基于改进的 TextRank 算法的计算机辅助定密研究

李晨庚[1] 谢四江[2]

[1](西安电子科技大学 陕西 西安 710071)

[2](北京电子科技学院 北京 100070)

摘 要 针对传统定密方式定密不严谨、定密尺度难以把握、经验难以积累等问题,提出基于改进的 TextRank 算法的计算机辅助定密方法,该方法通过定密规则的词性特点,将句向量分解为名词向量和非名词向量,构造基于词性的句向量,利用改进的 TextRank 算法对文档语句排序,获取在定密细则影响下的关键语句权重,计算文档密级分数,判断文档密级。实验结果表明,该方法比目前传统定密方式准确率有所提高。

关键词 计算机辅助定密 句向量 图模型

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.03.053

RESEARCH ON COMPUTER-AIDED SECRET-LEVEL CLASSIFICATION BASED ON IMPROVED TEXTRANK ALGORITHM

Li Chengeng[1] Xie Sijiang[2]

[1](*Xidian University, Xi'an 710071, Shaanxi, China*)

[2](*Beijing Electronic Science and Technology Institute, Beijing 100070, China*)

Abstract Aimed at the problems that the traditional document secret-level classification is not rigorous, its classification scale is difficult to grasp, and the experience is difficult to accumulate, an algorithm of computer-aided secret-level classification based on improved TextRank algorithm is proposed. According to the part-of-speech characteristics of the classification, this method decomposed the sentence vectors into noun vectors and non-noun vectors, and constructed the part-of-speech based sentence vectors. The improved TextRank algorithm was used to sort the document sentences, obtain the key sentence weights under the influence of this secret rules, calculate the document secret score, and judge the document secret-level. Experimental results show that the accuracy of this algorithm is higher than the current traditional secret-level classification method.

Keywords Computer-aided secret-level classification Sentence vector Graph model

0 引 言

保密工作是各国长久以来非常重要的国家安全基础工作。自《保密法》颁布以来,全国各单位与部门都依照保密法规定,制定相关保密细则,完善保密管理制度。不过由于各单位与部门的情况、领域各不相同,涉密事项种类繁多,定密方式单一,保密管理流程有所差异。定密大都依赖人工定密,指定专人负责文档的接收、制定定密细则、进行文档定密,以及分类管理。这样传统的定密方式造成定密工作效率低下,质量无法保证。

目前对定密工作的研究集中于定密制度、流程及管理上。高波等[1]研究对比不同国家定密制度,介绍和分析了美军的定密制度。陈翠玲等[2]总结了国内定密发展现状,结合借鉴国外先进理论,提出了改进国内定密工作的措施与方案。

在定密方式上,国内对定密工作的智能化与数字化研究并不多。张帆等[3]从基于可信度的不确定推理模型入手,提出基于可信度的不确定推理辅助定密方

收稿日期:2020-01-13。**李晨庚**,硕士,主研领域:信息安全,机器学习。**谢四江**,教授。

法,该方法通过专家打分的方式对从定密法规中抽取定密规则设定可信度,计算规则匹配后相关密级的可信度来确定涉密文档所属密级,这开启了计算机辅助定密方面的研究。吴国华等[4]提出了根据文档相似度快速查找定密依据的方法,基于 VSM 模型和匈牙利算法的文本相似度计算,确定最终文档密级。潘娅[5]提出一种基于中文文本分类技术的计算机辅助密级界定方法,将中文文本分类技术应用到文档定密任务上,通过对文档进行预处理,并用 TF-IDF 进行文档关键词权重计算,完成向量化过程,针对定密任务的特点,提出基于二叉树的多分类 SVM,完成计算机辅助定密,提高了定密的准确性。在定密方式的研究上,发展趋势是从人工定密转向计算机辅助定密,在定密方法上是从基于统计的方式到基于机器学习的方式。

