引用格式:彭曙蓉,郭丽娟,陈慧霞,等.基于特征增强的变电站保护装置录波通道同源匹配研究[J].电力科学与技术学报,2024,39(6):53-59.
Citation: PENG Shurong, GUO Lijuan, CHEN Huixia, et al. Matching of homologous recording channels of substation protection devices based on feature enhancement[J]. Journal of Electric Power Science and Technology, 2024, 39(6): 53-59.

基于特征增强的变电站保护装置录波通道同源匹配研究

彭曙蓉[1],郭丽娟[1],陈慧霞[1],王冠南[2]
(1.长沙理工大学电气与信息工程学院,湖南 长沙 410114;2.国网江西省电力有限公司电力科学研究院,江西 南昌 330096)

摘 要:继电保护装置是采集故障录波数据的重要设备之一。对比分析同源录波数据可以实现故障诊断,但是由于没有规范录波通道的命名格式,导致录波通道同源匹配困难。对此提出一种基于特征增强的保护装置录波通道同源匹配方法。首先,使用同义词替换、文本扩充和正则表达式算法对录波通道名称进行特征增强;然后,使用Pkuseg分词工具对录波通道名称进行分词,使用自定义的停用词表去除录波通道名称中的停用词;最后,使用TF-IDF方法将录波通道名称处理为数字向量形式,并使用余弦相似度算法计算录波通道名称之间的相似度,根据相似度大小判断同源录波通道。算例结果表明,所提方法能充分挖掘录波通道名称的关键信息并提高录波通道同源匹配的准确率。

关 键 词:保护装置录波同源匹配;文本匹配;特征增强;余弦相似度
DOI:10.19781/j.issn.1673-9140.2024.06.006 **中图分类号**:TM631 **文章编号**:1673-9140(2024)06-0053-07

Matching of homologous recording channels of substation protection devices based on feature enhancement

PENG Shurong[1], GUO Lijuan[1], CHEN Huixia[1], WANG Guannan[2]
(1.School of Electrical & Information Engineering, Changsha University of Science & Technology, Changsha 410114, China;
2.Electric Power Research Institute, State Grid Jiangxi Electric Power Co., Ltd., Nanchang 330096, China)

Abstract: Relay protection device is one of the important equipment for collecting fault recording data. Comparison and analysis of homologous recording data can realize fault diagnosis, but due to the lack of standardization for the naming format of the recording channel, it is difficult to match the homologous recording channels. In this paper, a matching method for homologous recording channels of protection devices based on feature enhancement is proposed. Firstly, the recording channel names are feature-enhanced by means of synonym replacement, text expansion, and regular expression algorithms. Then, the Pkuseg word segmentation tool is used to segment the recording channel names, and a customized stop word list is utilized to remove the stop words in the recording channel names. Finally, the term frequency-inverse document frequency (TF-IDF) method is used to process the recording channel names into the form of digital vectors, and the cosine similarity algorithm is used to calculate the similarity between the recording channel names. The homologous recording channels are judged according to their similarity. The example results show that the proposed method can fully explore the key information of the recording channel names and improve the accuracy of the matching of homologous recording channels.

Key words: matching of homologous recording channel of protection device; text matching; feature enhancement; cosine similarity

收稿日期:2023-03-18;**修回日期**:2023-12-11
基金项目:国网江西省电力有限公司科技项目(52182022000A);湖南省教育厅重点项目(20A021);国家自然科学基金(52177069)
通信作者:王冠南(1987—),男,硕士,高级工程师,主要从事继电保护技术及管理研究;E-mail:dell_wgn2006@163.com

电力系统发生故障时，继电保护装置和故障录波装置均会记录各电气量的变化情况，工作人员可基于这些录波文件进行故障诊断[1]。故障录波装置是专用的记录故障波形的装置，广泛应用于重要的变电站内，而继电保护装置则广泛配置于几乎所有的电气设备中，覆盖范围更广，因此，研究继电保护装置采集的波形是进行故障诊断的一个重要手段[2-5]。为维护电力系统安全稳定运行，在220 kV及以上变电站中，继电保护采用双重化配置，双套保护分别采集各自的录波数据，并将数据送至保信主站。对比同源双套录波的数据是进行故障分析的重要手段，为此，首先要进行同源通道匹配，由于没有相关的技术规范，不同的厂商有不同的录波通道命名习惯。目前，针对录波通道同源匹配技术的研究较少，实际工程中采用人工匹配搭配编辑距离算法的方式。但是，人工匹配工作量大、易出错，编辑距离算法准确性不高、自适应能力不强[6]，无法大范围推广使用。

