(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115879110 B
(45)授权公告日 2023.07.07

(21)申请号 202310084611.8
(22)申请日 2023.02.09
(65)同一申请的已公布的文献号
申请公布号 CN 115879110 A
(43)申请公布日 2023.03.31
(73)专利权人 北京金信网银金融信息服务有限公司
地址 102699 北京市大兴区欣雅街15号院1号楼3层303
(72)发明人 许会泉
(74)专利代理机构 北京海虹嘉诚知识产权代理有限公司 11129
专利代理师 张涛
(51)Int.Cl.
G06F 21/56(2013.01)
G06F 16/958(2019.01)
审查员 刘宇儒

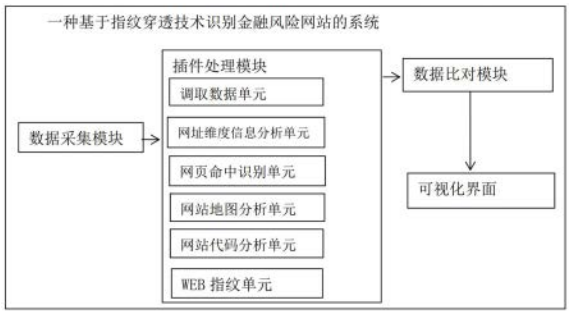
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种基于指纹穿透技术识别金融风险网站的系统

(57)摘要

本发明提供了一种基于指纹穿透技术识别金融风险网站的系统，包括数据采集模块、插件处理模块、数据比对模块、可视化界面；所述插件处理模块通过网址信息指纹、网站金融风险关系类型指纹、特征向量指纹、网站代码指纹作为该网站的WEB指纹、判断该网站是否存在金融风险；所述数据比对模块将该网站的WEB指纹与风险网站管理数据库进行比对，若问题特征一致，则认为该网站存在风险。本发明通过在多个层面构建WEB指纹，并利用了文本分类模型、相似度分析，静态检测分析、动态检测分析等多种分析方法，全面判断该网站是否存在金融风险，且该系统只需要输入目标网站的网址便可以实现对某个或某些金融风险网站的检测，检测结果清晰并可视化。



CN 115879110 B

1.一种基于指纹穿透技术识别金融风险网站的系统,其特征在于,包括:

数据采集模块:用户在系统可视化界面输入需要检测的网站的网址并传输给数据采集模块,所述数据采集模块通过爬虫抓取该网址的网页数据并进行网页解析得到目标数据,包括:制定URL,获取目标数据,并建立数据库,存储爬取的目标数据;

插件处理模块:所述插件处理模块包括:利用Websocket协议调取所述数据采集模块数据库中的目标数据的调取数据单元、通过网址维度信息分析形成网址信息指纹的网址维度信息分析单元、通过网页命中识别形成网站金融风险关系类型指纹的网页命中识别单元、通过网站地图分析形成特征向量指纹的网站地图分析单元、通过网站代码分析形成网站代码指纹的网站代码分析单元,将所述网址信息指纹、网站金融风险关系类型指纹、特征向量指纹、网站代码指纹作为该网站的WEB指纹、并判断该网站是否存在金融风险的WEB指纹单元;所述网站代码分析单元包括静态检测分析子单元和动态检测分析子单元;所述静态检测分析子单元是利用所述数据采集模块目标数据中的js源码转换为抽象语法树,并与已知的金融风险函数片段进行比较,若一致,则判断该网站存在相应的金融风险;所述动态检测分析子单元是通过事件调用链分析将所述数据采集模块目标数据中的网站代码运行在安全沙箱中,通过钩子工具集监测和触发网站功能和代码的运行,捕捉网页结构的变化所具有的网站功能,并与已知的金融风险网站所具有的功能进行相似度比对,判断该网站是否存在相应的金融风险;所述网页命中识别单元使用双向LSTM+Attention模型提取词级别上的词语特征的词提取子单元、在提取所述词语特征后,对句子特征采用Attention机制的句提取子单元;