在实际定密工作中,目前大多数采用的是基于关键词的方式。根据定密细则中的涉密关键词与文档进行比对,文档中出现涉密关键词和词对越多,则说明文档越有可能被定为相应密级。这种方式忽略了文档中会出现与关键词相似表达的问题,虽然文档中不含有涉密关键词,但语句表示的含义仍是涉密的。

综上,为解决以上辅助定密所产生的定密不严谨、效率低下等问题,本文提出了基于改进的 TextRank 算法的计算机辅助定密方法,主要做了以下几个工作:

1) 利用大规模预训练网络生成的 glove 词向量,基于词性构造具有语义信息的加权句向量。

2) 对待定密文档解析生成句子节点,并构造图模型,将定密细则融入图节点中,利用改进的 TextRank 算法进行排序计算,获取关键语句权重,进行辅助密级界定。

1　相关工作

1.1　词向量发展概况

在自然语言处理领域,文本作为非结构化的数据结构,要进行相关的计算就必须经过预处理转换为数值型数据,预处理的过程主要包括分词、去停用词、文本向量化等操作。其中如何进行文本向量化是自然语言处理领域非常热门的研究热点。词向量质量的好与坏对于下游文本分类等任务也起到非常大的作用。最初的文本向量化方式比较直接,有独热编码(One-hot)、词袋模型(Bag-of-words model)、TF-IDF(term frequency-inverse document frequency)等向量化方式。One-hot 将词语看作独立的存在,互相没有联系,也不包含任何语义语法信息,并且在词汇表非常大的时候造成维度灾难。Bow 和 TF-IDF 则是考虑了词语的分布情况,利用词语的统计特征构造向量。TF-IDF 不仅可以用来进行文本向量化,也可以用于关键词计算。但是,这些词向量表示方式都比较稀疏,并且忽略了语义信息。

1986 年 Hinton[6]提出词向量的分布式表示方式,这开启了新一次的研究热潮,其中最为经典的就是 Mikolov 等[7]提出的 Word2vec 算法。Word2vec 算法假设两个具有相似上下文的词语表达的意思也相近。根据这个假设,设计出 cbow 模型和 skip-gram 模型,cbow 模型通过上下文来预测中心词,skip-gram 模型通过中心词预测上下文,并在其之后的研究中优化了词向量生成方式,加入负采样及子采样等方式。词向量作为副产物在模型中生成。相比于 Word2vec 只考虑窗口大小的信息,Pennington 等[8]提出的 glove 词向量,除了考虑窗口大小的信息,还加入了全局信息,这使得 glove 词向量在质量上有所提高。

Bengio 等[9]提出了一种 N-gram 神经概率语言模型,通过多层前馈神经网络进行学习,利用极大似然法预测在该上下文条件下的当前中心词的条件概率。Peters 等[10]是神经语言模型中比较经典的模型,采用双向 LSTM 来进行监督学习,获取关于上下文的动态向量。

近几年词向量研究基本集中在大规模预训练网络上,以监督学习方式进行模型训练,然后通过 fine-turn 嫁接到任务网络之上。这个采用预训练方式进行文本向量化,不仅可以通过大规模的语料库获得词语统计信息,而且可以通过深度模型挖掘出语法语义等高级信息。Bert[11]是谷歌在 2018 年发布的大型预训练网络,在 11 项 NLP 任务中获得了 state-of-the-arts 的表现,Bert 是基于多层的 Transformer[12]的 Encoder-Decoder 模型。解决了传统神经语言模型无法长期依赖问题。并且通过 Masked 机制解决了“自己看到自己”的问题。本文的文本向量化考虑不同词向量质量以及不同任务下的表现,决定采用 glove 的预训练词向量进行句向量构造。

1.2　图排序模型

基于图模型的排序算法本质上是一种从整个图递归得出的全局信息来确定图内节点重要性的方法。比较有名的就是谷歌创始人 Brin 等[13]提出的 PageRank 算法以及 Jon Kleinberg 提出的 HITS 算法。