文本匹配是自然语言理解的核心问题，涉及信息检索、对话系统等多个领域[7-8]。传统的文本匹配主要从统计的角度出发，通过计算文本之间共有词的比例来判断相似文本，例如Jaccard相似系数算法[9-10]、Simhash相似度算法[11-12]、编辑距离(Levenshtein distance)[13-14]。这些方法结构简单，匹配速度快，但是缺乏对文本语义信息的挖掘，匹配效果不理想。随着深度学习的发展，文本匹配逐渐引入神经网络等算法，挖掘文本语义信息，提高匹配准确率。文献[15]提出了一种结合语义和中心词注意力机制的短文本相似度计算方法，利用双向门控递归单元(bidirectional gated recurrent unit，BiGRU)提取上下文信息，将短文本中反映主要信息的名词、动词作为中心词，更好地提取短文本的关键内容，把握语义信息。文献[16]结合双向长短期记忆神经网络、卷积神经网络和密集连接网络，充分挖掘文本的语义信息，提高农业问答系统的工作效率。文献[17]提出了一种结合加权微调基于变压器的双向编码器表示(bidirectional encoder representation from transformers，BERT)特征提取和孪生双向长短期记忆神经网络的文本相似度计算方法。虽然这些方法很大程度地提高了中文文本语义识别的准确率，但是模型的结构复杂、参数较多，匹配速度慢。

录波通道名称中包含有关键词信息，但是由于没有统一的命名格式，导致同源通道的命名不一致，增大了同源匹配的难度。录波通道命名时随机使用字母缩写、简写等形式，导致同源匹配时信息把握不准确、不完整，匹配准确率降低。

为了提高录波文件中同源通道匹配准确率，本文提出一种基于特征增强的保护装置录波通道同源匹配算法。首先，从配置文件中提取待匹配的录波通道名称；然后，通过同义词替换、文本扩充、正则表达式等算法对录波通道名称数据进行特征增强；接着，使用分词工具和去停用词算法对录波通道名称进行预处理；最后，使用余弦相似度算法计算录波通道名称之间的相似度，进而实现录波通道同源匹配。该方法充分考虑录波通道名称的特征，有针对性地解决同源录波通道匹配过程中的难点，提高同源录波通道匹配准确率，同时对多个不同变电站均有较好的匹配效果，保证算法的自适应能力。

1 同源通道匹配方法整体架构

基于特征增强的保护装置录波通道同源匹配算法的整体流程如图1所示。

特征增强主要解决录波通道命名不规范的问题。不同的继电保护设备厂家对录波通道命名的格式不同，因此在双套保护的录波数据间开展同源匹配存在一定困难。例如，有的厂家习惯使用字母及数字来表示各录波通道，有的厂家则会使用文字的形式来表示，还有的厂家会使用简写的形式。对此，本文使用特征增强算法使各录波通道的关键信息均能全面、相对一致地展示出来。

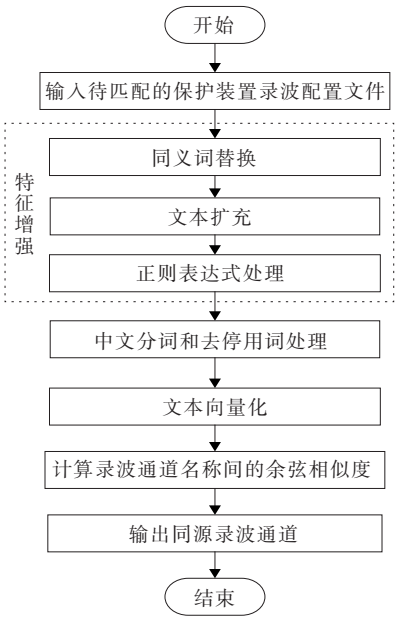

图1 录波通道同源匹配算法流程

Figure 1 Flow chart of matching algorithm for homologous recording channels

文本预处理则是针对录波通道名称中的停用词进行处理。停用词在录波通道名称中表现为无实际含义的字、词和符号,这些词汇会干扰相似度计算结果。经过预处理的录波通道名称将不再包含这些停用词,有助于提高相似度的计算精度。同时,文本预处理还将对录波通道名称进行分词,便于后续的文本向量化表示。

