(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116910603 A
(43)申请公布日 2023.10.20

(21)申请号 202310620030.1

(22)申请日 2023.05.29

(71)申请人 北京东方通网信科技有限公司
地址 100044 北京市海淀区中关村南大街2号1号楼19层A座2201

(72)发明人 李睿凡 黄永军 陈乔

(74)专利代理机构 北京辰权知识产权代理有限公司 11619
专利代理师 刘广达

(51)Int.Cl.
G06F 18/24(2023.01)
G06F 21/57(2013.01)
G06F 21/64(2013.01)
G06F 16/9532(2019.01)
G06F 16/955(2019.01)
G06F 40/30(2020.01)
G06V 20/40(2022.01)

权利要求书3页 说明书9页 附图4页

(54)发明名称

一种基于人工智能挖掘的网络内容风控管理系统

(57)摘要

本申请提供一种基于人工智能挖掘的网络内容风控管理系统，包括：新业务内容检测模块，用于通过工具自动获取新业务内容，并对所述新业务内容进行基于人工智能挖掘的审核、分类，所述新业务内容至少包括APP、公众号、微博、微信；网站内容监管模块，对预设网站库内的网站内容、暗链、外链、篡改、漏洞、可用性多个维度进行检测；企业内容安全治理模块，基于主动拨测、旁路检测、文件共享的内容采集技术，自动化获取文本、视频、图片、音频、复杂文档内容，并进行基于人工智能挖掘的内容审核；UGC内容审核模块，用于对用户原创内容进行审核，所述用户原创内容至少包括视频直播、婚恋交友、社区论坛、电商网站、和在线教育。本申请实现了对网络内容风控的及时、准确处理。

CN 116910603 A

1.一种基于人工智能挖掘的网络内容风控管理系统,其特征在于,包括:

新业务内容检测模块,用于通过工具自动获取新业务内容,并对所述新业务内容进行基于人工智能挖掘的审核、分类,所述新业务内容至少包括APP、公众号、微博、微信;

网站内容监管模块,对预设网站库内的网站内容、暗链、外链、篡改、漏洞、可用性多个维度进行检测;

企业内容安全治理模块,基于主动拨测、旁路检测、文件共享的内容采集技术,自动化获取文本、视频、图片、音频、复杂文档内容,并进行基于人工智能挖掘的内容审核;

UGC内容审核模块,用于对用户原创内容进行审核,所述用户原创内容至少包括视频直播、婚恋交友、社区论坛、电商网站、和在线教育。

2.根据权利要求1所述的系统,其特征在于,

所述通过工具自动获取新业务内容,并对所述新业务内容进行基于人工智能挖掘的审核、分类,包括:

通过爬取、流量还原、识别内容主动送检中的一种或多种自动获取所述新业务内容的历史挖掘数据;

根据合规数据库,对所述历史挖掘数据进行标注,得到标注后的历史挖掘数据集;

构建特征向量空间网络,并使用所述历史挖掘数据集输入所述特征向量空间网络进行训练,得到训练好的特征向量空间网络;

将实时获取的待挖掘数据输入所述训练好的特征向量空间网络,以判断所述待挖掘数据中是否包含违规内容。

3.根据权利要求2所述的系统,其特征在于,

所述对预设网站库内的网站内容、暗链、外链、篡改、漏洞、可用性多个维度进行检测,包括:

提取待检测网页的网页内容,对网页内容进行关键词匹配,判断网页内容中是否存在一个或多个预设关键词;若存在一个或多个预设关键词,发出匹配提示;

提取待检测网页的网页内容,对网页内容进行暗词匹配,判断网页内容中是否存在一个或多个预设暗词;若存在一个或多个预设暗词,则将一个或多个预设暗词确定为一个或多个目标暗词,并根据一个或多个目标暗词在合法网页中出现的目标概率得到网页异常参数;若网页异常参数大于标准参数,则确定检测到暗链;

获取待检测网站的外部链接信息以及外部链接所属网站域名信息;根据外部链接信息获取外部链接的页面内容信息,获取外部链接所属网站域名的特征信息;对外部链接的页面内容信息进行敏感内容识别,确定外部链接的页面内容的敏感内容分数;将外部链接所属网站域名的特征信息、相似度值和敏感内容分数对应的向量输入检测模型,获得外部链接的检测结果;

收集待监测网站的网页原始数据;从所述网页原始数据中提取页面元数据特征;将所述页面元数据特征输入到训练后的神经网络中进行篡改检测;根据篡改检测结果判断待监测网站是否存在篡改;

获取预设网站库里网站的每个网页的URL;基于每个网页的URL,确定所述预设网站库里网站的登录网页;检测所述登录网页是否存在验证码,得到第一检测结果;基于所述第一检测结果,检测所述目标网站是否存在漏洞;

获取待检测的网站的站点IP,并建立待检测的站点IP对应的人工智能模型;提取日常Web业务访问的目的IP,判断目的IP与人工智能模型的建模IP是否一致;若目的IP与人工智能模型的建模IP一致,则对符合建模IP的Web请求信息做关键信息统计,判断关键信息是否在预设的正常范围内;若关键信息不在预设的正常范围内,该网站不可用,若关键信息在预设的正常范围内,该网站可用。

4.根据权利要求3所述的系统,其特征在于,

所述基于主动拨测、旁路检测、文件共享的内容采集技术,自动化获取文本、视频、图片、音频、复杂文档内容,并进行基于人工智能挖掘的内容审核,包括:

建立拨测任务;选择拨测目标企业;对拨测目标企业进行验证访问;执行拨测任务,在每一次拨测任务的执行周期中,对选择的拨测目标企业进行主动测试获得文本、视频、图片、音频、复杂文档内容的各项指标;提取所述指标大于预设阈值的文本、视频、图片、音频、复杂文档内容;

将旁路检测设备的检测接口设置为二层接口,并将检测接口与交换机相连;交换机通过镜像将流量上送到所述旁路检测设备上检测;在检测接口上配置旁路检测功能,使旁路检测设备只检测而不转发流量;配置安全策略,引用需要的安全配置文件,对流量进行对应的内容安全检测;

通过预设文件共享端口获取待检测企业的文本、视频、图片、音频、复杂文档内容;

将所述文本、视频、图片、音频、复杂文档内容与基于人工智能构建的预设匹配数据库进行匹配度计算,将匹配度高于预设阈值的数据提取并给出预警提示。

5.根据权利要求4所述的系统,其特征在于,

当所述旁路检测设备只有一个接口接收镜像流量或者多个接口接收镜像流量但针对各个接口接收的流量配置相同的安全策略时,将接口加入任何安全区域,将安全策略的源安全区域和目的安全区域配置成any;当所述旁路检测设备有多个接口接收镜像流量,且要针对各个接口接收的流量配置不同的安全策略时,将接口加入不同的安全区域,将安全策略的源安全区域和目的安全区域配置成检测接口所在的安全区域。

6.根据权利要求1-5任一项所述的系统,其特征在于,

所述对用户原创内容进行审核,包括:

文本识别:对所述用户原创内容中的文本进行语义识别、语境分析;

图像识别:建立图片MD5指纹库,对所述用户原创内容中的图像进行图像识别;

视频识别:通过解码器对视频数据解码,提取视频特征,对所述用户原创内容中的视频进行识别;

音频识别:采用语言识别算法、关键词检索、声纹识别对所述用户原创内容中的音频进行识别;

将所述文本识别、图像识别、视频识别、音频识别得到的数据输入训练好的特征向量空间网络,以判断所述数据中是否包含违规内容。

7.根据权利要求6所述的系统,其特征在于,

所述图像识别具体包括以下步骤:获取所述用户原创内容中的图像的向量矩阵;对所述用户原创内容中的图像的向量矩阵进行字符串转换处理,得到所述用户原创内容中的图像的字符串;对所述用户原创内容中的图像的字符串进行加密,得到所述用户原创内容中

的图像的加密字符串;对所述用户原创内容中的图像的加密字符串,与所述用户原创内容中的图像的MD5值进行拼接操作,得到所述用户原创内容中的图像的组合数据;将所述用户原创内容中的图像的组合数据,发送至云端服务器;接收云端服务器根据组合数据反馈的图像识别结果。

8.根据权利要求7所述的系统,其特征在于,

所述对所述用户原创内容中的文本进行语义识别、语境分析,包括以下步骤:

基于长窗口构建语义模型,基于短窗口构建语境模型;使用用户原创内容历史语料库基于所述语义模型和语境模型训练得到所述用户原创内容的中文词向量模型;使用所述用户原创内容的中文词向量模型对实时输入的所述用户原创内容的文本进行识别并输出识别结果。

9.一种电子设备,包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,其特征在于,所述处理器运行所述计算机程序以实现如权利要求1-8任一项所述的系统。

10.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行实现如权利要求1-8中任一项所述的系统。

一种基于人工智能挖掘的网络内容风控管理系统

技术领域

[0001] 本申请涉及人工智能技术领域，尤其涉及一种基于人工智能挖掘的网络内容风控管理系统。

背景技术

[0002] 上世纪后期开始的互联网时代，极大的便利了人们的信息传输、提升了工作效率，产生了多种多样的经济模式。然而，由于互联网的自由性，世界范围内都存在着有些不法分子往往依靠网络进行一些公开或者隐蔽的不法活动，或者是虽然不违法但是违反社会公德的行为。

[0003] 对于网络内容的风险控制和管理，是互联网监管部门面临的棘手而必须要做的重要任务。目前的网络内容风控管理，往往需要大量的人工现场阅读、登录来进行监管。这种监管效率低，而且容易导致漏检。

发明内容

[0004] 有鉴于此，本申请的目的在于提出一种基于人工智能挖掘的网络内容风控管理系统，本申请能够针对性的解决现有的问题。

[0005] 基于上述目的，本申请还提出了一种基于人工智能挖掘的网络内容风控管理系统，包括：

[0006] 新业务内容检测模块，用于通过工具自动获取新业务内容，并对所述新业务内容进行基于人工智能挖掘的审核、分类，所述新业务内容至少包括APP、公众号、微博、微信；

[0007] 网站内容监管模块，对预设网站库内的网站内容、暗链、外链、篡改、漏洞、可用性多个维度进行检测；

[0008] 企业内容安全治理模块，基于主动拨测、旁路检测、文件共享的内容采集技术，自动化获取文本、视频、图片、音频、复杂文档内容，并进行基于人工智能挖掘的内容审核；