PageRank 的基本思想是“投票”“推荐”,当一个顶点链接到另一个节点时,算作对另一个节点的投票,票

数越高,该节点的重要性就越大。节点进行投票的重要性决定了投票本身的重要性。一个节点的重要性来源于为其投票的节点,以及为其投票节点的重要性。

如图 1 所示,A 节点有三个出链分别链接到 B、C、D 上,那么当用户访问 A 节点的时候,就有可能跳转到 B、C 或者 D 节点,跳转概率均为 1/3。B 节点有两个出链,链接到 A 和 D 节点上,跳转概率为 1/2。

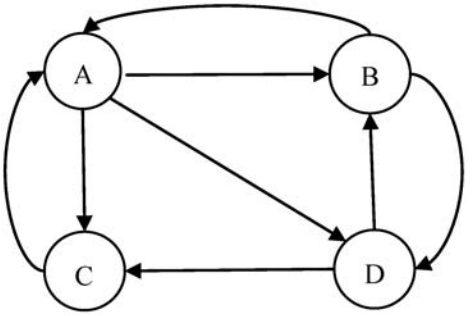

图 1 PageRank 节点示意图

抽象来说就是一个不带权的有向图 $G=(V,E)$,V 代表节点,E 代表节点之间的边,大小为 $|V\times V|$。对于给定节点 V_i,$ln(V_i)$ 表示所有链接至节点 V_i 的节点集合。$Out(V_i)$ 表示节点 V_i 所链接的节点集合。节点 V_i 的 PageRank 得分计算如下所示:

$$S(V_i)=(1-d)+d\times\sum_{j\in ln(V_i)}\frac{1}{|Out(V_j)|}S(V_j) \quad (1)$$

式中:d 是范围在 0~1 的阻尼系数,表示一个节点链接到另一个随机节点的概率,一般被启发式的设置为 0.85。然后通过迭代的方式进行计算,直至误差收敛到某一阈值以下。而对于无向图的排序来说只是令有向图中节点的入度等于出度即可。

HITS 是与 PageRank 同一时期提出的算法,全称是 Hyperlink-Induced Topic Search。在 HITS 算法中,每个节点被赋予两个属性:Hub 属性和 Authority 属性。同时节点被分为 Hub 节点和 Authority 节点。Hub 是中心的意思,所以 Hub 节点包含了许多指向 Authority 节点的链接。Authority 节点则是指包含实质内容的节点。公式如下所示:

$$HITS_{\mathrm{A}}(V_i)=\sum_{V_j\in In(V_i)}HITS_{\mathrm{H}}(V_j) \quad (2)$$

$$HITS_{\mathrm{H}}(V_i)=\sum_{V_j\in Out(V_i)}HITS_{\mathrm{A}}(V_j) \quad (3)$$

在 Web 搜索引擎的环境下,HITS 算法的目的是当用户给定某个查询,返回给用户高质量的 Authority 页面。Mihalcea 等[14]提出将 HITS 算法应用在文本关键句子抽取以及自动摘要任务上,达到了不错的应用效果。

2 模型设计

2.1 基于词性的加权句向量

句子作为一个完整的语义表示单元,如何进行向量化表示也是自然语言处理领域研究热点。不同于词语,语句含有的信息更丰富,拥有更精确的语义以及语法表示。Riedel 等[15]提出一种简单但是有效的 SIF 模型,SIF 模型将每个句子先表达为所含词语词嵌入的加权平均,然后把句子放在一起找最大主轴,最后从每个句子移除这个最大主轴。Conneau 等[16]提出 InferSent 模型,类比图像领域的 ImageNet,寻找出自然语言处理的“ImageNet”,在很多 NLP 任务中都取得了 state-of-the-arts 的成果。Ruckel 等[17]提出一个新的句子编码方式,基本思想就是给句子向量“求平均”,获得不同维度上的特征的句向量。除此之外,近年也有很多句子向量化的方法,例如 quick thought[18]等。

