(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119358564 A
(43)申请公布日 2025.01.24

(21)申请号 202411460565.8

(22)申请日 2024.10.18

(71)申请人 广东电网有限责任公司
地址 510000 广东省广州市越秀区东风东路757号
申请人 广东电网有限责任公司电力调度控制中心

(72)发明人 卢建刚 邓晓智 吴勤勤 杨云帆 潘垚鑫 古振威 杨晨威 李亚南 马腾腾 汤怿 张玉兵

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 邓健明

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/216(2020.01)
G06F 40/284(2020.01)
G06F 18/213(2023.01)
G06F 18/22(2023.01)
G06F 18/23(2023.01)
G06F 18/25(2023.01)

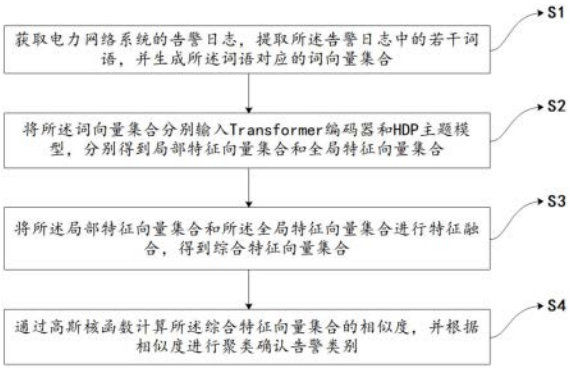
权利要求书2页 说明书10页 附图3页

(54)发明名称

一种告警信息分析方法、系统、设备及存储介质

(57)摘要

本发明公开了一种告警信息分析方法、系统、设备及存储介质,包括;获取电力网络系统的告警日志,提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合;将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合;将所述局部特征向量集合和所述全局特征向量集合进行特征融合,得到综合特征向量集合;通过高斯核函数计算所述综合特征向量集合的相似度,并根据相似度进行聚类确认告警类别。本申请可以提高电网攻击类型的识别准确性和全面性,以准确判断电网可能遭受的攻击类别。



CN 119358564 A

1.一种告警信息分析方法,其特征在于,包括:

获取电力网络系统的告警日志,提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合;

将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合;

将所述局部特征向量集合和所述全局特征向量集合进行特征融合,得到综合特征向量集合;

通过高斯核函数计算所述综合特征向量集合的相似度,并根据相似度进行聚类确认告警类别。

2.根据权利要求1所述的告警信息分析方法,其特征在于,所述获取电力网络系统的告警日志,具体为:

获取电力网络系统的日志记录;

对所述日志记录进行格式校正,并去除重复项,得到清洗后的日志记录;

通过设定告警关键词,保留含有所述告警关键词的日志记录,得到告警日志。

3.根据权利要求1所述的告警信息分析方法,其特征在于,所述提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合,具体为:

通过NLTK对所述告警日志中的若干词语进行分词处理和去除停用词处理,得到词语集合;

将所述词语集合输入Word2vec模型,得到所述告警日志中所述词语对应的词向量集合。

4.根据权利要求1所述的告警信息分析方法,其特征在于,所述得到局部特征向量集合,具体为:

通过向所述词向量集合中的向量填充信息,统一所述词向量集合中的向量的长度;

对所述告警日志中若干词语进行位置编码,得到所述告警日志中词语对应的位置向量;

将统一长度后的词向量集合与词语对应的所述位置向量进行融合,得到输入词向量集合;

将所述输入词向量集合输入Transformer编码器,得到注意力值集合;

将所述注意力值集合和所述输入词向量集合进行残差连接,并对残差连接结果进行层归一化处理,得到所述告警日志中局部特征向量集合。

5.根据权利要求4所述的告警信息分析方法,其特征在于,所述将所述输入词向量集合输入Transformer编码器,得到注意力值集合,具体为:

基于所述输入词向量集合,得到Query向量集合、Key向量集合和Value向量集合;

通过对所述Query向量集合和所述Key向量集合中的向量进行点积计算,得到所述告警日志中词语的注意力分数;

对所述注意力分数进行放缩,并通过softmax函数进行归一化处理,得到所述告警日志中词语的注意力权重;

基于所述注意力权重与所述Value向量集合,得到所述告警日志中词语的注意力值集合。

6.根据权利要求1所述的告警信息分析方法,其特征在于,所述得到全局特征向量集合,具体为:

计算所述告警日志中的语义特征权重集合;

基于所述语义特征权重集合提取所述告警日志中的关键词集合;

将所述关键词集合输入HDP主题模型,得到所述告警日志中的全局特征向量集合。

7.根据权利要求6所述的告警信息分析方法,其特征在于,所述计算所述告警日志中的语义特征权重集合,具体为:

基于所述词向量集合中若干向量之间的余弦相似度,得到语义相似性权重集合;

统计所述告警日志中词语共同出现的次数,得到词共现权重集合;

基于TF-IDF计算所述告警日志中词语的词频权重集合;

基于所述语义相似性权重集合、所述词共现权重集合和所述词频权重集合,得到所述告警日志中的语义特征权重集合。

8.一种告警信息分析系统,其特征在于,包括:获取模块、特征得到模块、融合模块和类别确认模块;

所述获取模块,用于获取电力网络系统的告警日志,提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合;

所述特征得到模块,用于将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合;

所述融合模块,用于将所述局部特征向量集合和所述全局特征向量集合进行特征融合,得到综合特征向量集合;

所述类别确认模块,用于通过高斯核函数计算所述综合特征向量集合的相似度,并根据相似度进行聚类确认告警类别。

9.一种终端设备,其特征在于,包括处理器、存储器以及存储在所述存储器中且被配置为由所述处理器执行的计算机程序,所述处理器执行所述计算机程序时实现如权利要求1至7任意一项所述的告警信息分析方法。

10.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质包括存储的计算机程序,其中,在所述计算机程序运行时控制所述计算机可读存储介质所在设备执行如权利要求1至7任一项所述的告警信息分析方法。