2　录波通道名称特征增强

2.1　录波通道名称数据分析

每一个保护装置录波文件中均包含有头文件(.HDR)、配置文件(.CFG)、数据文件(.DAT)和信息文件(.INF)4个相关联的文件[18]。配置文件中包含厂站名、通道总数和类型、通道名称等基本配置信息。配置文本分为数行,每一行以回车或换行结束,一行中的各个域由逗号分隔开,即使某个域中没有数据,也需要使用逗号与其他域区分。录波通道名称数据固定位于配置文件的第二列,可使用Python编程将其单独提取出来。

通道名称由各厂家设定,没有统一的标准,因此通道名称表达形式不统一,甚至有个别名称错误,给录波通道同源匹配带来挑战。保护装置录波文件中的各通道名称均包含电压电流、相别、变压器高、中、低压侧等关键信息。表1列举了部分同源录波通道。

由表1的示例可知,保护装置同源录波通道在命名时有多种情况。第一种情况是A、B双套录波通道的命名形式完全一致,这种情况在匹配时,即使不进行任何预处理,也可以准确匹配;第二种情况是A、B双套录波通道在命名时均包含了关键信息,但是关键信息的表达形式可能不一致,这种情况在匹配时则需要对录波通道名称进行适当处理,提高相似度计算的准确率。

表1　同源录波通道示例

Table 1　Examples of homologous recording channels

A套录波文件中的通道名称	B套录波文件中的同源通道名称
Ⅰ母UA(U1A)	Ⅰ母母线A相电压Ua11
Ⅱ母UA(U2A)	Ⅱ母母线A相电压Ua21
Ia	Ia
高压1侧Ih1a	高压1侧A相电流Ih1a1
高压间隙Ihj	高压侧间隙电流Ihj1

2.2　同义词替换和文本扩充

数据增强是一种针对小样本学习问题提出的训练样本强化方法[19],其在图像识别、语音修复等领域的应用较为广泛,但其在自然语言处理问题中仍有较大的发展空间。将数据增强应用于中文语义识别问题的关键点是如何保持文本语义的一致性,如果增强后的文本与原始文本语义不一致,将得不偿失。本文考虑使用同义词替换[20]和文本扩充[21]对录波通道名称进行数据增强,在保持关键信息不变的前提下,对录波通道名称进行处理,使其信息能完整、充分地体现出来。简单数据增强(easy data augmentation,EDA)是目前常用的数据增强方法,包括同义词替换、随机插入、随机交换和随机删除[22-23]。

规定录波通道名称的标准格式,对照标准格式,使用同义词替换和文本扩充算法对待匹配的录波通道名称进行处理,使得各录波通道名称完整、清晰地表达关键信息。为了验证同义词替换和文本扩充对录波通道名称信息的增强效果,本文从保护装置录波文件中提取录波通道名称进行实验。部分录波通道增强效果如表2所示。

表2　数据增强效果对比

Table 2　Comparison of data enhancement effects

原始通道名称	数据增强后的通道名称
Ⅰ母UA(U1A)	Ⅰ母A相电压UA(U1A)
Ⅱ母UA(U2A)	Ⅱ母A相电压UA(U2A)
高压1侧Ⅰh1a	高压1侧A相电流Ⅰh1a
高压间隙Ⅰhj	高压侧间隙电流Ⅰhj
Ⅰa	A相电流Ⅰa

2.3　正则表达式处理

正则表达式由一串特定意义的字符组成,能够在待匹配的文本中查找到希望被匹配的文本对象。正则表达式的字符分为普通字符和元字符2种,其中元字符主要用于对字符串模板进行设置[24]。常见的元字符如表3所示。