考虑到辅助定密任务的特殊性,文档定密的依据信息来源于定密细则,分析定密任务以及定密细则特点,本文提出了基于词性的加权句向量。通过改变名词性词向量以及非名词性词向量在句子中的表示权重,在保留语义语法信息的情况下,结合辅助定密任务特性,改变名词性的语义表达信息。在词向量的选择上,使用在百度百科语料库上训练的 glove 词向量,词向量维度为 300 维,可以在词的粒度上保留词语所含有的语义信息。在预训练词向量的基础上,本文基于词性,构建加权句向量 $\boldsymbol{S}$。

$$\boldsymbol{S}=[\alpha\sum\boldsymbol{W}_i+(1-\alpha)\sum\boldsymbol{W}_j]/|\boldsymbol{S}_w| \quad (4)$$

式中:α 是权值;$\boldsymbol{W}_i$ 是句子中的名词的向量化表示;$\boldsymbol{W}_j$ 表示句子中非名词的向量化表示;$|\boldsymbol{S}_w|$ 表示句子中含有词语的个数。通过设置权值 α,改变名词在句子中的表达权重。

2.2 改进的 TextRank 算法

根据 PageRank 算法改进而来的 TextRank 算法[19]是一种用于文本的基于图的排序算法,通过把文本分割成若干组成节点(单词、句子)并建立图模型,根据节点的 TextRank 得分对文本中的节点进行排序,仅利用单篇文档本身的信息即可实现关键词/句提取。与 LDA、HMM 等模型不同,TextRank 不需要事先对多篇文档进行学习训练,使用较为简洁、高效。

TextRank 算法是将文本解析成单词/句子节点。这时,图上的节点之间的关系不仅是简单的指向和被指向关系,是通过一个权重 ω_{ji} 来表示节点之间的链接

强度,因此是一个带权的无向图。此时 $In(V_i)=Out(V_i)$ = 全体词语/句子集合。公式如下所示:

$$WS(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}\frac{\omega_{ji}}{V_k\in Out(w_{jk})}WS(V_j) \tag{5}$$

式中:$WS(V_i)$表示图节点 V_i的 TextRank 得分;d 是一个常数。TextRank 算法用作文本摘要时,文本中的每一个句子 $\boldsymbol{S}_i$被作为图节点 V_i。$\boldsymbol{\omega}_{ji}$计算公式如下:

$$\boldsymbol{\omega}_{ji}=\frac{|\{w_k\mid w_k\in\boldsymbol{S}_i\&\ w_k\in\boldsymbol{S}_j\}|}{\log(|\boldsymbol{S}_i|)+\log(|\boldsymbol{S}_j|)} \tag{6}$$

式中:$\boldsymbol{S}_i$表示文档中第 i 个句子,w_k表示既属于 $\boldsymbol{S}_i$也属于 $\boldsymbol{S}_j$的单词,$|\boldsymbol{S}_i|$表示句子 $\boldsymbol{S}_i$的单词个数。

Pandit 等[20]在研究面向查询的文档摘要时,将查询语句融入图节点中,并且在迭代计算过程中,保持查询语句所在节点的得分始终为 1,以此获得在查询语句影响下的文档语句重要性排名。

本文根据定密任务特点,改进传统的 TextRank 算法中语句相似权值计算方式,在进行 TextRank 迭代时,将定密细则纳入图节点中,获取在定密细则影响下的文档句子排名。采用式(7)代替式(6)的相似度计算公式,相比于式(6),式(7)相似度计算方法更能体现语义层次的异同。具体计算公式如下:

$$\boldsymbol{\omega}_{ji}=q(\boldsymbol{S}_i,\boldsymbol{S}_j)+\cos(\boldsymbol{S}_i,\boldsymbol{S}_j) \tag{7}$$

$$q(\boldsymbol{S}_i,\boldsymbol{S}_j)=\begin{cases}1 & \boldsymbol{S}_i^{w}\cap\boldsymbol{S}_j^{w}\neq 0\\ 0 & \text{其他}\end{cases} \tag{8}$$