[0009] UGC内容审核模块，用于对用户原创内容进行审核，所述用户原创内容至少包括视频直播、婚恋交友、社区论坛、电商网站、和在线教育。

[0010] 进一步地，所述通过工具自动获取新业务内容，并对所述新业务内容进行基于人工智能挖掘的审核、分类，包括：

[0011] 通过爬取、流量还原、识别内容主动送检中的一种或多种自动获取所述新业务内容的历史挖掘数据；

[0012] 根据合规数据库，对所述历史挖掘数据进行标注，得到标注后的历史挖掘数据集；

[0013] 构建特征向量空间网络，并使用所述历史挖掘数据集输入所述特征向量空间网络进行训练，得到训练好的特征向量空间网络；

[0014] 将实时获取的待挖掘数据输入所述训练好的特征向量空间网络，以判断所述待挖掘数据中是否包含违规内容。

[0015] 进一步地，所述对预设网站库内的网站内容、暗链、外链、篡改、漏洞、可用性多个

维度进行检测，包括：

[0016] 提取待检测网页的网页内容；对网页内容进行关键词匹配，判断网页内容中是否存在一个或多个预设关键词；若存在一个或多个预设关键词，发出匹配提示；

[0017] 提取待检测网页的网页内容；对网页内容进行暗词匹配，判断网页内容中是否存在一个或多个预设暗词；若存在一个或多个预设暗词，则将一个或多个预设暗词确定为一个或多个目标暗词，并根据一个或多个目标暗词在合法网页中出现的目标概率得到网页异常参数；若网页异常参数大于标准参数，则确定检测到暗链；

[0018] 获取待检测网站的外部链接信息以及外部链接所属网站域名信息；根据外部链接信息获取外部链接的页面内容信息，获取外部链接所属网站域名的特征信息；对外部链接的页面内容信息进行敏感内容识别，确定外部链接的页面内容的敏感内容分数；将外部链接所属网站域名的特征信息、相似度值和敏感内容分数对应的向量输入检测模型，获得外部链接的检测结果；

[0019] 收集待监测网站的网页原始数据；从所述网页原始数据中提取页面元数据特征；将所述页面元数据特征输入到训练后的神经网络中进行篡改检测；根据篡改检测结果判断待监测网站是否存在篡改；

[0020] 获取预设网站库里网站的每个网页的URL；基于每个网页的URL，确定所述预设网站库里网站的登录网页；检测所述登录网页是否存在验证码，得到第一检测结果；基于所述第一检测结果，检测所述目标网站是否存在漏洞；

[0021] 获取待检测的网站的站点IP，并建立待检测的站点IP对应的人工智能模型；提取日常Web业务访问的目的IP，判断目的IP与人工智能模型的建模IP是否一致；若目的IP与人工智能模型的建模IP一致，则对符合建模IP的Web请求信息做关键信息统计，判断关键信息是否在预设的正常范围内；若关键信息不在预设的正常范围内，该网站不可用，若关键信息在预设的正常范围内，该网站可用。

[0022] 进一步地，所述基于主动拨测、旁路检测、文件共享的内容采集技术，自动化获取文本、视频、图片、音频、复杂文档内容，并进行基于人工智能挖掘的内容审核，包括：

[0023] 建立拨测任务；选择拨测目标企业；对拨测目标企业进行验证访问；执行拨测任务，在每一次拨测任务的执行周期中，对选择的拨测目标企业进行主动测试获得文本、视频、图片、音频、复杂文档内容的各项指标；提取所述指标大于预设阈值的文本、视频、图片、音频、复杂文档内容；

[0024] 将旁路检测设备的检测接口设置为二层接口，并将检测接口与交换机相连；交换机通过镜像将流量上送到所述旁路检测设备上检测；在检测接口上配置旁路检测功能，使旁路检测设备只检测而不转发流量；配置安全策略，引用需要的安全配置文件，对流量进行对应的内容安全检测；

[0025] 通过预设文件共享端口获取待检测企业的文本、视频、图片、音频、复杂文档内容；

[0026] 将所述文本、视频、图片、音频、复杂文档内容与基于人工智能构建的预设匹配数据库进行匹配度计算，将匹配度高于预设阈值的数据提取并给出预警提示。

[0027] 进一步地，当所述旁路检测设备只有一个接口接收镜像流量或者多个接口接收镜像流量但针对各个接口接收的流量配置相同的安全策略时，将接口加入任何安全区域，将安全策略的源安全区域和目的安全区域配置成any；当所述旁路检测设备有多个接口接收

镜像流量，且要针对各个接口接收的流量配置不同的安全策略时，将接口加入不同的安全区域，将安全策略的源安全区域和目的安全区域配置成检测接口所在的安全区域。

[0028] 进一步地，所述对用户原创内容进行审核，包括：

[0029] 文本识别：对所述用户原创内容中的文本进行语义识别、语境分析；

[0030] 图像识别：建立图片MD5指纹库，对所述用户原创内容中的图像进行图像识别；

[0031] 视频识别：通过解码器对视频数据解码，提取视频特征，对所述用户原创内容中的视频进行识别；

[0032] 音频识别：采用语言识别算法、关键词检索、声纹识别对所述用户原创内容中的音频进行识别；

[0033] 将所述文本识别、图像识别、视频识别、音频识别得到的数据输入训练好的特征向量空间网络，以判断所述数据中是否包含违规内容。

[0034] 进一步地，所述图像识别具体包括以下步骤：获取所述用户原创内容中的图像的向量矩阵；对所述用户原创内容中的图像的向量矩阵进行字符串转换处理，得到所述用户原创内容中的图像的字符串；对所述用户原创内容中的图像的字符串进行加密，得到所述用户原创内容中的图像的加密字符串；对所述用户原创内容中的图像的加密字符串，与所述用户原创内容中的图像的MD5值进行拼接操作，得到所述用户原创内容中的图像的组合数据；将所述用户原创内容中的图像的组合数据，发送至云端服务器；接收云端服务器根据组合数据反馈的图像识别结果。