一种告警信息分析方法、系统、设备及存储介质

技术领域

[0001] 本发明涉及信息技术领域,尤其涉及一种告警信息分析方法、系统、设备及存储介质。

背景技术

[0002] 当前电力系统与信息系统高度耦合,因此电力系统往往会面临网络攻击的风险,且随着电力系统信息化的不断发展,电力网络系统产生访问记录呈现爆发式增长,其中存在部分异常的访问记录即告警信息。

[0003] 在现有技术中,由于告警信息不存在明显的标注,往往需要借助人工进行筛选和判断,故存在异常诊断、诊断效率低下、存在遗漏和错看等问题。故如何从海量信息中识别出告警信息,通过分析进而判断出电力网络系统可能遭受的攻击,从而提高电力网络系统的维护效率是一个关键问题。

发明内容

[0004] 本申请提供了一种告警信息分析方法、系统、设备及存储介质,可以提高电网攻击类型的识别准确性和全面性,以准确判断电网可能遭受的攻击类别。

[0005] 第一方面,本申请提供了一种告警信息分析方法,包括:

[0006] 获取电力网络系统的告警日志,提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合;

[0007] 将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合;

[0008] 将所述局部特征向量集合和所述全局特征向量集合进行特征融合,得到综合特征向量集合;

[0009] 通过高斯核函数计算所述综合特征向量集合的相似度,并根据相似度进行聚类确认告警类别。

[0010] 本申请实施例通过提取所述告警日志中的若干词语,并生成对应的词向量集合,可以准确获取所述告警日志中每个词语中所蕴含的特征信息;通过将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合,这一过程可以充分挖掘告警日志的信息,准确把握单个词语蕴含的特征信息和单个词语在整体词语中所展现的主题信息,进而准确把握告警日志的语义信息;通过采用全局特征向量集合和局部特征向量集合融合的方式,从而更为全面更为准确地捕获告警日志的信息,以提高后续相似度计算的准确性;通过高斯核函数计算所述综合特征向量集合的相似度,并通过聚类将语义相似程度高的告警日志简化为特定类别,可以准确判断出电网可能遭受的网络攻击类别,进而提高电网攻击类型的识别准确性和全面性。

[0011] 进一步的,所述获取电力网络系统的告警日志,具体为:

[0012] 获取电力网络系统的日志记录;

[0013] 对所述日志记录进行格式校正,并去除重复项,得到清洗后的日志记录;

[0014] 设定告警关键词,并保留含有所述告警关键词的日志记录,得到告警日志。

[0015] 这样通过对日志记录依次进行清洗和过滤操作,可以去除格式错误的日志记录和正常的日志记录,保留告警日志,避免因为告警日志自身的问题产生错误分类。

[0016] 进一步的,所述提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合,具体为:

[0017] 通过NLTK对所述告警日志中的若干词语进行分词处理和去除停用词处理,得到词语集合;

[0018] 将所述词语集合输入Word2vec模型,得到所述告警日志中所述词语对应的词向量集合。

[0019] 这样通过对所述告警日志的若干词语进行分词,并生成对应的词向量集合,可以准确获取所述告警日志中每个词语中所蕴含的特征信息。

[0020] 进一步的,所述得到局部特征向量集合,具体为:

[0021] 通过向所述词向量集合中的向量填充信息,统一所述词向量集合中的向量的长度;

[0022] 对所述告警日志中若干词语进行位置编码,得到所述告警日志中词语对应的位置向量;

[0023] 将统一长度后的词向量集合与词语对应的所述位置向量进行融合,得到输入词向量集合;

[0024] 将所述输入词向量集合输入Transformer编码器,得到注意力值集合;

[0025] 将所述注意力值集合和所述输入词向量集合进行残差连接,并对残差连接结果进行层归一化处理,得到所述告警日志中局部特征向量集合。

[0026] 这样通过对所述词向量集合的向量进行填充,得到输入词向量集合,可以保证所述告警日志中的样本长度保持一致,同时引入将词向量集合与词语对应的所述位置向量进行融合,可以有效解决后续的Transformer编码器无法分辨词语的位置信息的问题;另外通过将所述注意力值集合和所述输入词向量集合进行残差连接以及层归一化处理,可以准确把握单个词语蕴含的特征信息,进而准确把握告警日志的语义信息。

[0027] 进一步的,所述将所述输入词向量集合输入Transformer编码器,得到注意力值集合,具体为:

[0028] 基于所述输入词向量集合,得到Query向量集合、Key向量集合和Value向量集合;

[0029] 通过对所述Query向量集合和所述Key向量集合中的向量进行点积计算,得到所述告警日志中词语的注意力分数;

[0030] 对所述注意力分数进行放缩,并通过函数进行归一化处理,得到所述告警日志中词语的注意力权重;

[0031] 基于所述注意力权重与所述向量集合,得到所述告警日志中词语的注意力值集合。

[0032] 这样通过将所述输入词向量集合输入Transformer编码器,可以捕获同一个句子中词语之间的语法和语义特征。

[0033] 进一步的,所述得到全局特征向量集合,具体为:

[0034] 计算所述告警日志中的语义特征权重集合;

[0035] 基于所述语义特征权重集合提取所述告警日志中的关键词集合;

[0036] 将所述关键词集合输入HDP主题模型,得到所述告警日志中的全局特征向量集合。

[0037] 这样通过HDP主题模型可以准确得到单个词语在整体词语中所展现的主题信息,可以准确把握告警日志的语义信息。

[0038] 进一步的,所述计算所述告警日志中的语义特征权重集合,具体为:

[0039] 基于所述词向量集合中若干向量之间的余弦相似度,得到语义相似性权重集合;

[0040] 统计所述告警日志中词语共同出现的次数,得到词共现权重集合;

[0041] 基于TF-IDF计算所述告警日志中词语的词频权重集合;

[0042] 基于所述语义相似性权重集合、所述词共现权重集合和所述词频权重集合,得到所述告警日志中的语义特征权重集合。

[0043] 这样通过考虑语义相似性权重、词共现权重和词频权重,进而确认语义特征权重,可以充分挖掘告警日志的信息,准确把握单个词语在整体词语中所展现的主题信息,进而准确把握告警日志的语义信息。