式中:$\cos(S_i,S_j)$表示句子 S_i、S_j的余弦相似度;S_i^{w}表示属于句子 $\boldsymbol{S}_i$中的所有单词。最终,计算文档密级分数 D_{score},作为文档密级判断依据,计算公式如下:

$$D_{\text{score}}=\max(WS(\boldsymbol{S}_i)\times\cos(\boldsymbol{S}_i,Z)) \tag{9}$$

式中:Z 表示定密细则;S_i代表待定密文档的第 i 个句子。将 D_{score}值与密级关联度 k 进行比较。大于 k 的文档句子定密为相应定密细则对应密级。这里 k 值是通过实验启发式的选择。

2.3　模型计算过程

模型的具体计算过程如下:

1)首先解析定密事项为细则句,对于一个待定密文档 D,将定密细则纳入文档 D 中,然后进行分词、去停用词等预处理。以句子为单位进行分隔,得到句子列表 $\boldsymbol{S}_i$。

2)构建加权句向量$\boldsymbol{S}_i$,并根据式(7)构建 $\boldsymbol{\omega}_{ji}$矩阵。

3)迭代计算 TextRank 得分,并在迭代过程中,保持定密细则的得分始终为 1。

4)计算文档密级分数 D_{score},并根据与密级关联度 k 的关系判断文档密级。

3　实　验

3.1　实验环境与数据

由于涉密数据难以获取,本文利用已有的公开政务数据,根据公开的地质材料定密细则内容,类比设计定密细则,实验数据来源于中华人民共和国中央人民政府信息公开网站。涉及民族宗教、国防、对外事务、财政金融审计等类别的政务文件。

定密细则的设定,则是根据网上已公开的地质资料定密细则进行类比设计,最终拟定实验用“机密级”以及“绝密级”细则。实验数据集选取了共 73 条来自不同类别的政务文件。首先,人工对实验数据集进行“密级”标注。分为普通级文件、机密级文件和绝密级文件。其中,普通级文件 45 条,机密级文件 18 条,绝密级文件 10 条。

3.2　实验运行环境

实验环境如表 1 所示。

表 1　实验运行环境

处理器	AMD A10 PRO-7800B R7 3.5 GHz
操作系统	Windows 10
运行内存	8 GB
编程语言	Python
工具库	Networkx,Jieba,Numpy

实验首先用 Jieba 分词工具对待定密文档 D 进行分词,删除停用词等无意义词汇。构造句向量的词向量来源于在百度百科语料集上训练的 glove 词向量,词向量大小为 300 维。算法迭代采用 Networkx 库函数,并修改代码,使得定密细则的得分在迭代过程中始终为 1,具体计算过程如图 2 所示。

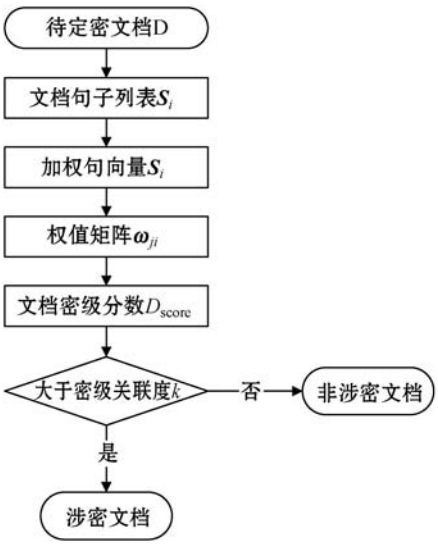

图 2　改进的 TextRank 算法流程

3.3 实验结果与分析

在改进的 TextRank 算法迭代计算中,将式(5)中的 d 设置为 0.85,实验在不同加权句向量权重 α 以及不同密级关联度 k 条件下的效果。通过计算文档定密的准确率来进行评价。具体实验结果如图 3 所示。