数据比对模块:将该网站的WEB指纹与风险网站管理数据库的问题特征进行比对,若问题特征一致,则认为该网站存在风险,所述问题特征包括目录特征、网站X文件、网站X代码、网站注释特征;

可视化界面:将所述WEB指纹作为系统信息显示在可视化界面,将数据采集模块解析到的目标数据分类并显示在可视化界面,所述分类类别包括主机信息、敏感信息、网络资产、附加信息。

2.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统,其特征在于,所述网址维度信息分析单元包括对域名所有者检测的域名所有者子单元、备案信息监测的备案信息子单元、cdn检测的cdn子单元、真实IP地址检测的真实IP地址子单元。

3.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统,其特征在于,所述网页命中识别单元根据网页内容对利用文本分类模型对该网站存在的金融风险进行分类:包括:利用softmax分类获取网站的金融风险关系类型以得到网站金融风险关系类型指纹的分类子单元。

4.由权利要求1或3所述的一种基于指纹穿透技术识别金融风险网站的系统,其特征在于,所述金融风险关系类型包括:赌博、诈骗、传销。

5.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统,其特征在于,所述网站地图分析单元包括:利用数据采集模块采集到的网站链接数据进行链接去重并形成站点地图、得到站点地图特征向量并形成该网站的特征向量指纹的站点地图子单元、按照站点地图的层级、通过计算两个特征向量的余弦得到特征向量的相似度的相似度子单元。

6.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统，其特征在于，所述主机信息包括：C段、DNS服务器、IP地址、主机名、其他IP地址、域名、域名WHOIS、是否使用CDN、根域名、真实IP地址。

7.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统，其特征在于，所述敏感信息包括：其他信息、手机号、特殊数字、邮箱。

8.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统，其特征在于，所述系统信息还包括：WEB服务器、开发语言、敏感目录、端口。

9.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统，其特征在于，所述网络资产包括：同服务器域名、备案者持有域名、子域名、注册者持有域名。

10.由权利要求1所述的一种基于指纹穿透技术识别金融风险网站的系统，其特征在于，所述风险网站管理数据库是定期从互联网渠道采集关于传销、赌博、诈骗广告的网站或新闻报道披露的风险网站并解析成WEB指纹，通过机器学习形成风险网站的问题特征并储存在风险网站管理数据库。

一种基于指纹穿透技术识别金融风险网站的系统

技术领域

[0001] 本发明涉及网站数据分析领域,特别涉及一种基于指纹穿透技术识别金融风险网站的系统。

背景技术

[0002] 公开号CN110796542A的专利“金融风险控制方法、金融风险控制装置和电子设备”,主要通过获取历史用户的APP下载序列信息和金融行为信息来创建用户风险控制模型,使用历史用户的所述APP下载序列向量数据和金融行为信息训练所述用户风险控制模型;使用所述用户风险控制模型计算所述目标用户的金融风险预测值。即现有技术是从用户自身的异常行为的角度监控金融风险,很可能是金融风险已经发生,用户财产已经造成了损失,不能及时的预测风险和避免风险。

发明内容

[0003] 为解决现有技术不能提供一套完善的针对金融风险网站的识别方法、进而加强对金融风险网站的监管的问题,本发明提供了一种基于指纹穿透技术识别金融风险网站的系统,通过在用户界面输入检测的目标网址,利用网址信息指纹、网站金融风险关系类型指纹、特征向量指纹、网站代码指纹作为该网站的WEB指纹,自动判断该网站是否存在金融风险。

[0004] 一种基于指纹穿透技术识别金融风险网站的系统,包括:

[0005] 数据采集模块:用户在系统可视化界面输入需要检测的网站的网址并传输给数据采集模块,所述数据采集模块通过爬虫抓取网页数据并进行网页解析得到目标数据,包括:制定URL,获取目标数据,并建立数据库,存储爬取的目标数据;