[0044] 第二方面,本申请提供了一种告警信息分析系统,包括:获取模块、特征得到模块、融合模块和类别确认模块;

[0045] 所述获取模块,用于获取电力网络系统的告警日志,提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合;

[0046] 所述特征得到模块,用于将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合;

[0047] 所述融合模块,用于将所述局部特征向量集合和所述全局特征向量集合进行特征融合,得到综合特征向量集合;

[0048] 所述类别确认模块,用于通过高斯核函数计算所述综合特征向量集合的相似度,并根据相似度进行聚类确认告警类别。

[0049] 本申请实施例通过提取所述告警日志中的若干词语,并生成对应的词向量集合,可以准确获取所述告警日志中每个词语中所蕴含的特征信息;通过将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合,这一过程可以充分挖掘告警日志的信息,准确把握单个词语蕴含的特征信息和单个词语在整体词语中所展现的主题信息,进而准确把握告警日志的语义信息;通过采用全局特征向量集合和局部特征向量集合融合的方式,从而更为全面更为准确地捕获告警日志的信息,以提高后续相似度计算的准确性;通过高斯核函数计算所述综合特征向量集合的相似度,并通过聚类将语义相似程度高的告警日志简化为特定类别,可以准确判断出电网可能遭受的网络攻击类别,进而提高电网攻击类型的识别准确性和全面性。

[0050] 第三方面,本申请提供了一种终端设备,其特征在于,包括处理器、存储器以及存储在所述存储器中且被配置为由所述处理器执行的计算机程序,所述处理器执行所述计算机程序时实现如本申请所述的告警信息分析方法。

[0051] 第四方面,本申请提供了一种计算机可读存储介质,其特征在于,所述计算机可读存储介质包括存储的计算机程序,其中,在所述计算机程序运行时控制所述计算机可读存储介质所在设备执行如本申请所述的告警信息分析方法。

附图说明

[0052] 图1是本申请提供的一种告警信息分析方法的一种实施例的流程示意图;

[0053] 图2是本申请提供的确认告警类别的场景示意图;

[0054] 图3是本申请提供的一种告警信息分析方法的另一种实施例的流程示意图;

[0055] 图4是本申请提供的一种告警信息分析系统的一种实施例的结构示意图;

[0056] 图5是本申请提供的一种终端设备的结构示意图。

具体实施方式

[0057] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0058] 应当理解,文中所使用的步骤编号仅是为了方便描述,不作为对步骤执行先后顺序的限定。

[0059] 应当理解,在本发明说明书中所使用的术语仅仅是出于描述特定实施例的目的而并不意在限制本发明。如在本发明说明书和所附权利要求书中所使用的那样,除非上下文清楚地指明其他情况,否则单数形式的“一”、“一个”及“该”意在包括复数形式。

[0060] 术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在,但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

[0061] 术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合,并且包括这些组合。

[0062] NLTK是一个基于python的分词开源项目,可以将连续的字词序列按照一定的规范划分为独立词语序列,具有简单分词、命令行分词、词性标注以及词位置查询等功能,NLTK自带停词库,包含了文本中常见但对文本表达无实质帮助的停用词。

[0063] Word2Vec模型是一种词嵌入技术,可以将词语映射到向量空间,能够将词语转化为稠密的低维向量并保留词中的语义关系,有CBOW(continuous bag-of-word)和Skip-Gram(Continuous skip-gram Model)两种算法,均通过浅层神经网络对文本数据进行训练,将文本词语输入到模型中可以得到相应的词向量。

[0064] Transformer的编码器使用了注意力机制,因为注意力机制可以捕获同一个句子中词语之间的语法和语义特征,保留句子中长距离的依赖特征,Transformer通过编码整个输入序列并输出注意力编码,编码器由多层相同结构堆叠而成,每层结构主要由多头注意力和前馈神经网络两个组件构成,Transformer使用的位置编码为绝对位置编码。

[0065] TF-IDF(词频-逆向文件频率)是一种文本挖掘常用的加权技术,用于评估一个词对于文本的重要程度。TF代表词频,关键字在文本中出现的频率越高其重要程度越高;IDF代表词语的普遍程度,包含该词语的文件出现频率越高其重要程度越低。

[0066] 请参照图1,为本发明实施例提供的一种告警信息分析方法的流程示意图,包括步骤S1至步骤S4:

[0067] 步骤S1、获取电力网络系统的告警日志,提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合;包括步骤S11和步骤S12;

[0068] 具体的,步骤S11、获取电力网络系统的告警日志,具体为:

[0069] 获取电力网络系统的日志记录;对所述日志记录进行格式校正,并去除重复项,得到清洗后的日志记录;通过设定告警关键词,保留含有所述告警关键词的日志记录,得到告警日志。

[0070] 需要说明的是,对所述日志记录进行格式校正时需要先设计正确格式的正则表达式,通过正则表达式去除格式有误的日志记录;而去除重复项的具体操作为准备一个空的列表,遍历所有的日志记录,当日志记录不存在于列表,则将其添加到列表中,反之不进行添加,当遍历完成后也就得到了清洗后的日志记录。

[0071] 这样通过对日志记录依次进行清洗和过滤操作,可以去除格式错误的日志记录和正常的日志记录,保留告警日志,避免因为告警日志自身的问题产生错误分类。

[0072] 具体的,步骤S12、提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合,具体为:

[0073] 通过NLTK对所述告警日志中的若干词语进行分词处理和去除停用词处理,得到词语集合$\{w_1,w_2,\cdots,w_p,\cdots,w_q\}$;其中,$w_q$是指去除停用词后保留第q个词语,q为包含的词语总数;

[0074] 将所述词语集合$\{w_1,w_2,\cdots,w_p,\cdots,w_q\}$输入Word2vec模型,得到所述告警日志中所述词语对应的词向量集合$\{v_a,v_b,\cdots,v_p,\cdots,v_q\}$,其中,$v_q$是第q个词语$w_q$对应的词向量。

