 年由 MIKOLOV T 等人提出[7]，同样用于学习文档表示。该模型在构建的过程中，在获取上下文单词信息的同时，增加了一个段落标记，能够更精确地表示原始文本。但是在实际应用中 Doc2Vec 需要大量数据进行训练才能有较好的效果，当数据量不足时，提取信息不充分，结果产生的偶然性较大。

主题模型由于可以发掘深层次的语义信息，因此在构建文本表示模型时也可以达到较好的效果。2003 年 BLEI D M、JORDAN M I 等人提出了潜在狄利克雷分布[8]，首次将狄利克雷先验分布加入到文档、主题、词的多项式分布中，效果显著。LDA 是一种从大量文档中发现潜在主题的概率主题模型，它从文本的统计学特性入手，将文本语料库映射到各个主题空间中，从而发掘文本中各主题与词语之间的对应关系，得到文本的主题分布[9]。它通常被认为是一种通过对不同主题中的单词进行分组的特征约简方法，因此可以将文档映射到更低的维度空间。但 LDA 没有考虑词语的前后文关联，构建出的文本向量比较稀疏，在表示原始文本的信息方面效果一般。

本文尝试将 LDA 和 Doc2Vec 进行融合。LDA 从每个文档到所有主题的全局关系建模，而 Doc2Vec 则通过从目标单词的上下文中学习来捕获这些关系。发挥这两种模型各自的优点，从而产生比传统模型更高的准确率判断。

1 背景算法介绍

1.1 LDA 模型

LDA 模型是一种主题模型，它将语料库中每篇文档的主题表示为概率分布的形式[10]。同时它也是一种无监督学习算法，在训练时不需要手工标注训练集，通过对文本中隐含的主题信息进行建模，描述相关文本，保留其本质的统计信息。该方法能更加高效地处理数据量大的语料库。LDA 模型是在概率隐性语义索引(Probabilistic Latent Semantic Analysis，PLSA)的基础上加层贝叶斯框架，它有 3 层生成式贝叶斯网络结构，分别为文档层、主题层和特征词层，其拓扑结构如图 1 所示。

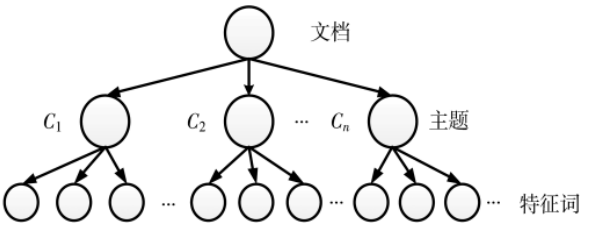

图 1 LDA 模型隐含主题的拓扑结构示意图

LDA 模型由参数(α,β)确定，其中 α 表示文档集合中隐含主题之间的相对强弱，β 表示所有隐含主题自身的概率分布。LDA 模型图如图 2 所示，其中 θ_i 表示文档

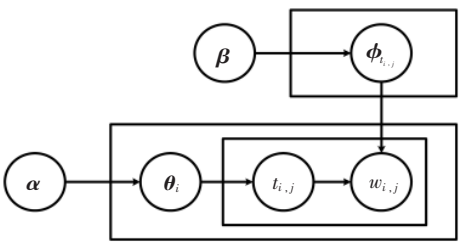

图 2 LDA 模型图

主题的概率分布，$\phi_{t_{i,j}}$ 表示特定主题下特征词的概率分布，t 表示主题，w 表示特征词。

LDA 概率主题模型生成文本的过程如下：

(1)按照先验概率 $P(d_i)$选择一篇文档 d_i；

(2)从 Dirichlet 分布 α 中取样生成文档 d_i 的主题分布 θ_i，即主题分布 θ_i 由超参数为 α 的 Dirichlet 分布生成；

(3)从主题的多项式分布 θ_i 中取样生成文档 d_i 第 j 个词的主题 $t_{i,j}$；

(4)从 Dirichlet 分布 β 中取样生成主题 $t_{i,j}$ 对应的词语分布 $\phi_{t_{i,j}}$，即词语分布由超参数为 β 的 Dirichlet 分布生成；

(5)从词语的多项式分布 $\phi_{t_{i,j}}$ 中采样最终生成词语 $w_{i,j}$；

(6)利用吉布斯采样[11]，可以在文本集已知的条件下，使用参数估计获取参数值，即通过 θ 和 ϕ 可以推断出文档中隐含的主题信息，并预测出任何具有主题比例分布的新文档。