[0006] 插件处理模块:所述插件处理模块包括:利用Websocket协议调取所述数据采集模块数据库中的数据的调取数据单元、通过网址维度信息分析形成网址信息指纹的网址维度信息分析单元、通过网页命中识别形成网站金融风险关系类型指纹的网页命中识别单元、通过网站地图分析形成特征向量指纹的网站地图分析单元、通过网站代码分析形成网站代码指纹的网站代码分析单元,将所述网址信息指纹、网站金融风险关系类型指纹、特征向量指纹、网站代码指纹作为该网站的WEB指纹、并判断该网站是否存在金融风险的WEB指纹单元;

[0007] 数据比对模块:将该网站的WEB指纹与风险网站管理数据库进行比对,若问题特征一致,则认为该网站存在风险,所述问题特征包括目录特征、网站X文件、网站X代码、网站注释特征;

[0008] 可视化界面:将所述WEB指纹作为系统信息显示在可视化界面,将数据采集模块解析到的目标数据分类并显示在可视化界面,所述分类类别包括主机信息、敏感信息、网络资产、附加信息。

[0009] 优选地,所述网址维度信息分析单元包括对域名所有者检测的域名所有者子单

元、备案信息监测的备案信息子单元、cdn检测的cdn子单元、真实IP地址检测的真实IP地址子单元。

[0010] 优选地，所述网页命中识别单元根据网页内容对利用文本分类模型对该网站存在的金融风险进行分类：包括：使用双向LSTM+Attention模型提取词级别上的词语特征的词提取子单元、在提取所述词语特征后，对句子特征采用Attention机制的句提取子单元、利用softmax分类获取网站的金融风险关系类型以得到网站金融风险关系类型指纹的分类子单元。

[0011] 优选地，所述金融风险关系类型包括：赌博、诈骗、传销。

[0012] 优选地，所述网站地图分析单元包括：利用数据采集模块采集到的网站链接数据进行链接去重并形成站点地图、得到站点地图特征向量并形成该网站的特征向量指纹的站点地图子单元、按照站点地图的层级、通过计算两个特征向量的余弦得到特征向量的相似度的相似度子单元。

[0013] 优选地，所述网站代码分析单元包括静态检测分析子单元和动态检测分析子单元。

[0014] 优选地，所述静态检测分析子单元是利用所述数据采集模块目标数据中的js源码转换为抽象语法树，并与已知的金融风险函数片段进行比较，若一致，则判断该网站存在相应的金融风险。

[0015] 优选地，所述动态检测分析子单元是通过事件调用链分析将所述数据采集模块目标数据中的网站代码运行在安全沙箱中，通过钩子工具集监测和触发网站功能和代码的运行，捕捉网页结构的变化所具有的网站功能，并与已知的金融风险网站所具有的功能进行相似度比对，判断该网站是否存在相应的金融风险。

[0016] 优选地，所述主机信息包括：C段、DNS服务器、IP地址、主机名、其他IP地址、域名、域名WHOIS、是否使用CDN、根域名、真实IP地址。

[0017] 优选地，所述敏感信息包括：其他信息、手机号、特殊数字、邮箱。

[0018] 优选地，所述系统信息还包括：WEB服务器、开发语言、敏感目录、端口。

[0019] 优选地，所述网络资产包括：同服务器域名、备案者持有域名、子域名、注册者持有域名。

[0020] 优选地，所述风险网站管理数据库是定期从互联网渠道采集关于传销、赌博、诈骗广告的网站或新闻报道披露的风险网站并解析成WEB指纹，通过机器学习形成风险网站的问题特征并储存在风险网站管理数据库。