[0075] 需要说明的是,NLTK自带停词表可以去除词语集合中无意义词汇,即遍历词语集合中的所有词语,检查是否在停词表中,若存在则将其从词语集合中移除。

[0076] 这样通过对所述告警日志的若干词语进行分词,并生成对应的词向量集合,可以准确获取所述告警日志中每个词语中所蕴含的特征信息。

[0077] 步骤S2、将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合;

[0078] 具体的,所述得到局部特征向量集合,具体为:

[0079] 确认词向量集合$\{v_a,v_b,\cdots,v_p,\cdots,v_q\}$中词向量的最长长度,通过向所述词向量集合中的向量填充信息,统一所述词向量集合中的向量的长度;

[0080] 用正弦和余弦函数的组合交替对所述告警日志中若干词语进行位置编码,得到每个位置对应的编码元素,得到所述告警日志中词语对应的位置向量;其中所述正弦和余弦函数分别为:

$$PE(pos,2j)=sin\left(\frac{pos}{10000^{2j/d_k}}\right);$$

[0081]

$$PE(pos,2j+1)=cos\left(\frac{pos}{10000^{2j/d_k}}\right);$$

[0082] 式中,pos是位置索引,j是维度索引,d_k是词向量维度。

[0083] 将统一长度后的词向量集合与词语对应的所述位置向量进行逐元素相加求和,以实现向量融合,得到输入词向量集合$\{input_{i1},input_{i2},\cdots,input_{im},\cdots,input_{in}\}$;

[0084] 将所述输入词向量集合输入Transformer编码器,得到注意力值集合;具体为:

[0085] 基于所述输入词向量集合$\{input_{i1},input_{i2},\cdots,input_{im},\cdots,input_{in}\}$,分别乘以训

练所得的三个权重矩阵W_q,W_k,W_v,得到Query向量集合、Key向量集合和Value向量集合;

[0086] 通过对所述Query向量集合和所述Key向量集合中的向量进行点积计算,得到所述告警日志中词语的注意力分数{$Score_{1,1}$,$Score_{1,2}$,..,$Score_{1,q}$};其中,点积计算的公式为:

[0087] $$Score(Word_m, Word_n) = \sum_{j=1}^{d} Q_{m,j} \cdot K_{n,j};$$

[0088] 式中,$Word_m$为目标词语,$Word_n$为评分词语,$Q_{m,j}$为目标词语Query向量的第j个元素,$K_{n,j}$为评分词语Key向量的第j个元素。

[0089] 对所述注意力分数{$Score_{1,1}$,$Score_{1,2}$,..,$Score_{1,q}$}进行放缩,其中放缩公式为:

[0090] $$ScoreScale(Word_m, Word_n) = \frac{Score(Word_m, Word_n)}{\sqrt{d_k}};$$

[0091] 式中,ScoreScale($Word_m$,$Word_n$)为放缩结果,d_k为词向量维度,Score($Word_m$,$Word_n$)为注意力分数。

[0092] 并通过softmax函数进行归一化处理,得到所述告警日志中词语的注意力权重;其中,归一化的计算公式为:

[0093] $$SoftWeight(Word_m, Word_n) = \frac{e^{ScoreScale(Word_m, Word_n)}}{\sum_{j=1}^{p} e^{ScoreScale(Word_m, Word_j)}};$$

[0094] 其中,SoftWeight($Word_m$,$Word_n$)为注意力权重,ScoreScale($Word_m$,$Word_n$)为放缩结果,p为词语总数。

[0095] 需要说明的是,由于之前进行了向量填充操作,为了不让注意力放在填充位置上将注意力得分为0的替换为负无穷,会将其权重计算为0,而注意力权重在0到1之间。

[0096] 基于所述注意力权重与所述Value向量集合,进行加权求和后,得到所述告警日志中词语的注意力值集合{$Atten_{i,1}$,$Atten_{i,2}$,…,$Atten_{i,m}$},其中,注意力值计算求和公式为:

[0097] $$Atten_{i,m} = \sum_{j=1}^{p} SoftWeight(Word_m, Word_n) \times Value_j;$$

[0098] 其中,$Atten_{i,m}$为告警日志l_i第m个词语的注意力向量,SoftWeight($Word_m$,$Word_n$)为注意力权重,$Value_j$为第j个Value向量。

[0099] 最后将所述注意力值集合和所述输入词向量集合进行残差连接,并对残差连接结果进行层归一化处理,得到所述告警日志中局部特征向量集合{$Atten_1$,$Atten_2$,…,$Atten_N$}。

[0100] 需要说明的是,局部特征向量是对单个告警日志进行上下文特征提取后的表示向量。

[0101] 这样通过对所述词向量集合的向量进行填充,得到输入词向量集合,可以保证所述告警日志中的样本长度保持一致,同时引入将词向量集合与词语对应的所述位置向量进行融合,可以有效解决后续的Transformer编码器无法分辨词语的位置信息的问题;另外通过将所述注意力值集合和所述输入词向量集合进行残差连接以及层归一化处理,可以准确把握单个词语蕴含的特征信息,进而准确把握告警日志的语义信息。

[0102] 具体的,所述得到全局特征向量集合,具体为:

[0103] 计算所述告警日志中的语义特征权重集合;具体为:

[0104] 基于所述词向量集合中若干向量之间的余弦相似度,得到语义相似性权重集合;其中,语义相似性权重函数为:

[0105] $$SimWeight(w_i,Log)=\frac{\sum_{j=1}^{p}WS(w_i,w_j)-WS(w_i,w_i)}{p-1};$$

[0106] 式中,SimWeight(w_i,Log)为告警日志Log中w_i的语义相似性权重,WS(w_i,w_j)为告警日志Log中两个词语w_i,w_j的余弦相似度,p为词语总数。

[0107] 统计所述告警日志中词语共同出现的次数,得到词共现权重集合;其中,词共现权重函数为:

[0108] $$CoWeight(w_i,Log)=\frac{\sum_{j=1}^{k}WordCo(w_i,w_j)-WordCo(w_i,w_i)}{WordCo(w_i,w_i)\times(k-1)};$$