1.2 Doc2Vec 模型

Doc2Vec 是在 Word2Vec 模型基础上衍生出来的模型，Word2Vec 有两种模型，分别是 CBOW 模型和 Skip-gram 模型。通过神经网络进行训练，前者利用中心词前后 c 个词来预测中心词，后者利用中心词去预测该词前后 c 个词[12]。相对于 Word2Vec，Doc2Vec 多了一个段落标记。段落标记可以看作词汇的一种表现形式，它能记住前后文或相关文档主题中遗漏的内容[7]。因此，该模型通常被称为段落向量的分布式存储模型(PV-DM)，如图 3 所示，该模型是 CBOW 模型的衍生，通过上下文的单词预测中心词。

另一种叫作分布式词袋的文档向量模型(PV-DBOW)如图 4 所示，该模型是 Skip-gram 模型的衍生，通过训练段落标记，实现对小窗口范围内上下文单词的预测。

Doc2Vec 与 Word2Vec 模型相比，在输入层增加了段

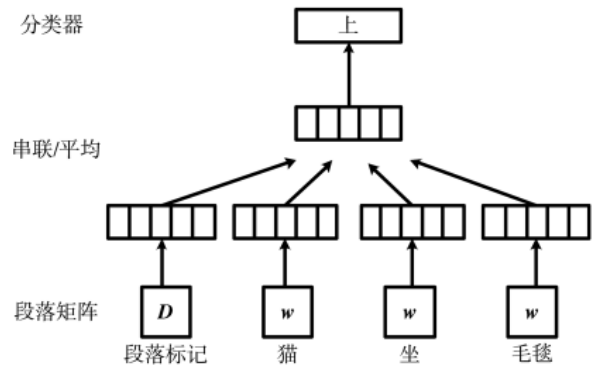

图 3 PV-DM 模型图

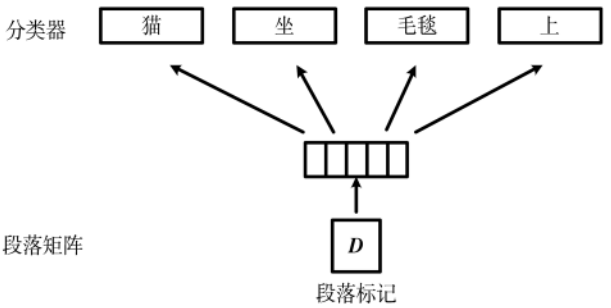

图 4 PV-DBOW 模型图

落标记，该标记是相应段落的向量表示，与词向量进行串联或取平均值作为输入。算法本身有两个关键步骤：(1)在已知的文档集中，通过训练，得到词向量 w、softmax 权重 U、b 和文档向量 D；(2)在推断阶段，在保持 w、U、b 不变的情况下对 D 进行梯度下降，增加更多的 D 的列来获取新的文档向量 D。文档向量的一个重要优势是它们可以从未标记的数据中学习，因此可以很好地用于没有足够标记数据的任务。文档向量还解决了词袋模型的一些关键缺陷。首先，文档向量获取了单词的语义信息，这就表示词义相近的词在向量空间中会更加接近；其次，文档向量和高纬度 n-gram 模型[13]类似，保留了段落的大量信息，包括单词顺序。但 n-gram 模型会创建一个非常高维的表示方法，所以这种表示方法往往不能很好地推广。简而言之，Doc2Vec 比 n-gram 词袋模型具有更高的性能。

2 HybridDL 模型

LDA 模型是通过概率的形式对主题和其对应的特征词进行抽取，从而来表示文档。构建出来的文本向量比较稀疏，具有一定的不确定性。为了更准确地表达出词语之间的语义关系，本文将 LDA 模型与 Doc2Vec 模型进行融合，同时也使 Doc2Vec 文本表示模型对主题信息更加敏感。新方法如图 5 所示，将单词、文档和主题投射到高维语义空间中。文档向量被视为单个向量，它是文档中所有单词的质心。在构造主题向量的过程中，在每个主题中使用一个高概率单词子集来表示主题，然后将它们的概率重新标为单词的权重。因此，不同的词汇对