表3　常见正则表达式元字符

Table 3　Common regular expression metacharacters

元字符	作用
.(点)	通配符,匹配任何字符
d	匹配0~9任何数字
w	匹配非数字非字母的任何字符
s	等价[.],即匹配任何一个字符
$E\{n\}$	表达式E匹配n次
$E\{n,m\}$	表达式E匹配n次或m次
$E\{n-m\}$	表达式E匹配n次至m次
^	匹配字符串开始的位置
$	匹配字符串结束的位置

正则表达式能够实现数据的匹配、替换、删除和提取功能[25]。将设定的正则表达式与待匹配的文本进行比较，即可实现匹配功能，还可以根据匹配的结果进行进一步删除、替换和提取工作，如图2所示。

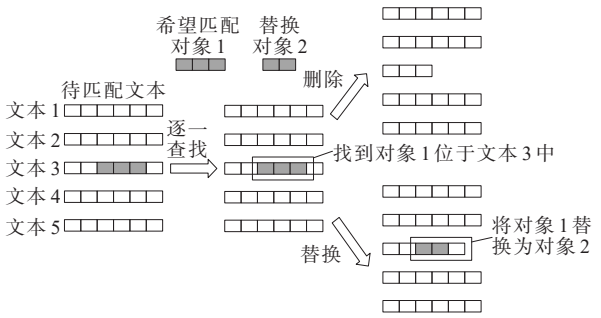

图2 正则表达式处理示意图

Figure 2 Regular expression processing

对于录波通道名称中的一些表达形式不统一的情况，例如大小写不一致、罗马数字和字母使用不统一等，可编写对应的正则表达式，对录波通道名称进行匹配、替换，以统一表达形式，提高相似度计算的准确率。

2.4 录波通道名称向量化处理

计算文本相似度的前提是将文本转化为向量形式。当前，文本向量化都是在词语的基础上开展的，因此，首先要对录波通道名称文本进行分词。目前，常用的中文分词算法主要有Pkuseg分词和Jieba分词2种，这2种算法都能实现准确地分词。录波通道名称中包含了大量的专有名词以及地方特色词汇，难以整理出完整的词库用于分词。Jieba分词支持自定义的分词词库，且过程简单、速度快，但是由于录波通道分词词典难以完整地整理，导致Jieba分词效果不佳。因此，本文考虑使用Pkuseg分词算法。文献[26]提出使用已正确分词的录波通道名称，随机生成训练集、测试集和验证集，训练得到录波通道名称专用的分词模型。

词频-逆文档频率(term frequency-inverse document frequency，TF-IDF)是常用的基于统计的文本向量化方法，被用来统计文本中词语出现的频率，越重要的词语在特定文本中出现的频率越高，而在其他文本中出现的频率越低。词频统计(term frequency，TF)是词语在特定文本中出现的频率，逆文本词频统计(inverse document frequency，IDF)是词语在其他文本中出现的频率。词语w_i在文档d_j中出现了$n_{i,j}$次，则词频统计$T_{\mathrm{TF},i,j}$和逆文本词频统计$T_{\mathrm{IDF},i}$计算如下：

$$T_{\mathrm{TF},i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$T_{\mathrm{IDF},i}=\log\frac{|D|}{1+\left|j:w_i\in d_j\right|} \tag{2}$$

式中，$|D|$为所有文档的数量；$\left|j:w_i\in d_j\right|$表示包含词语$w_i$的文档数量。

TF-IDF计算如下：

$$T_{\mathrm{TF\text{-}IDF},i}=T_{\mathrm{TF},i,j}\cdot T_{\mathrm{IDF},i} \tag{3}$$

本文将分词后的录波通道名称使用TF-IDF方法转化为数字向量形式，用于后续的相似度计算。

3 算例分析

为了验证基于特征增强的录波通道同源匹配方法的有效性，本文基于Python语言建立保护装置录波通道同源匹配模型，使用不同变电站的不同保护装置记录的录波通道名称进行验证。

3.1 录波通道名称特征增强结果

从配置文件的第二列中提取录波通道名称，去除配置文件中其他信息，便于后续匹配同源录波通道。由于录波通道命名没有规范格式，各同源录波通道表达形式存在不统一的现象，同时录波通道名称中含有停用词，影响录波通道同源匹配。另外，录波通道命名时存在使用简写、字母缩写等情况，也会降低同源通道间的相似度。因此，本文在进行同源匹配之前，先对录波通道名称进行特征增强。