[0109] 式中,CoWeight(w_i,Log)为告警日志Log中w_i的词共现权重,WordCo(w_i,w_j)为词语w_i与词语w_j的词共现次数,k为告警日记总数。

[0110] 基于TF-IDF计算所述告警日志中词语的词频权重集合;其中,词频权重的计算公式为:

[0111] $$FreWeight(w_i,Log)=\frac{n_{i,j}}{\sum_k n_{k,j}}\times log\frac{|D|}{1+|j:w_i\in l_j|};$$

[0112] 式中,FreWeight(w_i,Log)为告警日志Log中w_i的词频权重,$n_{i,j}$为词语w_i在告警日志l_j中出现的次数,$n_{k,j}$为词语w_k在告警日志l_j中出现的次数,|D|为所有告警日志的数量,|j:$w_i\in l_j$|表示包含词语w_i的告警日志数目。

[0113] 需要说明的是,词频权重值FreWeight由TF-IDF值表示,TF为词语在告警日志中出现的频率,IDF表示词语在文档中的普遍程度。

[0114] 基于所述语义相似性权重集合、所述词共现权重集合和所述词频权重集合,得到所述告警日志中的语义特征权重集合,其中,语义特征权重的计算公式为:

[0115] SemFeaWeight(w_i,Log)=

[0116] SimWeight(w_i,Log)*CoWeight(w_i,Log)*FreWeight(w_i,Log);

[0117] 式中,SemFeaWeight(w_i,Log)为告警日志Log中词语w_i的语义特征权重,SimWeight(w_i,Log)为告警日志Log中w_i的语义相似性权重,CoWeight(w_i,Log)为告警日志Log中w_i的词共现权重,FreWeight(w_i,Log)为告警日志Log中w_i的词频权重。

[0118] 需要说明的是,所述语义相似性权重为某个词语与告警日志中其他词语语义相似度求和后的平均值;所述词共现权重为某个词语与告警日志中其余词语一起出现的次数的平均值,并除以该词语在文档中出现的次数表示;所述词频权重为,词语在告警日志中出现的频率和词语在文档中的普遍程度计算表示。

[0119] 基于所述语义特征权重集合提取所述告警日志中的关键词集合{$Word_1$,$Word_2$,…,$Word_k$};

[0120] 需要说明的是,所述关键词为根据语义特征权重大小进行排序,且根据告警日志中包含的词语数量定义关键词保留比例,进而确认保留的关键词数量。

[0121] 将所述关键词集合{$Word_1$,$Word_2$,…,$Word_k$}输入HDP主题模型,得到所述告警日志中的全局特征向量集合{vec_1,vec_2,…,vec_i,…,vec_N}。

[0122] 需要说明的是,全局特征向量是对告警日志进行主题信息特征提取后的表示向量。

[0123] 这样通过考虑语义相似性权重、词共现权重和词频权重,进而确认语义特征权重,

可以充分挖掘告警日志的信息,并通过HDP主题模型可以准确得到单个词语在整体词语中所展现的主题信息,可以准确把握告警日志的语义信息。

[0124] 步骤S3、将所述局部特征向量集合和所述全局特征向量集合进行特征融合,得到综合特征向量集合;

[0125] 具体的将局部特征向量集合$\{Atten_1,Atten_2,\cdots,Atten_N\}$和全局特征向量集合$\{vec_1,vec_2,\cdots,vec_i,\cdots,vec_N\}$进行拼接完成特征融合,得到最终的综合特征向量集合$\{Atten_1:vec_1,Atten_2:vec_2,\cdots,Atten_N:vec_N\}$。

[0126] 这样通过采用全局特征向量集合和局部特征向量集合融合的方式,从而更为全面更为准确的捕获告警日志的信息,以提高后续相似度计算的准确性。

[0127] 步骤S4、通过高斯核函数计算所述综合特征向量集合的相似度,并根据相似度进行聚类确认告警类别;

[0128] 具体的,通过高斯核函数计算所述综合特征向量集合的相似度,并对样本在图空间中进行谱聚类,通过多次训练调整谱聚类的模型参数,获得最优的谱聚类结果,进而确认告警日志所对应的类别,自然也就确定了网络攻击类别;其中,确认告警类别的场景示意图如图2所示;

[0129] 其中,所述高斯核函数的计算公式为:

[0130] $$s(x_i,x_j)=e^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}};$$

[0131] 式中,x_i,x_j为两个向量样本,$||x_i-x_j||$为两个向量样本间的欧氏距离,σ为高斯核函数带宽参数。

[0132] 需要说明的是,谱聚类是一种聚类模型,能够处理非线性聚类结构。谱聚类作为聚类模型的一大特点是引入图论知识,将数据转化为空间中的点,以点之间的远近和边的权重标记文本数据的相似性,通过图的特征分解完成聚类。同时,谱聚类具备得到全局最优解的优点。

[0133] 这样通过高斯核函数计算所述综合特征向量集合的相似度,并通过聚类将语义相似程度高的告警日志简化为特定类别。

[0134] 需要说明的是,本申请可以针对电力系统网络攻击方法主要考虑暴力破解攻击、窃听攻击、XSS攻击、SQL注入攻击、Dos攻击、DDos攻击、MITM攻击、会话劫持攻击等攻击类别的告警日志作为核心聚类数据,经过聚类处理后得到相应簇,对应得到所属类别,完成聚类。

[0135] 本申请实施例通过提取所述告警日志中的若干词语,并生成对应的词向量集合,可以准确获取所述告警日志中每个词语中所蕴含的特征信息;通过将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合,这一过程可以充分挖掘告警日志的信息,准确把握单个词语蕴含的特征信息和单个词语在整体词语中所展现的主题信息,进而准确把握告警日志的语义信息;通过采用全局特征向量集合和局部特征向量集合融合的方式,从而更为全面更为准确地捕获告警日志的信息,以提高后续相似度计算的准确性;通过高斯核函数计算所述综合特征向量集合的相似度,并通过聚类将语义相似程度高的告警日志简化为特定类别,可以准确判断出电网可能遭受的网络攻击类别,进而提高电网攻击类型的识别准确性和全面性。