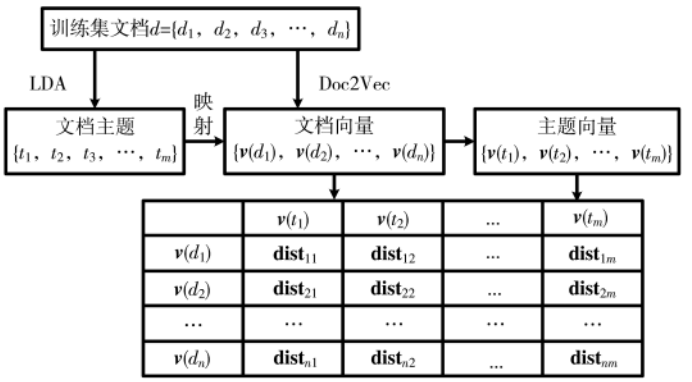

图 5 HybridDL 模型步骤图

主题有不同的贡献。测量每个文档到主题的欧氏距离，以便用距离分布表示文档。

HybridDL 模型的具体构建步骤如下：

(1)给出训练集 $d\{d_1, d_2, d_3, \cdots, d_n\}$。

(2)通过 LDA 模型训练 d，得到主题 $\{t_1, t_2, t_3, \cdots, t_m\}$ 以及在每个主题下各特征词出现的概率。

(3)通过 Doc2Vec 训练得到文档向量 $\{\boldsymbol{v}(d_1), \boldsymbol{v}(d_2), \boldsymbol{v}(d_3), \cdots, \boldsymbol{v}(d_n)\}$。

选取主题 t_i 中 h 个高频率词出现的概率 $\boldsymbol{\phi}_i$，通过式(1)标注权重。

$$\omega_i = \frac{\boldsymbol{\phi}_i}{\sum_{n=1}^{h} \boldsymbol{\phi}_n} \tag{1}$$

主题向量由高频词在 Doc2Vec 空间下的坐标乘上相应权重获得。

$$\boldsymbol{v}(t_i) = \sum_{n=1}^{h} \omega_{i_n} \boldsymbol{w}(d_i) \tag{2}$$

每个文档都可以表示为一个语义空间中从文档到所有主题的距离分布，其中文本向量与单个主题向量计算距离公式为：

$$\text{distance}(\boldsymbol{v}(d_i), \boldsymbol{v}(t_i)) = |\boldsymbol{v}(d_i) - \boldsymbol{v}(t_i)| \tag{3}$$

在实验过程中，若需要求文本 d_i 与文本 d_2 的相似度，训练完模型后，输入文本 d_1，通过 HybridDL 模型得到 $\boldsymbol{v}(t_1)$，通过 Doc2Vec 模型得到 $\boldsymbol{v}(d_1)$，通过式(3)计算出每个主题向量与文本向量的距离，再将其整合得到距离分布向量 $\mathbf{dist}(\boldsymbol{v}(d_1), \boldsymbol{v}(t_1))$，即图 5 中 $\mathbf{dist}_{11}$，从而用来表示文本 d_1。再输入文本 d_2，进行同样的操作，得到 $\mathbf{dist}(\boldsymbol{v}(d_2), \boldsymbol{v}(t_2))$，对这两个距离分布向量求余弦相似度，如式(4)所示。

$$\begin{aligned} \text{Sim} &= \text{sim}(\mathbf{dist}(\boldsymbol{v}(d_1), \boldsymbol{v}(t_1)), \mathbf{dist}(\boldsymbol{v}(d_2), \boldsymbol{v}(t_2))) \\ &= \frac{\mathbf{dist}(\boldsymbol{v}(d_1), \boldsymbol{v}(t_1)) \cdot \mathbf{dist}(\boldsymbol{v}(d_2), \boldsymbol{v}(t_2))}{|\mathbf{dist}(\boldsymbol{v}(d_1), \boldsymbol{v}(t_1))||\mathbf{dist}(\boldsymbol{v}(d_2), \boldsymbol{v}(t_2))|} \end{aligned} \tag{4}$$

而传统利用 Doc2Vec 模型求文本相似度的方法是直接求 $\boldsymbol{v}(d_1)$、$\boldsymbol{v}(d_2)$ 的余弦相似度。

3 实验设计与结果分析

为了验证本文提出的方法的有效性，选用了搜狗语料库与网易新闻语料库，并将语料库分为健康、教育、金

融、军事、旅游、文化、医学、娱乐8类。获取了原始文本之后，通过jieba分词对文本进行分词与去除停用词的操作，命名为T1、T2、T3、T4、T5、T6、T7、T8。使用的语料集与文本数如表1所示。

在8类语料库中各类分别选10组文本，一共80组文本作为测试语料库，测试语料库中的每组语料经人工判定评为56组相似与24组不相似，将其输入到各个模型中，Doc2Vec训练模型采用PV-DM模型，窗口设置为5，向量维度设置为100，得到相似度的数值，部分数据如表2所示。