首先，给定录波通道命名模板，加载通用词典，对录波通道名称进行同义词替换以及文本扩充。同义词替换可以统一录波通道中同义的表达形式，文本扩充可以处理通道名称中简写和英文缩写的情况。然后，归纳录波通道名称中罗马数字和字母混用、字母大小写不一致等情况，编写相应的正则表达式，使用正则表达式的匹配、替换功能，统一这些表达形式。经过特征增强后的录波通道名称的特征更加显著，关键信息表述更加完整、准确。

特征增强处理后的效果对比如表4所示。从表4中可以看出，经过同义词替换、文本扩充和正则表达式处理，录波通道名称发生以下变化：1)同义的表达形式被统一，如“高”“低1分支”统一表达为“高压侧”“低压1分支”；2)针对输入法不同而出现的字母、罗马数字混用情况，将字母统一表达为罗马数字形式；3)对于一些仅使用了字母缩写的录波通道名称，补充其关键信息，例如对“UA”通道补充文字表述“A相电压”。

表 4　特征增强效果对比

Table 4　Comparison of feature enhancement effects

原始录波通道名称	特征增强后的录波通道名称
高ⅠA	高压侧A相电流Ⅰa
高压1侧Ⅰh1a	高压1侧A相电流Ⅰh1a
低1分支C相电压Ul1c1	低压1分支C相电压Ul1c
Ⅰ母UA(U1A)	Ⅰ母A相电压Ua
Ⅱ母UA(U2A)	Ⅱ母A相电压Ua

3.2　录波通道名称预处理

经过特征增强,录波通道名称已经具有统一的表达形式,关键信息表述清晰,但是其中还可能包含了一些停用词,例如无实际含义的符号、字词等。另外,文本的向量化表示也需要在词语的维度上进行,因此需要对录波通道名称进行分词和去停用词操作。

本文使用文献[26]提出的方法,建立录波通道专用分词模型,结合Pkuseg分词工具进行分词,在分词结果的基础上,加载自定义的停用词表去除保护装置录波通道名称中的停用词。表5展示了Pkuseg分词效果。

表 5　Pkuseg分词效果

Table 5　Pkuseg word segmentation effects

原始录波通道名称	Pkuseg分词结果
低压1分支C相电压Ul1c	低压/1/分支/C相/电压/Ul1c
Ⅰ母A相电压Ua	Ⅰ母/A相/电压/Ua
高压侧A相电流Ⅰa	高压侧/A相/电流/Ⅰa
保护电流A相Ⅰa	保护/电流/A相/Ⅰa
通道1对侧A相电流	通道1/对侧/A相/电流
通道2A相差动电流	通道2/A相/差动/电流
对侧A相电流	对侧/A相/电流

3.3　录波通道同源匹配

按照本文所提方法,采用余弦相似系数计算经过预处理的保护装置录波通道名称文本数据间的相似度,依据各自相似度值的大小判断同源录波通道。算例分析采用Python进行,电脑配置为i7-7500U,2.7 GHz,12 GB。

A套录波中的各录波通道与B套录波中的各录波通道之间的相似度大小使用余弦相似度算法进行计算,相似度最大的组合被选择作为同源录波通道输出。为了验证余弦相似系数匹配的效果,本文引入Jaccard相似度、Simhash相似度和编辑距离这几种匹配算法,分析它们的匹配结果。

为了量化模型的匹配效果,引入以下评价指标。同源通道正确匹配的对数为T_P,非同源通道被错误匹配为同源通道的对数为F_P。模型匹配精确率(precision ratio,PR)[26-28]定义如下:

$$T_{PR}=\frac{T_P}{T_P+F_P}\times 100\% \tag{4}$$

实验得到的几种匹配算法的评价指标比较如表6所示。由表6可知,针对保护装置录波通道同源匹配问题,这几种算法中,余弦相似度和特征增强相结合的方法匹配效果最好,能达到96.8%的匹配精确率。Jaccard相似度算法的核心是计算共有词数在文本总词数中的占比。Simhash相似度能很好地反映文本局部的差异,对于局部不同的相似文本,可以得到相似的散列值。编辑距离是计算由一个文本得到另一个文本所需的最少编辑次数。这3种算法的原理及匹配过程都比较简单,但是都没有关注文本的语义信息,而且由于保护装置录波通道名称没有规范的命名格式,因此很容易影响算法的相似度计算,导致匹配效果不佳。而余弦相似度算法建立在文本向量化的基础上,匹配精确率的提高反映了使用TF-IDF向量化算法的效果较好。余弦相似度和特征增强算法相结合,充分挖掘了录波通道名称的关键词信息,提高了同源匹配的精确率,而且匹配所需要的时间也较短。