[0136] 本申请还提供了图3以方便理解,图3是本申请提供的一种告警信息分析方法的另一种实施例的流程示意图,其中图3所涉及的步骤已经在上述详细展开,此处不再赘述。

[0137] 请参照图4,图4为本发明实施例提供的一种告警信息分析系统的结构示意图,包括:获取模块01、特征得到模块02、融合模块03和类别确认模块04;

[0138] 所述获取模块01,用于获取电力网络系统的告警日志,提取所述告警日志中的若干词语,并生成所述词语对应的词向量集合;

[0139] 所述特征得到模块02,用于将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合;

[0140] 所述融合模块03,用于将所述局部特征向量集合和所述全局特征向量集合进行特征融合,得到综合特征向量集合;

[0141] 所述类别确认模块04,用于通过高斯核函数计算所述综合特征向量集合的相似度,并根据相似度进行聚类确认告警类别。

[0142] 上述告警信息分析系统内的各模块之间信息交互、执行过程等内容,由于与本发明第一方面的告警信息分析方法的实施例基于同一构思,所实现的技术效果基本相同,具体内容可参见本发明方法实施例一中的叙述,此处不再赘述。

[0143] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的模块可以是或者也可以不是物理上分开的,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方法的目的。

[0144] 图5为一种终端设备的结构示意图。如图5所示,该实施例的终端设备5包括:至少一个处理器501(图5中仅示出一个)处理器、存储器502以及存储在存储器502中并可在至少一个处理器501上运行的计算机程序503,处理器501执行计算机程序503时实现上述任意方法实施例中的步骤。

[0145] 终端设备5可以是智能手机、笔记本电脑、平板电脑和桌上型计算机等计算设备。该终端设备可包括但不仅限于处理器501、存储器502。本领域技术人员可以理解,图5仅仅是终端设备5的举例,并不构成对终端设备5的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件,例如还可以包括输入输出设备、网络接入设备等。

[0146] 所称处理器501可以是中央处理单元(Central Processing Unit,CPU),该处理器501还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0147] 存储器502在一些实施例中可以是终端设备5的内部存储单元,例如终端设备5的硬盘或内存。存储器502在另一些实施例中也可以是终端设备5的外部存储设备,例如终端设备5上配备的插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。进一步地,存储器502还可以既包括终端设备5的内部存储单元也包括外部存储设备。存储器502用于存储操作系统、应用程序、引导装载程序(BootLoader)、数据以及其他程序等,例如计算机程序的程序代码等。存储器502还可以用于暂时地存储已经输出或者将要输出的数据。

[0148] 另外,本发明还提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时,实现如上述实施例一所述的告警信息分析方法。

[0149] 本申请实施例提供了一种计算机程序产品,当计算机程序产品在终端设备上运行时,使得终端设备执行时实现上述各个方法实施例中的步骤。

[0150] 在本申请所提供的几个实施例中,可以理解的是,流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分,模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意的是,在有些作为替换的实现方式中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。

[0151] 功能如果以软件功能模块的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台终端设备执行本申请各个实施例方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0152] 综上,本发明提供的一种告警信息分析方法、系统、设备及存储介质,通过提取所述告警日志中的若干词语,并生成对应的词向量集合,可以准确获取所述告警日志中每个词语中所蕴含的特征信息;通过将所述词向量集合分别输入Transformer编码器和HDP主题模型,分别得到局部特征向量集合和全局特征向量集合,这一过程可以充分挖掘告警日志的信息,准确把握单个词语蕴含的特征信息和单个词语在整体词语中所展现的主题信息,进而准确把握告警日志的语义信息;通过采用全局特征向量集合和局部特征向量集合融合的方式,从而更为全面更为准确地捕获告警日志的信息,以提高后续相似度计算的准确性;通过高斯核函数计算所述综合特征向量集合的相似度,并通过聚类将语义相似程度高的告警日志简化为特定类别,可以准确判断出电网可能遭受的网络攻击类别,进而提高电网攻击类型的识别准确性和全面性。

[0153] 以上所述的具体实施例,对本发明的目的、技术方案和有益效果进行了进一步的详细说明,应当理解,以上所述仅为本发明的具体实施例而已,并不用于限定本发明的保护范围。特别指出,对于本领域技术人员来说,凡在本发明的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