[0035] 进一步地，所述对所述用户原创内容中的文本进行语义识别、语境分析，包括以下步骤：

[0036] 基于长窗口构建语义模型，基于短窗口构建语境模型；使用用户原创内容历史语料库基于所述语义模型和语境模型训练得到所述用户原创内容的中文词向量模型；使用所述用户原创内容的中文词向量模型对实时输入的所述用户原创内容的文本进行识别并输出识别结果。

[0037] 总的来说，本申请的优势及给用户带来的体验在于：

[0038] 1、本申请实现了对网络内容风控的及时、准确、高效处理，不需要大量的人工现场阅读、登录来进行监管网站、APP、微博微信等新兴业务内容，实现了对各种门户网站、企业内容、UGC用户原创内容的实时监管，监管效率高，而且不容易漏检。

[0039] 2、高效的采集能力：海量数据采集能力，支持WEB、APP、短视频等类型的数据采集；文本识别：语义识别、语境分析，文本识别率达到98％以上；图像识别：精准识别图片中不良内容。建立图片MD5指纹库，图片识别率达到93％以上；视频识别：解码器对视频数据解码，提取视频特征，视频识别率达到95％以上；音频识别：采用语言识别算法、关键词检索、声纹识别，业内领先的语音识别能力。

附图说明

[0040] 在附图中，除非另外规定，否则贯穿多个附图相同的附图标记表示相同或相似的部件或元素。这些附图不一定是按照比例绘制的。应该理解，这些附图仅描绘了根据本申请公开的一些实施方式，而不应将其视为是对本申请范围的限制。

[0041] 图1示出本申请的系统架构原理示意图。

[0042] 图2示出根据本申请实施例的基于人工智能挖掘的网络内容风控管理系统的构成图。

[0043] 图3示出新业务内容检测模块的具体实现方法示意图。

[0044] 图4示出根据本申请实施例的网站内容监管模块功能示意图。

[0045] 图5示出了本申请一实施例所提供的一种电子设备的结构示意图。

[0046] 图6示出了本申请一实施例所提供的一种存储介质的示意图。

具体实施方式

[0047] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

[0048] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0049] 图1示出本申请的系统架构原理示意图。本申请的实施例中，网络风险风控系统主要通过爬取、流量还原、识别内容主动送检等方式自动获取web网页、媒体客户端、社交通讯、短视频、管道流量等内容，并进行内容安全识别，识别类别包括涉黄、涉暴、违禁、挂马、暗链、谩骂、违规、低俗、欺诈、灌水等。采用的技术手段包括但不限于：文本视频、图像识别、视频识别、音频识别、相似度识别、复杂文档提取、格式预处理等。

[0050] 申请实施例提供了一种基于人工智能挖掘的网络内容风控管理系统，如图2所示，该系统包括：

[0051] 新业务内容检测模块101，用于通过工具自动获取新业务内容，并对所述新业务内容进行基于人工智能挖掘的审核、分类，所述新业务内容至少包括APP、公众号、微博、微信；

[0052] 网站内容监管模块102，对预设网站库内的网站内容、暗链、外链、篡改、漏洞、可用性多个维度进行检测；

[0053] 企业内容安全治理模块103，基于主动拨测、旁路检测、文件共享的内容采集技术，自动化获取文本、视频、图片、音频、复杂文档内容，并进行基于人工智能挖掘的内容审核；

[0054] UGC内容审核模块104，用于对用户原创内容进行审核，所述用户原创内容至少包括视频直播、婚恋交友、社区论坛、电商网站、和在线教育。

[0055] 以下详细介绍每个模块的具体实现方式及技术细节：

[0056] 新业务内容检测模块101中，通过工具自动获取新业务内容，并对所述新业务内容进行基于人工智能挖掘的审核、分类，如图3所示包括以下步骤：

[0057] S1、通过爬取、流量还原、识别内容主动送检中的一种或多种自动获取所述新业务内容的历史挖掘数据；爬取可以采用网络爬虫的方式，在此不再赘述。网络流量还原是一个将捕获到的流量包根据协议标准逐层分析，最终得到网络中各主机收发的数据类型和内容。tcp/ip协议簇具有完全开放化、独立与网络硬件系统、能实现网络地址统一分配和高层协议标准化的特点，适应了世界范围内数据通信的需要，可以提供多种多样可靠的网络服务。因此，对于tcp/ip流量的还原是网络流量还原的重要组成部分。

[0058] 识别内容主动送检是APP、公众号、微博、微信等运营商根据监管部门要求，通过预设送检端口将需要识别内容发送到预设服务器中，以进行进一步合规检测使用。

[0059] S2、根据合规数据库，对所述历史挖掘数据进行标注，得到标注后的历史挖掘数据集；这个步骤中的标注，可以采用人工标注的方式，也可以采用机器标注的方式进行，在此不做限制。合规数据库为预设数据库，属于新业务内容的合规数据集合，在该数据库中，可以包含不合规数据集，例如涉黄、涉暴、违禁、挂马、暗链、谩骂、违规、低俗、欺诈、灌水等内容。

[0060] S3、构建特征向量空间网络，并使用所述历史挖掘数据集输入所述特征向量空间网络进行训练，得到训练好的特征向量空间网络；这个步骤中，特征向量空间网络的构建形式可以采用多种形式，可以采用LSTM神经网络，也可以采用向量机的形式，也可以采用知识图谱网络的形式，也可以通过搭建人工智能模型而获取，在此不作限制。所述特征向量空间网络有多个节点组成，各个节点之间连接后组成空间网络。