[0021] 本发明提供了一种基于指纹穿透技术识别金融风险网站的系统，包括数据采集模块、插件处理模块、数据比对模块、可视化界面；所述插件处理模块通过网址信息指纹、网站金融风险关系类型指纹、特征向量指纹、网站代码指纹作为该网站的WEB指纹、判断该网站是否存在金融风险；所述数据比对模块将该网站的WEB指纹与风险网站管理数据库进行比对，若问题特征一致，则认为该网站存在风险。本发明通过在多个层面构建WEB指纹，并利用了文本分类模型、相似度分析，静态检测分析、动态检测分析等多种分析方法，全面判断该网站是否存在金融风险，且该系统只需要输入目标网站的网址便可以实现对某个或某些金融风险网站的检测，检测结果清晰并可视化，更有效的实现金融风险网站的金融监管。不需要过多的现场摸排，线上多渠道搜索信息的方式，就可以初步定位出存在金融风险的清晰

画像,为金融监管部门打击传网络非法金融活动提供了一种快捷且精准的系统。

附图说明

[0022] 图1一种基于指纹穿透技术识别金融风险网站的系统。

[0023] 图2在一种基于指纹穿透技术识别金融风险网站的系统中输入被检测网站网址的可视化界面。

[0024] 图3在一种基于指纹穿透技术识别金融风险网站的系统中显示金融风险网站识别结果的可视化界面。

具体实施方式

[0025] 下面结合附图和实施例对本发明做进一步说明。

[0026] 如图1所示,一种基于指纹穿透技术识别金融风险网站的系统,包括:

[0027] 数据采集模块:如图2所示,用户在系统可视化界面输入需要检测的网站的网址并传输给数据采集模块,所述数据采集模块通过爬虫抓取网页数据并进行网页解析得到目标数据,包括:制定URL,获取目标数据,并建立数据库,存储爬取的目标数据;

[0028] 插件处理模块:所述插件处理模块包括:利用Websocket协议调取所述数据采集模块数据库中的数据的调取数据单元、通过网址维度信息分析形成网址信息指纹的网址维度信息分析单元、通过网页命中识别形成网站金融风险关系类型指纹的网页命中识别单元、通过网站地图分析形成特征向量指纹的网站地图分析单元、通过网站代码分析形成网站代码指纹的网站代码分析单元,将所述网址信息指纹、网站金融风险关系类型指纹、特征向量指纹、网站代码指纹作为该网站的WEB指纹、并判断该网站是否存在金融风险的WEB指纹单元;

[0029] 数据比对模块:将该网站的WEB指纹与风险网站管理数据库进行比对,若问题特征一致,则认为该网站存在风险,所述问题特征包括目录特征、网站X文件、网站X代码、网站注释特征;

[0030] 可视化界面:将所述WEB指纹作为系统信息显示在可视化界面,将数据采集模块解析到的目标数据分类并显示在可视化界面,所述分类类别包括主机信息、敏感信息、网络资产、附加信息。

[0031] 优选地,所述网址维度信息分析单元包括对域名所有者检测的域名所有者子单元、备案信息监测的备案信息子单元、cdn检测的cdn子单元、真实IP地址检测的真实IP地址子单元。

[0032] 优选地,所述网页命中识别单元根据网页内容对利用文本分类模型对该网站存在的金融风险进行分类:包括:使用双向LSTM+Attention模型提取词级别上的词语特征的词提取子单元、在提取所述词语特征后,对句子特征采用Attention机制的句提取子单元、利用softmax分类获取网站的金融风险关系类型以得到网站金融风险关系类型指纹的分类子单元。

[0033] 优选地,所述文本分类模型还可以使用Fasttext模型、TextCNN、TextRNN等模型,所述Fasttext模型包括输入层、隐含层、输出层共三层。输入词向量,输出label,隐含层是对多个词向量的叠加平均。CBOW的输入是目标单词的上下文,Fasttext的输入是多个单词

及其n-gram特征,这些单词用来表示单个文档CBOW的输入单词使用one-hot编码,Fasttext的输入特征时使用embedding编码;CBOW的输出是目标词汇,Fasttext的输出是文档对应的类别。所述TextCNN只有一层convolution,一层max-pooling, 最后将输出外接softmax来n分类。所述TextRNN一般取前向/反向LSTM在最后一个时间步长上隐藏状态,然后进行拼接,在经过一个softmax层进行一个多分类;或者取前向/反向LSTM在每一个时间步长上的隐藏状态,对每一个时间步长上的两个隐藏状态进行拼接concat,然后对所有时间步长上拼接后的隐藏状态取均值,再经过softmax层进行分类。