表 6　算法评价指标比较

Table 6　Comparison of evaluation indexes of algorithms

匹配算法	T_{PR}/%	匹配用时/s
Jaccard相似度	94.5	0.18
Simhash相似度	87.2	0.43
编辑距离	89.5	0.32
余弦相似度	95.2	0.15
余弦相似度+特征增强	96.8	0.33

为了验证本文所提方法的自适应能力,本文获取多个变电站的保护装置录波文件进行算例分析。图3展示了63所变电站保护装置录波通道的同源匹配情况。可以发现,余弦相似度和特征增强相结合的方法匹配效果最好,曲线较其他2种算法的精确率更高,且有更多的点落在100%。这说明,余弦相似度和特征增强相结合的方法的匹配精确率较其他2种算法的更高,且在更多的变电站内达到了100%的匹配精确率,证明了其自适应能力更强。

综合上述算例分析结果可知,基于特征增强的保护装置录波通道同源匹配算法不仅精确率更高,而且自适应能力更强,能够在多个变电站内都有较好的匹配效果。

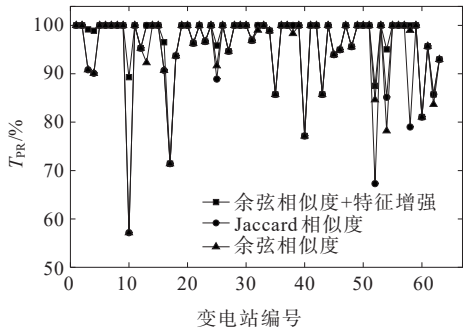

图3 不同变电站内同源匹配结果对比

Figure 3 Comparison of homologous matching results in different substations

4 结语

本文针对变电站保护装置录波通道同源匹配问题,分析了录波通道名称的特点。录波通道名称中包含了关键词信息,但是命名格式不统一,且包含停用词。根据录波通道名称的特征,本文提出了基于特征增强的保护装置录波通道同源匹配方法。该方法使用同义词替换、文本扩充和正则表达式算法对录波通道名称进行特征增强;使用Pkuseg分词工具和去停用词操作对录波通道名称进行预处理;使用TF-IDF算法将录波通道名称向量化;使用余弦相似度算法计算录波通道名称之间的相似度。将本文所提方法与其他方法进行比较发现,本文所提方法具有更高的匹配精确率,且自适应能力更强。

参考文献:

[1] 叶远波,刘宏君,张兆云,等.基于广域信息的继电保护实时评价研究[J].电力系统保护与控制,2021,49(13):150-157.
YE Yuanbo, LIU Hongjun, ZHANG Zhaoyun, et al. Research on real-time evaluation of relay protection based on wide area information[J]. Power System Protection and Control,2021,49(13):150-157.

[2] 雷明,陈一椋,刘峰,等.D5000继电保护设备在线监视及分析应用提升[J].电网技术,2020,44(3):1197-1202.
LEI Ming, CHEN Yicong, LIU Feng, et al. Online monitoring and analysis system based on D5000 for relay protections[J]. Power System Technology, 2020, 44(3):1197-1202.

[3] 严敬汝,臧谦,赵宇皓,等.基于配网录波特征库的故障识别与保护定值整定及实现[J].电力科学与技术学报,2023,38(2):248-254.
YAN Jingru, ZANG Qian, ZHAO Yuhao, et al. Implementation of distribution network fault identification and protection setting based on characteristic recording data map[J]. Journal of Electric Power Science and Technology,2023,38(2):248-254.