[0061] S4、将实时获取的待挖掘数据输入所述训练好的特征向量空间网络，以判断所述待挖掘数据中是否包含违规内容。训练好的特征向量空间网络可以认为构成了一个判别器或者分类器，当新的实时获取的待挖掘数据输入所述训练好的特征向量空间网络后，通过其判断和分类功能，能够判断所述待挖掘数据中是否包含违规内容，例如，待挖掘数据的文本、视频等内容中如果含有合规数据库中的不合规数据集里的内容，那么就可以判断为包含有不合规内容，并予以提示或者报警等。

[0062] 网站内容监管模块102中，如图4所示，所述对预设网站库内的网站内容、暗链、外链、篡改、漏洞、可用性多个维度进行检测，包括：

[0063] 提取待检测网页的网页内容；对网页内容进行关键词匹配，判断网页内容中是否存在一个或多个预设关键词；若存在一个或多个预设关键词，发出匹配提示；

[0064] 提取待检测网页的网页内容；对网页内容进行暗词匹配，判断网页内容中是否存在一个或多个预设暗词；若存在一个或多个预设暗词，则将一个或多个预设暗词确定为一个或多个目标暗词，并根据一个或多个目标暗词在合法网页中出现的目标概率得到网页异常参数；若网页异常参数大于标准参数，则确定检测到暗链；

[0065] 获取待检测网站的外部链接信息以及外部链接所属网站域名信息；根据外部链接信息获取外部链接的页面内容信息，获取外部链接所属网站域名的特征信息；对外部链接的页面内容信息进行敏感内容识别，确定外部链接的页面内容的敏感内容分数；将外部链接所属网站域名的特征信息、相似度值和敏感内容分数对应的向量输入检测模型，获得外部链接的检测结果；

[0066] 收集待监测网站的网页原始数据；从所述网页原始数据中提取页面元数据特征；将所述页面元数据特征输入到训练后的神经网络中进行篡改检测；根据篡改检测结果判断待监测网站是否存在篡改；

[0067] 获取预设网站库里网站的每个网页的URL；基于每个网页的URL，确定所述预设网站库里网站的登录网页；检测所述登录网页是否存在验证码，得到第一检测结果；基于所述第一检测结果，检测所述目标网站是否存在漏洞；

[0068] 获取待检测的网站的站点IP，并建立待检测的站点IP对应的人工智能模型；提取日常Web业务访问的目的IP，判断目的IP与人工智能模型的建模IP是否一致；若目的IP与人工智能模型的建模IP一致，则对符合建模IP的Web请求信息做关键信息统计，判断关键信息是否在预设的正常范围内；若关键信息不在预设的正常范围内，该网站不可用，若关键信息

在预设的正常范围内，该网站可用。

[0069] 企业内容安全治理模块103中，基于主动拨测、旁路检测、文件共享的内容采集技术，自动化获取文本、视频、图片、音频、复杂文档内容，并进行基于人工智能挖掘的内容审核，包括：

[0070] 建立拨测任务；选择拨测目标企业；对拨测目标企业进行验证访问；执行拨测任务，在每一次拨测任务的执行周期中，对选择的拨测目标企业进行主动测试获得文本、视频、图片、音频、复杂文档内容的各项指标；提取所述指标大于预设阈值的文本、视频、图片、音频、复杂文档内容；主动拨测监控是通过遍布全球各地的节点进行分布式主动监控的工具。模拟用户访问域名、页面URL、API等，检测网络链路质量，监控网站的事务可用性，主动感知用户端应用访问体验，领先一步发现问题，降低用户流失风险。

[0071] 将旁路检测设备的检测接口设置为二层接口，并将检测接口与交换机相连；交换机通过镜像将流量上送到所述旁路检测设备上检测；在检测接口上配置旁路检测功能，使旁路检测设备只检测而不转发流量；配置安全策略，引用需要的安全配置文件，对流量进行对应的内容安全检测。例如DeviceA只有一个接口接收镜像流量或者多个接口接收镜像流量但针对各个接口接收的流量配置相同的安全策略时，可以将接口加入任何安全区域，将安全策略的源安全区域和目的安全区域配置成any。DeviceA有多个接口接收镜像流量，且要针对各个接口接收的流量配置不同的安全策略时，需要将接口加入不同的安全区域，将安全策略的源安全区域和目的安全区域配置成检测接口所在的安全区域。

[0072] 通过预设文件共享端口获取待检测企业的文本、视频、图片、音频、复杂文档内容；

[0073] 将所述文本、视频、图片、音频、复杂文档内容与基于人工智能构建的预设匹配数据库进行匹配度计算，将匹配度高于预设阈值的数据提取并给出预警提示。

[0074] UGC内容审核模块104中，对用户原创内容进行审核，包括：

[0075] 所述对用户原创内容进行审核，包括：

[0076] 文本识别：对所述用户原创内容中的文本进行语义识别、语境分析，包括以下步骤：

[0077] 基于长窗口构建语义模型，基于短窗口构建语境模型；使用用户原创内容历史语料库基于所述语义模型和语境模型训练得到所述用户原创内容的中文词向量模型；使用所述用户原创内容的中文词向量模型对实时输入的所述用户原创内容的文本进行识别并输出识别结果。