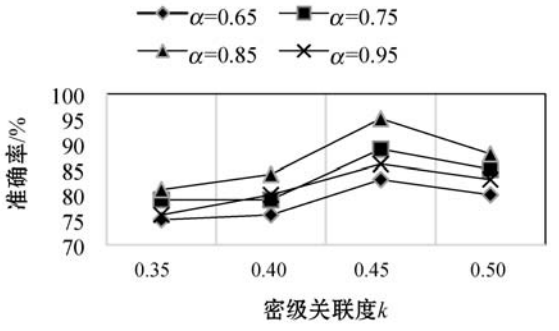

图 3 不同 k 的准确率变化曲线

图 4 是在相同的数据集下,将传统基于关键词的定密方法与改进的 TextRank 算法在准确率、精确率以及召回率上进行对比。

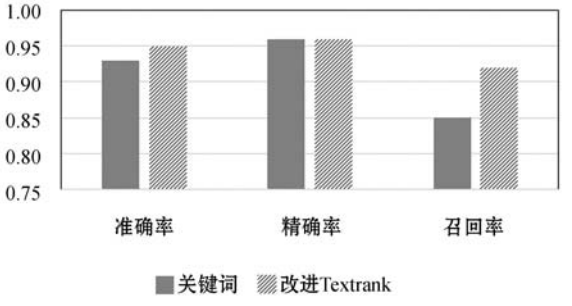

图 4 算法效果对比

由图 3 可以看出,总体上在相同的权值 α 下,准确率随着 k 值的增加而不断增加,在 $k=0.45$ 左右达到最大值。而在相同的 k 下,整体上准确率随着 α 增大而增大。因为随着 α 的增大,句向量的构造中名词词性的词汇权重越来越大。但是在 $\alpha=0.95$ 时,准确率并没有随着 α 增大而继续增大。本文认为当名词词汇权重过于大时,忽视了句子本身的语法信息,导致准确率下降。由图 4 可以看出改进的 TextRank 算法相比于传统基于关键词的辅助定密,在准确率和召回率有一定的提升。

4 结 语

本文提出了一种基于改进的 TextRank 算法的计算机辅助定密方法,该方法结合构造的加权句向量和改进 TextRank 算法,在对文档中的句子权重计算时,将定密细则纳入算法迭代过程中,并在迭代过程中保持定密细则权重始终为 1,从而生成在定密细则影响下的文档句子权重。通过文档句子权重与定密细则相似度判断文档密级,解决了传统基于关键词辅助定密准确率不足,质量差的问题。实验结果表明所提出的改进的 TextRank 方法相对比传统基于关键词的定密方式在准确率上有一定的提升。

近年来,信息安全成为国家安全的重点,涉密电子政务文件的保密定密工作重要性尤为凸显。如何结合定密细则,设计更高效、速度更快的辅助定密模型,是往后工作的研究方向与重点。

参 考 文 献

[1] 高波. 美军定密制度研究[D]. 长沙:国防科学技术大学,2007.

[2] 陈翠玲. 我国定密工作研究[D]. 泉州:华侨大学,2016.

[3] 张帆,卢昱. 基于可信度的不确定推理辅助定密[C]//中国电子学会通信学分会. 2009 全国计算机网络与通信学术会议论文集,2009:212 - 216.

[4] 吴国华,霍晨晨. 一种根据文档相似度快速查找定密依据的方法[J]. 保密科学技术,2014(7):12 - 15.

[5] 潘娅. 一种基于中文文本分类技术的计算机辅助密级界定方法[J]. 电子测试,2016(6):50 - 51.

[6] Hinton G E. Learning distributed representations of concepts [C]//Eighth Conference of the Cognitive Science Society, 1989.

[7] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]//International Conference on Learning Representations, 2013.

[8] Pennington J, Socher R, Manning C D, et al. Glove: Global vectors for word representation[C]//Empirical Methods in Natural Language Processing, 2014: 1532 - 1543.

[9] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137 - 1155.