[4] 来智浩,高钰琛,翟常营,等.基于信号路径跟踪算法与并发分层架构的热过载保护装置自动校验方法研究[J].高压电器,2022,58(11):121-127.
LAI Zhihao, GAO Yuchen, ZHAI Changying, et al. Research on automatic verification method of thermal overload protection device based on signal path tracking algorithm and concurrent hierarchical architecture[J]. High Voltage Apparatus,2022,58(11):121-127.

[5] 王业,崔玉,陆兆沿,等.基于CNN图像识别算法的保护装置智能巡视技术[J].电力工程技术,2022,41(6):252-257.
WANG Ye, CUI Yu, LU Zhaoyan, et al. Intelligent inspection technology of protection device based on convolution neural network image recognition algorithm [J]. Electric Power Engineering Technology, 2022, 41(6): 252-257.

[6] 简开宇,史涯晴,黄松,等.业务流程模型相似度研究综述[J].计算机科学,2023,50(6):338-350.
JIAN Kaiyu,SHI Yaqing,HUANG Song,et al.Review on similarity of business process models[J]. Computer Science,2023,50(6):338-350.

[7] 庞亮,兰艳艳,徐君,等.深度文本匹配综述[J].计算机学报,2017,40(4):985-1003.
PANG Liang, LAN Yanyan, XU Jun, et al. A survey on deep text matching[J]. Chinese Journal of Computers, 2017,40(4):985-1003.

[8] 宋丹,陆奎,戴旭凡.基于改进的卷积神经网络邮件分类算法研究[J].重庆工商大学学报(自然科学版),2022,39(3): 20-25.
SONG Dan, LU Kui, DAI Xufan. Research on mail classification algorithm based on improved convolutional neural network[J]. Journal of Chongqing Technology and Business University(Natural Science Edition),2022,39(3):20-25.

[9] 田星,郑瑾,张祖平.基于词向量的Jaccard相似度算法[J].计算机科学,2018,45(7):186-189.
TIAN Xing, ZHENG Jin, ZHANG Zuping. Jaccard text similarity algorithm based on word embedding[J]. Computer Science,2018,45(7):186-189.

[10] SRINIVASARAO U, KARTHIKEYAN R, SARANGI P K, et al. Enhanced movie recommendation and sentiment analysis model achieved by similarity method through cosine and Jaccard similarity algorithms[C]//2022 International Conference on Computing,Communication, and Intelligent Systems (ICCCIS). Greater Noida, India. IEEE,2022:214-218.

[11] 张航,盛志伟,张仕斌,等.Simhash算法在文本去重中的应用[J].计算机工程与应用,2020,56(11):246-251.
ZHANG Hang, SHENG Zhiwei, ZHANG Shibin, et al. Application of Simhash algorithm in text deduplication [J].Computer Engineering and Applications,2020,56(11): 246-251.

[12] 李玫,高庆,马森,等.面向代码相似性检测的相似哈希改进方法[J].软件学报,2021,32(7):2242-2259.
LI Mei, GAO Qing, MA Sen, et al. Enhanced Simhash algorithm for code similarity detection[J]. Journal of Software,2021,32(7):2242-2259.

[13] 赵志靖,江荻.基于编辑距离的语言分类研究[J].语言研究,2020,40(2):43-50.
ZHAO Zhijing,JIANG Di.Language classification study based on levenshtein distance[J]. Studies in Language and Linguistics,2020,40(2):43-50.

[14] 熊安萍,詹妮,邹毅,等.大数据环境下一种基于模式匹配的实体统一方法[J].计算机应用与软件,2018,35(8):87-92+97.
XIONG Anping, ZHAN Ni, ZOU Yi, et al. A method of entity resolution based on pattern matching in big data environment[J]. Computer Applications and Software, 2018,35(8):87-92+97.

[15] JI M Y,ZHANG X H.A short text similarity calculation method combining semantic and headword attention mechanism[J]. Scientific Programming, 2022, 2022(1): 8252492.

[16] 金宁,赵春江,吴华瑞,等.基于多语义特征的农业短文本匹配技术[J].农业机械学报,2022,53(5):325-331.
JIN Ning, ZHAO Chunjiang, WU Huarui, et al. Agricultural short text matching technology based on multi-semantic features[J]. Transactions of the Chinese Society for Agricultural Machinery,2022,53(5):325-331.