[0078] 将所述用户原创内容历史语料库任一条语料s分词后得到的语料序列为$s=\{w_1,\cdots w_{t-1},w_t,w_{t+1},\cdots w_N\}$，其中，$w_t$为分词后序列中的第t个词语，设$w_t$为待预测的目标词，$t=1,\cdots N$，N为语料序列中的总词语数目；以目标词$w_t$为中心构建窗口，定义短窗口为：

[0079] $window_s=\{w_{t\pm ds}\in s\,|\,1<d_s\leqslant\theta\}$

[0080] 其中，ds表示文本短窗口中的词语到目标词w_t的距离，设文本短窗口的距离阈值为θ，$window_s$表示由邻近目标词w_t的上下文组成的词语集合；

[0081] 定位长窗口为：

[0082] $window_l=\{w_{t\pm dl}\in s\,|\,\theta<d_l\leqslant\gamma\}$

[0083] 其中，d_l代表长窗口中的词语到目标词的距离，最小值为θ+1，最大值为β，$\beta\leqslant N$，$window_l$表示由距离目标词w_t距离较远的上下文组成，且不包括短窗口中的内容。

[0084] 上述的语义模型、语境模型、中文词向量模型可以采用本领域常用的人工智能模型，例如神经网络模型等，在此不再赘述。

[0085] 图像识别：建立图片MD5指纹库，对所述用户原创内容中的图像进行图像识别；MD5值就像是这个文件的“数字指纹”。每个文件的MD5值是不同的，如果任何人对文件做了任何改动，其MD5值也就是对应的“数字指纹”就会发生变化，比如下载服务器针对一个文件预先提供一个MD5值，用户下载完该文件后，用算法重新计算下载文件的MD5值，通过比较这两个值是否相同，就能判断下载的文件是否出错，或者说下载的文件是否被篡改了。

[0086] MD5实际上一种有损压缩技术，压缩前文件一样MD5值一定一样，反之MD5值一样并不能保证压缩前的数据是一样的，在密码学上发生这样的概率是很小的。

[0087] 图像识别具体包括以下步骤：获取所述用户原创内容中的图像的向量矩阵；对所述用户原创内容中的图像的向量矩阵进行字符串转换处理，得到所述用户原创内容中的图像的字符串；对所述用户原创内容中的图像的字符串进行加密，得到所述用户原创内容中的图像的加密字符串；对所述用户原创内容中的图像的加密字符串，与所述用户原创内容中的图像的MD5值进行拼接操作，得到所述用户原创内容中的图像的组合数据；将所述用户原创内容中的图像的组合数据，发送至云端服务器；接收云端服务器根据组合数据反馈的图像识别结果。

[0088] 视频识别：通过解码器对视频数据解码，提取视频特征，对所述用户原创内容中的视频进行识别；由于视频识别是本领域较为广泛使用的技术，本申请的视频识别步骤可以采用本领域较为公知的技术即可，在此不再赘述。

[0089] 音频识别：采用语言识别算法、关键词检索、或声纹识别对所述用户原创内容中的音频进行识别；由于音频识别是本领域较为广泛使用的技术，本申请的音频识别步骤可以采用本领域较为公知的技术即可，在此不再赘述。

[0090] 将所述文本识别、图像识别、视频识别、音频识别得到的数据输入训练好的特征向量空间网络，以判断所述数据中是否包含违规内容。这个步骤类似于新业务内容检测模块101中的实现过程，在此不再赘述。

[0091] 请参考图5，其示出了本申请的一些实施方式所提供的一种电子设备的示意图。如图5所示，所述电子设备20包括：处理器200，存储器201，总线202和通信接口203，所述处理器200、通信接口203和存储器201通过总线202连接；所述存储器201中存储有可在所述处理器200上运行的计算机程序，所述处理器200运行所述计算机程序时执行本申请前述任一实施方式所提供的基于人工智能挖掘的网络内容风控管理系统。

[0092] 其中，存储器201可能包含高速随机存取存储器(RAM:Random Access Memory)，也可能还包括非不稳定的存储器(non-volatile memory)，例如至少一个磁盘存储器。通过至少一个通信接口203(可以是有线或者无线)实现该系统网元与至少一个其他网元之间的通信连接，可以使用互联网、广域网、本地网、城域网等。

[0093] 总线202可以是ISA总线、PCI总线或EISA总线等。所述总线可以分为地址总线、数据总线、控制总线等。其中，存储器201用于存储程序，所述处理器200在接收到执行指令后，执行所述程序，前述本申请实施例任一实施方式揭示的所述基于人工智能挖掘的网络内容风控管理系统可以应用于处理器200中，或者由处理器200实现。

[0094] 处理器200可能是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述

方法的各步骤可以通过处理器200中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器200可以是通用处理器，包括中央处理器(Central Processing Unit，简称CPU)、网络处理器(Network Processor，简称NP)等；还可以是数字信号处理器(DSP)、专用集成电路(ASIC)、现成可编程门阵列(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。可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。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。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器201，处理器200读取存储器201中的信息，结合其硬件完成上述方法的步骤。

[0095] 本申请实施例提供的电子设备与本申请实施例提供的基于人工智能挖掘的网络内容风控管理系统出于相同的发明构思，具有与其采用、运行或实现的方法相同的有益效果。

[0096] 本申请实施方式还提供一种与前述实施方式所提供的基于人工智能挖掘的网络内容风控管理系统对应的计算机可读存储介质，请参考图6，其示出的计算机可读存储介质为光盘30，其上存储有计算机程序(即程序产品)，所述计算机程序在被处理器运行时，会执行前述任意实施方式所提供的基于人工智能挖掘的网络内容风控管理系统。