[10] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//North American Chapter of the Association for Computational Linguistics, 2018: 2227 - 2237.

[11] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB]. arXiv: 1810.04805, 2018.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000 - 6010.

[13] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998,30: 107 - 117.

(下转第 345 页)

管理员多出“图像标注”和“数据上传”按钮,可以实现完成标记并上传病虫害数据的功能。

4 结　语

本系统针对病虫害对象,基于 Android 开发建设专用作物的病虫害识别系统,采用主动采集与网络数据抓取方式,经过智能筛选及异构数据融合等技术手段,实现数据有效获取并保证其真实可靠。实现病虫害识别、病虫害情况上报、病虫害预警等功能,为科研机构、政府、企业、农用品经销商以及消费者提供决策支持和信息服务。受限于样本与技术,针对算法模型的进一步优化升级,提升识别检测准确度,将应用范围进一步扩大是未来工作的研究方向。

参 考 文 献

[1] 李勤斌. 农业病虫害防治现状与方法研究[J]. 种子科技,2019,37(4):118 - 121.

[2] 温芝元,曹乐平. 基于为害状色相多重分形的桔柑病虫害图像识别[J]. 农业机械学报,2014,45(3):262 - 267.

[3] 郭小清,范涛杰,舒欣. 基于改进 Multi-Scale AlexNet 的番茄叶部病害图像识别[J]. 农业工程学报,2019,35(13):162 - 169.

[4] 于洪涛,袁明新,王琪,等. 基于 VGG-F 动态学习模型的苹果病虫害识别[J]. 科学技术与工程,2019,19(32):249 - 253.

[5] 师韵,黄文准,张善文. 基于二维子空间的苹果病害识别方法[J]. 计算机工程与应用,2017,53(22):180 - 184.

[6] 屈赟,陶哺,王政嘉,等. 基于 Android 的苹果叶部病害识别系统设计[J]. 河北农业大学学报,2015,38(6):102 - 106.

[7] 赵建敏,李艳,李琦,等. 基于卷积神经网络的马铃薯叶片病害识别系统[J]. 江苏农业科学,2018,46(24):251 - 255.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB]. arXiv:1409.1556, 2014.

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770 - 778.

[10] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2015.

[11] Redmon J, Farhadi A. Yolov3: An incremental improvement [EB]. arXiv:1804.02767, 2018.

[12] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2016: 779 - 788.

[13] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2017: 7263 - 7271.

[14] 王岩,吴晓富. 深度神经网络训练中适用于小批次的归一化算法[J]. 计算机科学,2019,46(S2):273 - 276.

[15] 林涛,赵璨. 最近邻优化的 k-means 聚类算法[J]. 计算机科学,2019,46(S2):216 - 219.

[16] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017(6):1137 - 1149.

[17] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Springer, 2016: 21 - 37.

[18] Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]//Proceedings of the European Conference on Computer Vision(ECCV),2018: 385 - 400.

(上接第 340 页)

[14] Mihalcea R. Graph-based ranking algorithms for sentence extraction, applied to text summarization[C]//Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 2004.

[15] Riedel B, Augenstein I, Spithourakis G P, et al. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task[EB]. arXiv:1707.03264, 2017.

[16] Conneau A, Kiela D, Schwenk H, et al. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data[EB]. arXiv: 1705.02364, 2017.

[17] Ruckle A, Eger S, Peyrard M, et al. Concatenated p-mean word embeddings as universal cross-lingual sentence representations[EB]. arXiv:1803.01400, 2018.

[18] Logeswaran L, Lee H. An efficient framework for learning sentence representations[C]//International Conference on Learning Representations, 2018.

[19] Mihalcea R, Tarau P. TextRank: Bringing order into text [C]//Empirical Methods in Natural Language Processing, 2004: 404 - 411.

[20] Pandit S R, Potey M A. A query specific graph based approach to multi-document text summarization: simultaneous cluster and sentence ranking[C]//International Conference on Machine Intelligence and Research Advancement,2014.