表1 训练语料库文本数

组号	类别	文本数
T1	健康	4 672
T2	教育	11 354
T3	金融	4 532
T4	军事	4 612
T5	旅游	5 105
T6	文化	4 976
T7	医学	2 857
T8	娱乐	44 349

表2 测试集部分语料相似度

组号	文本1	文本2	Doc2Vec相似度	HybridDL相似度
1	S1	S2	0.778 5	0.790 3
2	S3	S4	0.924 7	0.940 3
3	S5	S6	0.648 6	0.576 1
4	S7	S8	0.868 9	0.889 3
5	S9	S10	0.713 3	0.741 2
6	S11	S12	0.675 4	0.692 2
7	S13	S14	0.614 7	0.590 0
8	S15	S16	0.657 7	0.673 1

根据表2可以看出，通过HybridDL模型得到的文本相似度比起传统的Doc2Vec模型来说对于相似的文本判断得更为准确。

经实验，将判定相似度的阈值设置为0.65，采用准确率P、召回率R、F值来考察模型性能。

先得到FN、FP、TN、TP的值，再通过式(5)~式(7)分别计算出准确率P、召回率R以及F值。

FN、FP、TN、TP定义如下：

FN：被判定为不相似文本，但实际上是相似文本；

FP：被判定为相似文本，但实际上是不相似文本；

TN：被判定为不相似文本，实际上也是不相似文本；

TP：被判定为相似文本，实际上也是相似文本。

准确率：

$$P=\frac{TP}{TP+FP} \tag{5}$$

召回率：

$$R=\frac{TP}{TP+FN} \tag{6}$$

F值：

$$F=\frac{2PR}{P+R} \tag{7}$$

实验效果对比如表3所示。

表3 Doc2Vec模型与HybridDL模型相似度检测效果

模型	P	R	F
Doc2Vec模型	0.891 9	0.589 3	0.709 7
HybridDL模型	0.902 4	0.660 7	0.762 9

表3实验结果表明，采用HybridDL模型计算文本相似度在召回率和F值上有明显提升。主要原因在于HybridDL模型不仅能准确分析出文档中词项的语义信息，还可以得到文档中的主题分布情况，这些因素对在不同主题下的文本相似度检测很有帮助。

4 结论

本文提出了HybridDL模型，利用Doc2Vec对文档训练得到文档向量，训练后的结果更能表达词语之间的语义，并利用LDA模型提取出文档的主题及其特征词。再将LDA主题词加权求和映射到Doc2Vec文档向量中，得到新的主题向量。每个文档都可以表示为一个语义空间中从文档到所有主题的距离分布。实验结果表明，HybridDL模型对比传统的Doc2Vec模型计算文本相似度在召回率和F值上有明显提升。该模型的提出，提供了一种新的文档向量表示方法思路，能有效提升文本相似度计算的准确率。在后续工作中将使用更加复杂的神经网络结构并且训练更多语料库进行进一步研究。

参考文献

[1] 冀宇轩.文本向量化表示方法的总结与分析[J].电子世界，2018，22(3)：10-12.

[2] 殷晓雨，阿力木江·艾沙，库尔班·吾布力.基于卷积递归模型的文本分类研究[J].电子技术应用，2019，45(10)：29-32，36.

[3] MAAS A L，DALY R E，PHAM P T，et al.Learning word vectors for sentiment analysis[C].The Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies.Portland，Oregon：Association for Computational Linguistics，2011.

[4] 幸凯.基于卷积神经网络的文本表示建模方法研究[D].武汉：华中科技大学，2017.

[5] BENGIO Y，DUCHARME R，VINCENT P，ed al.A neural probabilistic language model[J].Journal of Machine Learning Research，2003，3：1137-1155.

[6] MIKOLOV T，CHEN K，CORRADO G，et al.Efficient estimation of word representations in vector space[C].International Conference on Learning Representations.Portland，Oregon：Association for Computational Linguistics，2013.

[7] QUOC L，MIKOLOV T.Distributed representations of sentences and documents[C].The Proceedings of the 31st International Conference on Machine Learning.Portland，Oregon：Association for Computational Linguistics，2014.