[0097] 需要说明的是，所述计算机可读存储介质的例子还可以包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他光学、磁性存储介质，在此不再一一赘述。

[0098] 本申请的上述实施例提供的计算机可读存储介质与本申请实施例提供的基于人工智能挖掘的网络内容风控管理系统出于相同的发明构思，具有与其存储的应用程序所采用、运行或实现的方法相同的有益效果。

[0099] 需要说明的是：

[0100] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备有固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述，构造这类系统所要求的结构是显而易见的。此外，本申请也不针对任何特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本申请的内容，并且上面对特定语言所做的描述是为了披露本申请的最佳实施方式。

[0101] 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本申请的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

[0102] 类似地，应当理解，为了精简本申请并帮助理解各个发明方面中的一个或多个，在上面对本申请的示例性实施例的描述中，本申请的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本申请要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，

遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本申请的单独实施例。

[0103] 本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0104] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本申请的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0105] 本申请的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本申请实施例的虚拟机的创建系统中的一些或者全部部件的一些或者全部功能。本申请还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者系统程序(例如，计算机程序和计算机程序产品)。这样的实现本申请的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

[0106] 应该注意的是上述实施例对本申请进行说明而不是对本申请进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本申请可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干系统的单元权利要求中，这些系统中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

[0107] 以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到其各种变化或替换，这些都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。


审核对象
互联网网站
应用程序
社交媒体
短视频
两微一端
流文件
格式文件
识别引擎
文本视频
图像识别
视频识别
音频识别
相似度识别
复杂文档提取
格式预处理
个性化分类
识别类型
涉黄
谩骂
违规
低俗
欺诈
灌水
高效治理
审核服务
自定义流程
人审、机审灵活配置
多级审核
7*24小时服务
审核标准制定

图1


新业务内容检测模块101
网站内容监管模块102
企业内容安全治理模块103
UGC内容审核模块104

图2

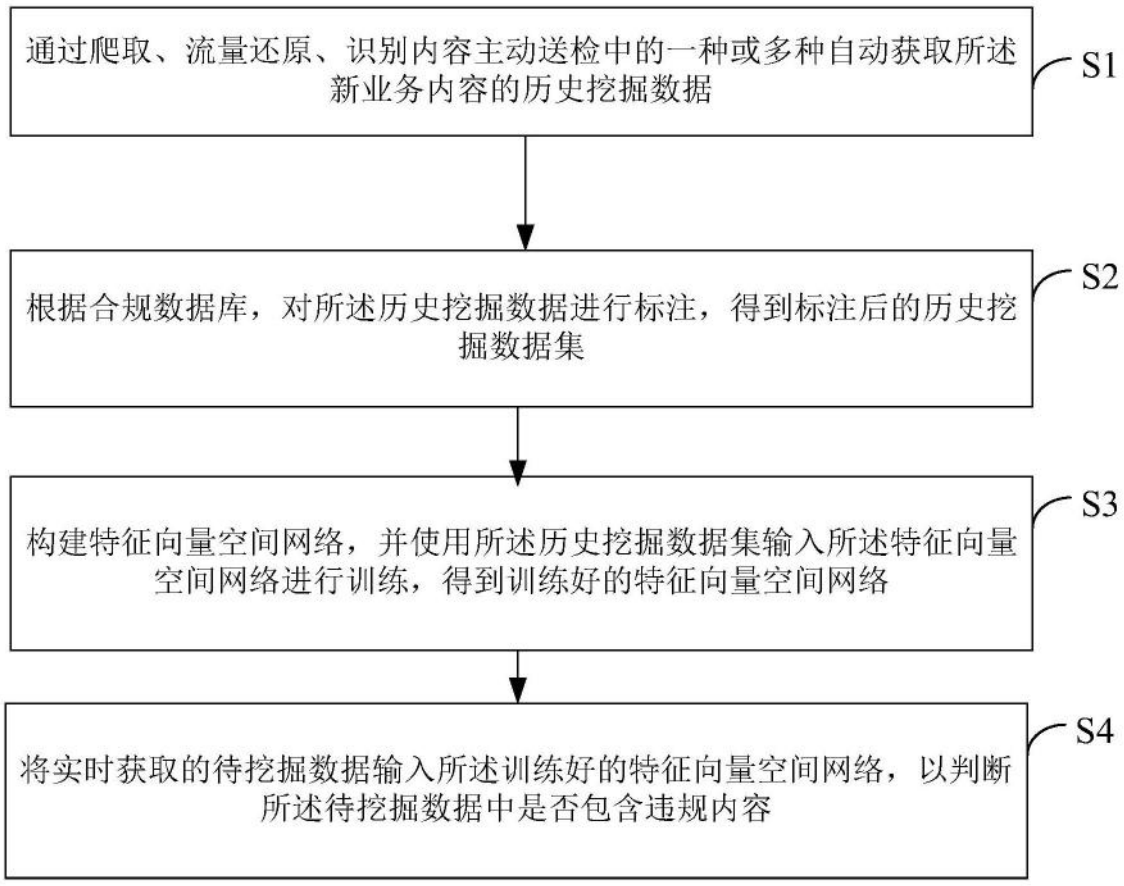
通过爬取、流量还原、识别内容主动送检中的一种或多种自动获取所述新业务内容的历史挖掘数据
S1
根据合规数据库，对所述历史挖掘数据进行标注，得到标注后的历史挖掘数据集
S2
构建特征向量空间网络，并使用所述历史挖掘数据集输入所述特征向量空间网络进行训练，得到训练好的特征向量空间网络
S3
将实时获取的待挖掘数据输入所述训练好的特征向量空间网络，以判断所述待挖掘数据中是否包含违规内容
S4