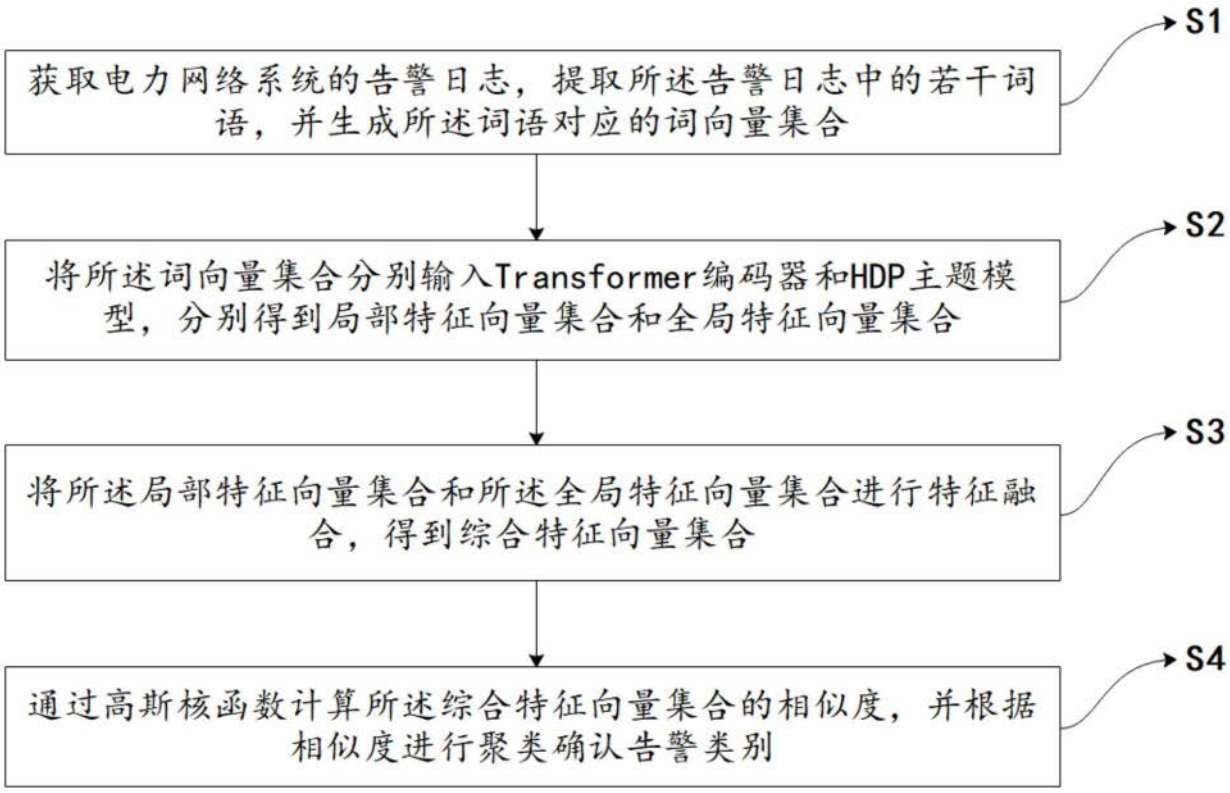

S1
获取电力网络系统的告警日志，提取所述告警日志中的若干词语，并生成所述词语对应的词向量集合
S2
将所述词向量集合分别输入Transformer编码器和HDP主题模型，分别得到局部特征向量集合和全局特征向量集合
S3
将所述局部特征向量集合和所述全局特征向量集合进行特征融合，得到综合特征向量集合
S4
通过高斯核函数计算所述综合特征向量集合的相似度，并根据相似度进行聚类确认告警类别

图1

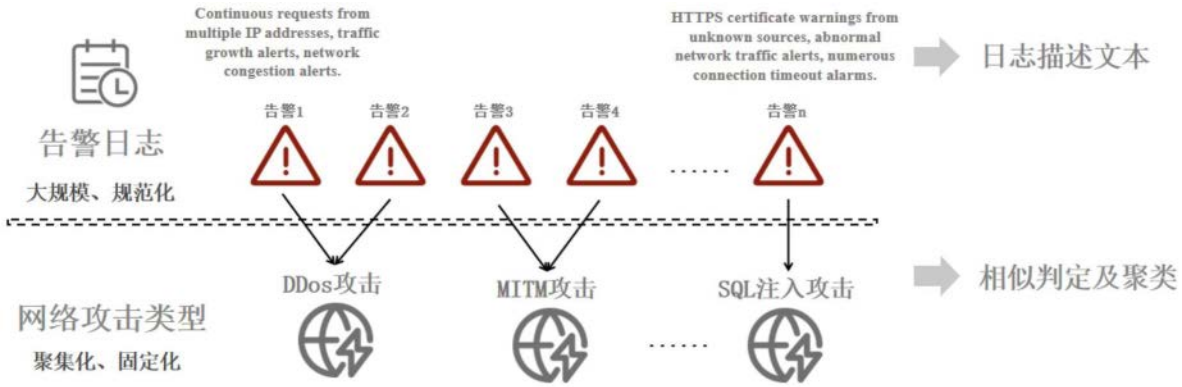

Continuous requests from multiple IP addresses, traffic growth alerts, network congestion alerts.
HTTPS certificate warnings from unknown sources, abnormal network traffic alerts, numerous connection timeout alarms.
日志描述文本
告警日志
大规模、规范化
告警1
告警2
告警3
告警4
告警n
DDos攻击
MITM攻击
SQL注入攻击
网络攻击类型
聚集化、固定化
相似判定及聚类