[0034] 优选地,所述金融风险关系类型包括:赌博、诈骗、传销。

[0035] 优选地,所述网站地图分析单元包括:利用数据采集模块采集到的网站链接数据进行链接去重并形成站点地图、得到站点地图特征向量并形成该网站的特征向量指纹的站点地图子单元、按照站点地图的层级、通过计算两个特征向量的余弦得到特征向量的相似度的相似度子单元。

[0036] 优选地,所述网站代码分析单元包括静态检测分析子单元和动态检测分析子单元。

[0037] 优选地,所述静态检测分析子单元是利用所述数据采集模块目标数据中的js源码转换为抽象语法树,并与已知的金融风险函数片段进行比较,若一致,则判断该网站存在相应的金融风险。

[0038] 优选地,所述动态检测分析子单元是通过事件调用链分析将所述数据采集模块目标数据中的网站代码运行在安全沙箱中,通过钩子工具集监测和触发网站功能和代码的运行,捕捉网页结构的变化所具有的网站功能,并与已知的金融风险网站所具有的功能进行相似度比对,判断该网站是否存在相应的金融风险。

[0039] 如图3所示,在所述系统可视化界面显示金融风险的识别结果。

[0040] 优选地,所述主机信息包括:C段、DNS服务器、IP地址、主机名、其他IP地址、域名、域名WHOIS、是否使用CDN、根域名、真实IP地址。

[0041] 优选地,所述敏感信息包括:其他信息、手机号、特殊数字、邮箱。

[0042] 优选地,所述系统信息还包括:WEB服务器、开发语言、敏感目录、端口。

[0043] 优选地,所述网络资产包括:同服务器域名、备案者持有域名、子域名、注册者持有域名。

[0044] 优选地,所述风险网站管理数据库是定期从互联网渠道采集关于传销、赌博、诈骗广告的网站或新闻报道披露的风险网站并解析成WEB指纹,通过机器学习形成风险网站的问题特征指标并储存在风险网站管理数据库。

[0045] 应当指出,以上所述具体实施方式可以使本领域的技术人员更全面地理解本发明创造,但不以任何方式限制本发明创造。因此,尽管本说明书参照附图和实施例对本发明创造已进行了详细的说明,但是,本领域技术人员应当理解,仍然可以对本发明创造进行修改或者等同替换,总之,一切不脱离本发明创造的精神和范围的技术方案及其改进,其均应涵盖在本发明创造专利的保护范围当中。