图3

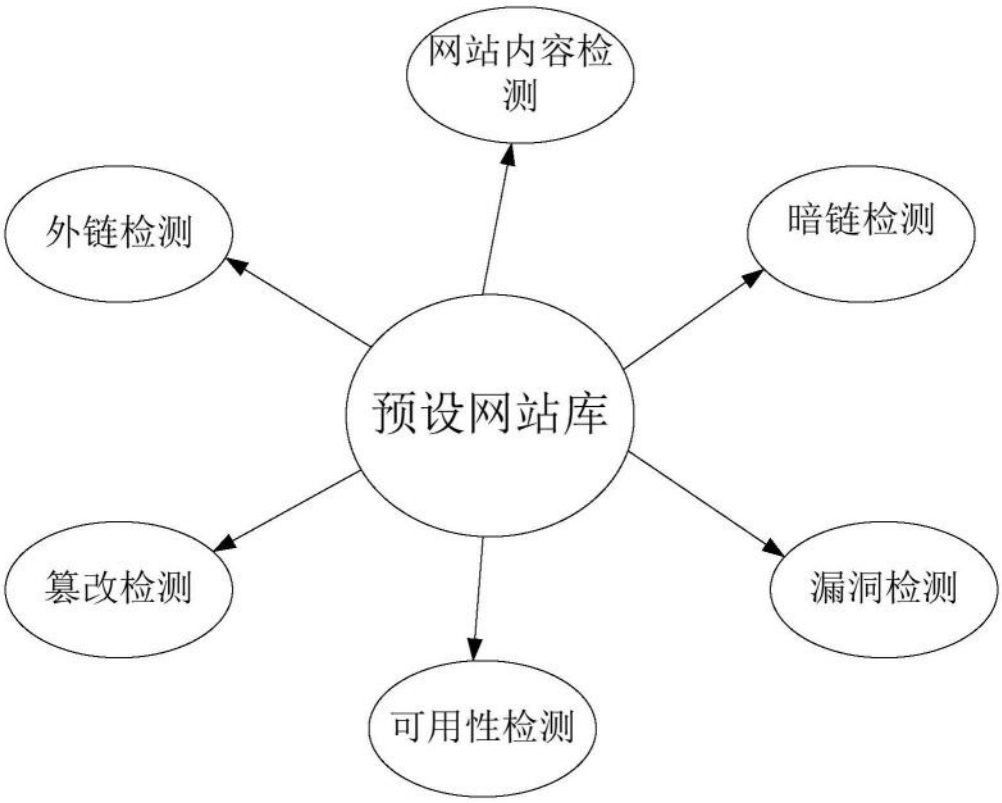
网站内容检测
外链检测
暗链检测
预设网站库
篡改检测
漏洞检测
可用性检测

图4

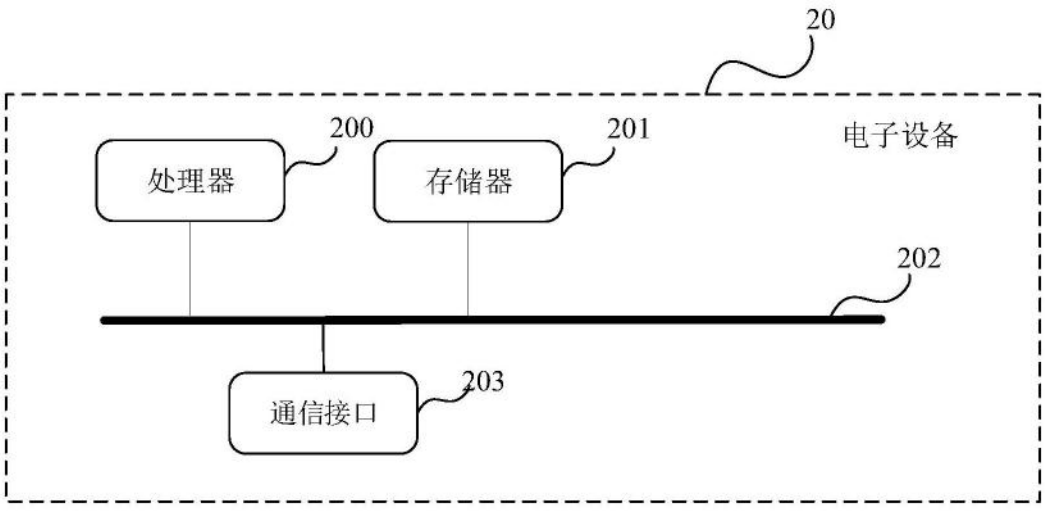
20
200
处理器
201
存储器
电子设备
202
通信接口
203

图5

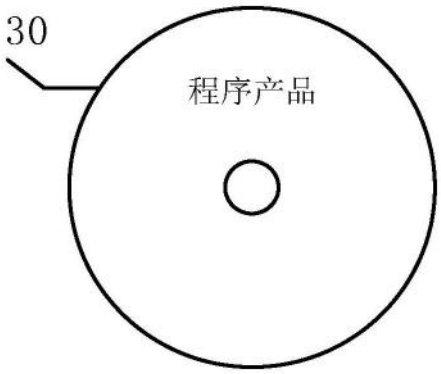

图6