图2

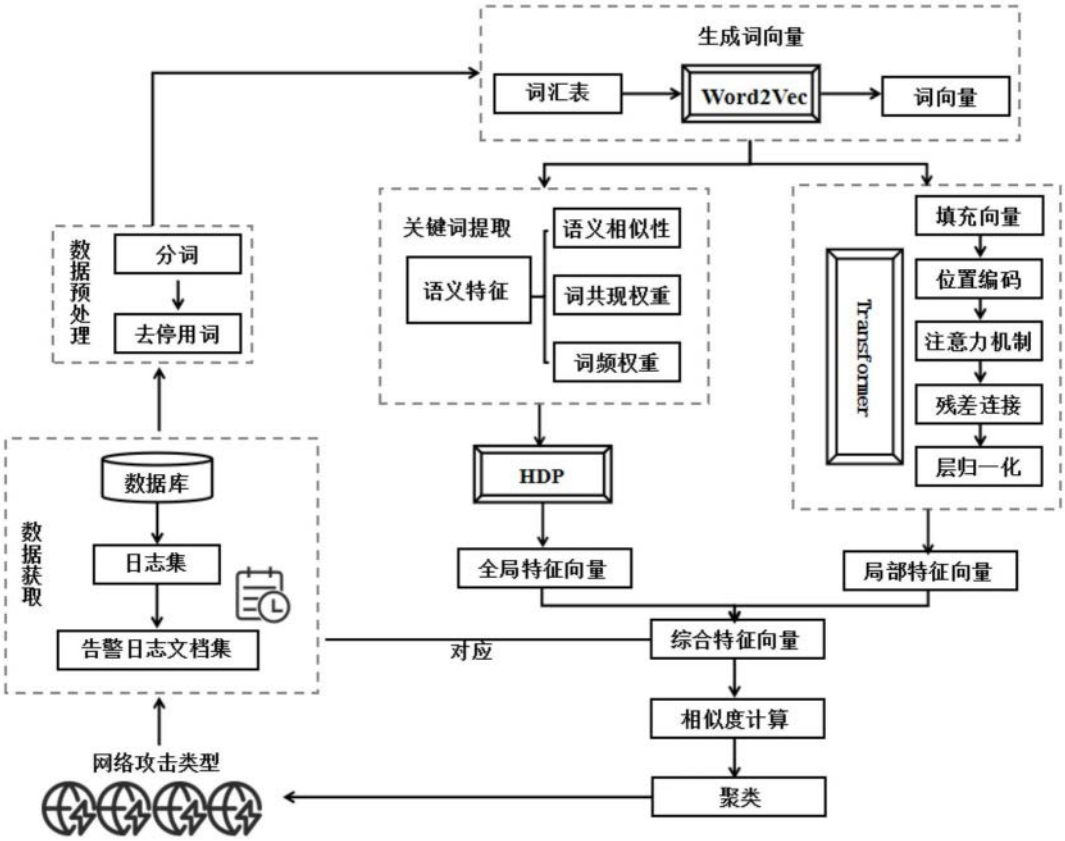

生成词向量
词汇表
Word2Vec
词向量
数据预处理
分词
去停用词
关键词提取
语义特征
语义相似性
词共现权重
词频权重
Transformer
填充向量
位置编码
注意力机制
残差连接
层归一化
数据获取
数据库
日志集
告警日志文档集
HDP
全局特征向量
局部特征向量
对应
综合特征向量
相似度计算
聚类
网络攻击类型

图3

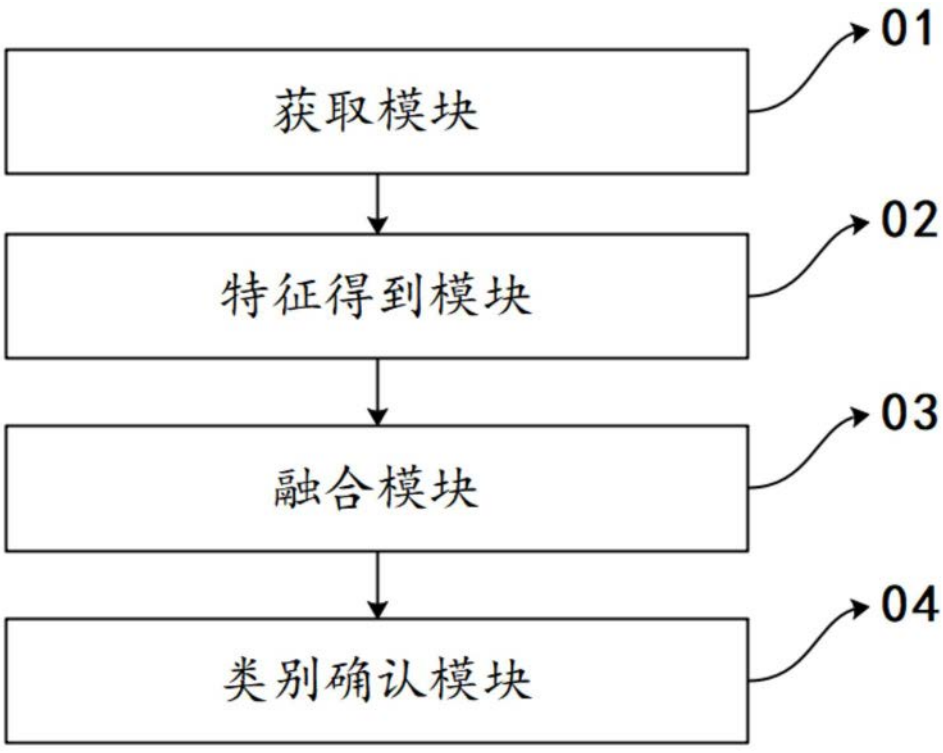

获取模块
01
特征得到模块
02
融合模块
03
类别确认模块
04

图4

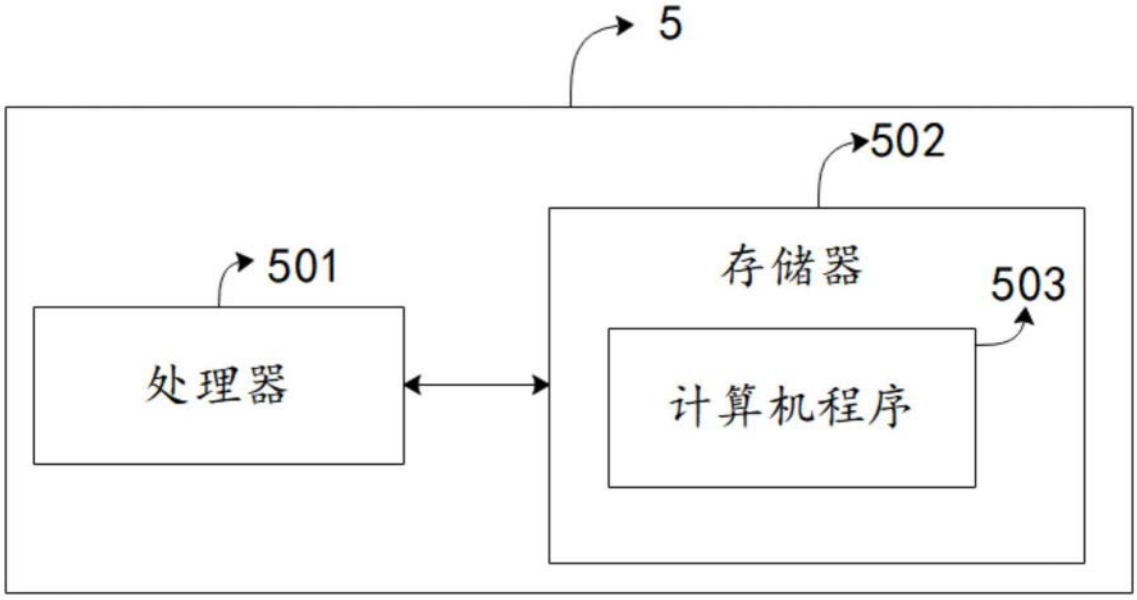


图5