[17] VIJI D,REVATHY S.A hybrid approach of weighted fine-tuned BERT extraction with deep Siamese Bi-LSTM model for semantic text similarity identification[J]. Multimedia Tools and Applications,2022,81(5):6131-6157.

[18] 陈泗贞,梁竞雷,卢迪勇,等.基于COMTRADE模型的电力系统多源故障数据融合分析方法[J].电力科学与技术学报,2019,34(3):92-100.
CHEN Sizhen, LIANG Jinglei, LU Diyong, et al. Multi-source fault data comprehensive analysis method for power system based on COMTRADE model[J].Journal of Electric Power Science and Technology,2019,34(3):92-100.

[19] 李牧南,王良,赖华鹏.中文科技政策文本分类:增强的TextCNN视角[J].科技管理研究,2023,43(2):160-166.
LI Munan, WANG Liang, LAI Huapeng. Text classification of Chinese S&T policies: enhanced TextCNN perspective[J]. Science and Technology Management Research,2023,43(2):160-166.

[20] 尤丛丛,高盛祥,余正涛,等.基于同义词数据增强的汉越神经机器翻译方法[J].计算机工程与科学,2021,43(8):1497-1502.
YOU Congcong,GAO Shengxiang,YU Zhengtao,et al.A Chinese-Vietnamese neural machine translation method based on synonym data augmentation[J]. Computer Engineering & Science,2021,43(8):1497-1502.

[21] 吕晓锋,赵书良,高恒达,等.基于异质信息网的短文本特征扩充方法[J].计算机科学,2022,49(9):92-100.
LYU Xiaofeng, ZHAO Shuliang, GAO Hengda, et al. Short texts feautre enrichment method based on heterogeneous information network[J]. Computer Science,2022,49(9):92-100.

[22] WEI J,ZOU K.EDA:easy data augmentation techniques for boosting performance on text classification tasks[EB/OL].2019:1901.11196.https://arxiv.org/abs/1901.11196v2

[23] 杨鹤.面向渔业标准知识图谱构建的实体识别与关系抽取[D].大连:大连海洋大学,2022.
YANG He.Name entity recognition and relation extraction for fishery standard knowledge graph construction[D]. Dalian:Dalian Ocean University,2022.

[24] 刘洋,赵庆志,王宏甲,等.基于正则表达式的译码方法研究[J].制造业自动化,2022,44(8):48-50.
LIU Yang, ZHAO Qingzhi, WANG Hongjia, et al. Research on decoding method based on regular expression[J].Manufacturing Automation,2022,44(8):48-50.

[25] 胡军伟,秦奕青,张伟.正则表达式在Web信息抽取中的应用[J].北京信息科技大学学报(自然科学版),2011,26(6):86-89.
HU Junwei, QIN Yiqing, ZHANG Wei. Regular expression and its applications to Web information extraction[J]. Journal of Beijing Information Science & Technology University,2011,26(6):86-89.

[26] 戴志辉,杨鑫,刘悦,等.基于增量学习优化的故障录波文件通道名称识别方法[J].电力系统保护与控制,2023,51(4):148-156.
DAI Zhihui, YANG Xin, LIU Yue, et al. Recognition method of fault recorder file channel name based on incremental learning optimization[J]. Power System Protection and Control,2023,51(4):148-156.

[27] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学报,2006,17(9):1848-1859.
SU Jinshu, ZHANG Bofeng, XU Xin. Advances in machine learning based text categorization[J].Journal of Software,2006,17(9):1848-1859.

[28] 丘浩,张炜,彭博雅,等.基于YOLOv3的特定电力作业场景下的违规操作识别算法[J].电力科学与技术学报,2021,36(3):195-202.
QIU Hao, ZHANG Wei, PENG Boya, et al. Illegal operation recognition algorithm based on YOLOv3 in specific power operation scenario[J]. Journal of Electric Power Science and Technology,2021,36(3):195-202.

[29] 吴海涛,代尚林,乔中伟,等.基于RBF-SVM智能配变终端的网络安全态势评估[J].电力科学与技术学报,2021,36(5):35-40.
WU Haitao, DAI Shanglin, QIAO Zhongwei, et al. Research on network security situation awareness of intelligent distribution transformer terminal unit based on RBF-SVM[J].Journal of Electric Power Science and Technology,2021,36(5):35-40.