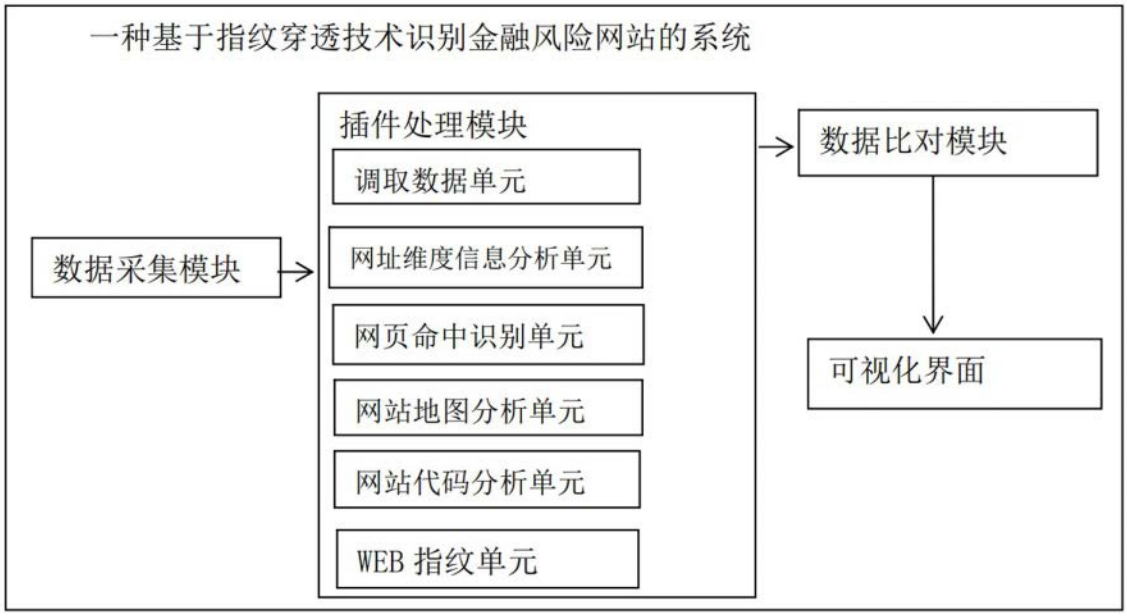

一种基于指纹穿透技术识别金融风险网站的系统
插件处理模块
调取数据单元
网址维度信息分析单元
网页命中识别单元
网站地图分析单元
网站代码分析单元
WEB 指纹单元
数据采集模块
数据比对模块
可视化界面

图 1

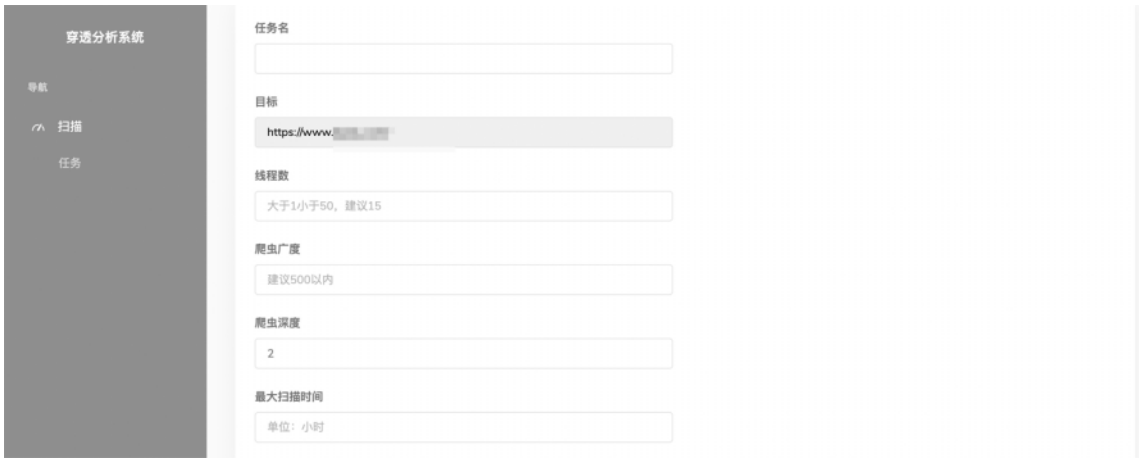

穿透分析系统
导航
扫描
任务
任务名
目标
https://www.
线程数
大于1小于50，建议15
爬虫广度
建议500以内
爬虫深度
2
最大扫描时间
单位：小时

图 2


穿透分析系统
导航
扫描
任务
穿透分析系统
导航
扫描
任务
扫描
任务
任务名
https://www
目标
https://www
基本信息
主机信息
C段
DNS服务器
IP地址
47.93.213.65
主机名
其他IP地址
域名
域名WHOIS
是否使用CDN
是否使用WAF
根域名
真实IP地址
敏感信息
其他信息
手机号
特殊数字
邮箱
系统信息
WEB指纹
WEB服务器
开发语言
敏感目录
端口
网络资产
同服务器域名
备案者持有域名
子域名
注册者持有域名

图 3