[8] BLEI D M，NG A Y，JORDAN M I.Latent dirichlet allocation[J].Journal of Machine Learning Research，2003，3：

(下转第35页)

LBP 纹理特征以及使用卷积神经网络 CNN 的人脸活体检测方法进行对比，结果如表 2 所示。表中 FRR(False Reject Rate)为错误拒绝率，FAR(False Accept Rate)为错误接受率，HTRE(Half Total Error Rate)为半错误率。半错误率指的是错误接受率和错误拒绝率总和的一半。

表 2 跨数据集活体检测准确率对比

检测方法	FRR/%	FAR/%	HTRE/%
LBP	11.4	80.3	45.9
CNN	47.2	44.3	45.7
改进方法	0.5	10.02	5.26

通过对比发现，LBP 和 CNN 检测方法都取得了较高的半错误率，表明在进行跨数据集活体检测时这两种方法的泛化性能较差。而改进方法的 HTER 值为 5.26%，表明其检测准确率高于前两种方法。

4 结论

本文针对现有活体检测方法在跨数据集时泛化能力较差从而导致检测准确率低的问题，提出了基于检测环境中的上下文信息并使用单类支持向量机对检测对象进行分类的活体检测方法。通过实验发现，改进方法的泛化能力优于 LBP 和 CNN 方法。但在某些活体检测场景中可能没有可以利用的上下文信息，此时如何实现准确率较高的跨数据集活体检测将是后续研究的方向。

参考文献

[1] 孙霖.人脸识别中的活体检测技术研究[D].杭州：浙江大学，2010.

[2] PAN G，SUN L，WU Z，et al.Eyeblink-based anti-spoofing in face recognition from a generic webcamera[C].IEEE 11th International Conference on Computer Vision，IEEE，2007：1-8.

[3] LI J，YU W，KUANG G Y，et al.A compound face recognition system design[J].Journal of National University of Defense Technology(China)，2003，25(3)：45-48.

[4] MÄÄTTÄ J，HADID A，PIETIKÄINEN M.Face spoofing detection from single images using micro-texture analysis[C].2011 International Joint Conference on Biometrics(IJCB).IEEE，2011：1-7.

[5] TIAGO D F P，ANDRÉ A，JOSÉ MARIO D M，et al.LBP-TOP based countermeasure against face spoofing attacks[M].Computer Vision-ACCV 2012 Workshops.Springer Berlin Heidelberg，2013.

[6] KIM W，SUH S，HAN J J.Face liveness detection from a single image via diffusion speed model[J].IEEE Transactions on Image Processing，2015，24(8)：2456-2465.

[7] Yang Jianwei，Lei Zhen，LI S Z.Learn convolutional neural network for face anti-spoofing[J].Computer Science，2014，9218：373-384.

[8] 李冰，王宝亮，由磊，等.应用并联卷积神经网络的人脸防欺骗方法[J].小型微型计算机系统，2017，38(10)：2187-2191.

[9] 黄海新，张东.基于深度学习的人脸活体检测算法[J].电子技术应用，2019，45(8)：44-47.

[10] 黄睿，陆许明，邬依林.基于 TensorFlow 深度学习手写体数字识别及应用[J].电子技术应用，2018，44(10)：12-16.

[11] 尹传环，牟少敏，田盛丰，等.单类支持向量机的研究进展[J].计算机工程与应用，2012，48(12)：1-5.

[12] 刘学艺，李平，郜传厚.极限学习机的快速留一交叉验证算法[J].上海交通大学学报，2011，45(8)：1140-1145.

(收稿日期：2019-12-09)

作者简介：

闫龙(1989-)，男，硕士研究生，主要研究方向：人脸活体检测、机器学习、软件架构。

胡晓鹏(1972-)，男，博士，副教授，主要研究方向：软件架构、机器学习。

(上接第 31 页)

993-1022.

[9] 王振振，何明，杜永萍.基于 LDA 主题模型的文本相似度计算[J].计算机科学，2013，4(12)：229-232.

[10] 曾祥坤，张俊辉，石拓，等.基于主题提取模型的交通违法行为文本数据的挖掘[J].电子技术应用，2019，45(6)：41-45.

[11] 徐佳俊，杨飏，姚天昉，等.基于 LDA 模型的论坛热点话题识别和追踪[J].中文信息学报，2016，30(1)：43-49.

[12] 唐明，朱磊，邹显春.基于 Word2vec 的一种文档向量表示[J].计算机科学，2016，43(6)：214-217.

[13] BROWN P F，DESOUZA P V，MERCER R L.Class-based n-gram models of natural language[J].Journal Computational Linguistics，1992，18(4)：467-479.

(收稿日期：2019-11-20)

作者简介：

肖晗(1994-)，男，硕士研究生，主要研究方向：自然语言处理。

毛雪松(1975-)，男，博士，教授，主要研究方向：智能驾驶路径规划、环境信息感知和智能光设备。

朱泽德(1985-)，男，博士，副研究员，主要研究方向：自然语言处理。

版权声明