(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110929525 B
(45)授权公告日 2022.08.05

(21)申请号 201911012231.3
(22)申请日 2019.10.23
(65)同一申请的已公布的文献号
申请公布号 CN 110929525 A
(43)申请公布日 2020.03.27
(73)专利权人 三明学院
地址 365000 福建省三明市三元区荆东路25号
(72)发明人 余建 林志兴
(74)专利代理机构 厦门智慧呈睿知识产权代理事务所(普通合伙) 35222
专利代理师 陈槐萱
(51)Int.Cl.
G06F 40/30(2020.01)
G06K 9/62(2022.01)
G06Q 40/02(2012.01)

(56)对比文件
CN 107622443 A,2018.01.23
CN 109658222 A,2019.04.19
KR 101999213 B1,2019.07.11
谭天骄 等.P2P网络借贷平台风险预警研究.《金融与经济》.2019,第77-83页.
Maoguang Wang等.Research on Financial Network Loan Risk Control Model based on Prior Rule and Machine Learning Algorithm.《ICMAI 2019》.2019,第76-79页.

审查员 林菁

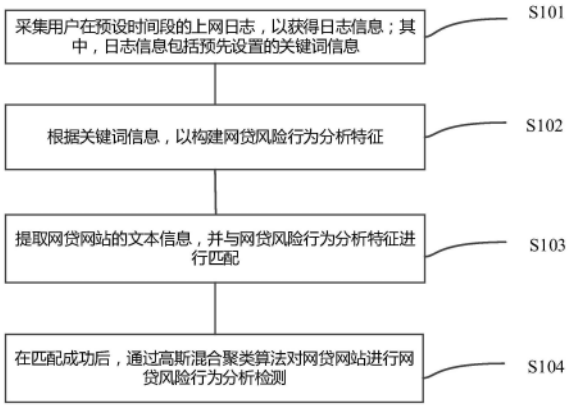
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种网贷风险行为分析检测方法、装置、设备和存储介质

(57)摘要

本发明公开了一种网贷风险行为分析检测方法、装置、设备及计算机存储介质，方法包括：采集用户在预设时间段的上网日志，以获得日志信息；其中，所述日志信息包括预先设置的关键词信息；根据所述关键词信息，以构建网贷风险行为分析特征；提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配；在匹配成功后，通过高斯混合聚类算法对所述网贷网站进行网贷风险行为分析检测。本发明根据多维度挖掘恶意访问的表现特征，结合高斯混合聚类算法对网贷行为做出分析判断，提高了识别精度以及效率。



CN 110929525 B

1.一种网贷风险行为分析检测方法,其特征在于,包括:

采集用户在预设时间段的上网日志,以获得日志信息;其中,所述日志信息包括预先设置关键词的关键词信息;

根据所述关键词信息,以构建网贷风险行为分析特征;

提取网贷网站的文本信息,并与所述网贷风险行为分析特征进行匹配;

在匹配成功后,通过高斯混合聚类算法对所述网贷网站进行网贷风险行为分析检测:具体为:

在匹配成功后,对所述匹配后的文本信息进行距离度量以及性能指标测量;

将距离度量以及性能指标测量后的文本信息,基于高斯混合聚类算法,检测所述网贷网站进行网贷风险行为;

采用VDM对匹配后的文本信息进行距离度量,距离度量表达式为:$VDM_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$;$m_{u,a}$为在属性u 上取值为a的文本样本数,$m_{u,a,i}$为在第i个样本簇中属性u 上取值为a的样本数,k为文本特征样本簇数,$VDM_p(a,b)$为VDM度量距离。

2.根据权利要求1所述的网贷风险行为分析检测方法,其特征在于,在所述根据所述关键词信息,以构建网贷风险行为分析特征的步骤之后,在所述提取网贷网站的文本信息,并与所述网贷风险行为分析特征进行匹配的步骤之前,还包括:

基于深度包检测的应用识别算法,识别所有网站的应用标签类型;

根据所述网站的应用标签类型,将网站进行区分分类,以获取分类后的网贷网站。

3.根据权利要求1所述的网贷风险行为分析检测方法,其特征在于,提取网贷网站的文本信息,并与所述网贷风险行为分析特征进行匹配,具体为:

通过多变潜在语义索引文本敏感特征抽取算法对网贷网站的文本信息进行提取,并将提取后的文本信息与所述网贷风险行为分析特征进行匹配。

4.一种网贷风险行为分析检测装置,其特征在于,包括:

采集单元,用于采集用户在预设时间段的上网日志,以获得日志信息;其中,所述日志信息包括预先设置关键词的关键词信息;

构建单元,用于根据所述关键词信息,以构建网贷风险行为分析特征;

提取单元,用于提取网贷网站的文本信息,并与所述网贷风险行为分析特征进行匹配;

检测单元,用于在匹配成功后,通过高斯混合聚类算法对所述网贷网站进行网贷风险行为分析检测:具体为:

在匹配成功后,对所述匹配后的文本信息进行距离度量以及性能指标测量;

将距离度量以及性能指标测量后的文本信息,基于高斯混合聚类算法,检测所述网贷网站进行网贷风险行为;

采用VDM对匹配后的文本信息进行距离度量,距离度量表达式为:$VDM_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$;$m_{u,a}$为在属性u 上取值为a的文本样本数,$m_{u,a,i}$为在第i个样本簇中属性u 上取值为a的样本数,k为文本特征样本簇数,$VDM_p(a,b)$为VDM度量距离。

5.根据权利要求4所述的网贷风险行为分析检测装置,其特征在于,

识别单元，用于基于深度包检测的应用识别算法，识别所有网站的应用标签类型；

区分分类单元，用于根据所述网站的应用标签类型，将网站进行区分分类，以获取分类后的网贷网站。

6.一种网贷风险行为分析检测设备，包括处理器、存储器以及存储在所述存储器中且被配置由所述处理器 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3任一项所述网贷风险行为分析检测方法。

7.一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3中任意一项所述的网贷风险行为分析检测方法。

一种网贷风险行为分析检测方法、装置、设备和存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种网贷风险行为分析检测方法、装置、设备和存储介质。

背景技术

[0002] 目前，随着当前网贷平台的技术越来越成熟，以及各类网贷网站的急剧增加，导致依靠人工来评估网贷行为产生的风险不再有效。因此，出现了各种基于行为的校园贷风险行为分析检测技术，例如通过建立SVM、Logit、判别分析模型来识别网贷问题平台，并利用比较问题平台和正常平台的各项识别指标的均值来判读正常平台与问题平台；通过机器语言算法先得出一套平台风险的最优指标组合，利用所选的变量对其指标进行因子分析并得到其指标值，然后对多家的平台按指标分配后得到综合得分并进行评价排序，得到排名最前的50家网贷平台，最后根据模型建立的平台风险评价体系进行风险预测。但是上述方法中通过对比分析法及指标分配法对网贷网站进行识别，其识别精度、效率相对低且智能化水平较低。

发明内容

[0003] 针对上述问题，本发明的目的在于提供一种网贷风险行为分析检测方法、装置、设备和存储介质，本发明根据多维度挖掘恶意访问的表现特征，结合高斯混合聚类算法对网贷行为做出分析判断，提高了识别精度以及效率。

[0004] 本发明第一方面提供了一种网贷风险行为分析检测方法，包括：

[0005] 采集用户在预设时间段的上网日志，以获得日志信息；其中，所述日志信息包括预先设置关键词的关键词信息；

[0006] 根据所述关键词信息，以构建网贷风险行为分析特征；

[0007] 提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配；

[0008] 在匹配成功后，通过高斯混合聚类算法对所述网贷网站进行网贷风险行为分析检测。

[0009] 优选地，在所述根据所述关键词信息，以构建网贷风险行为分析特征的步骤之后，在所述提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配的步骤之前，还包括：

[0010] 基于深度包检测的应用识别算法，识别所有网站的应用标签类型；

[0011] 根据所述网站的应用标签类型，将网站进行区分分类，以获取分类后的网贷网站。

[0012] 优选地，提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配，具体为：

[0013] 通过多变潜在语义索引文本敏感特征抽取算法对网贷网站的文本信息进行提取，并将提取后的文本信息与所述网贷风险行为分析特征进行匹配。

[0014] 优选地，所述在匹配成功后，通过高斯混合聚类算法对所述网贷网站进行网贷风

险行为分析检测，具体为：

[0015] 在匹配成功后，对所述匹配后的文本信息进行距离度量以及性能指标测量；将距离度量以及性能指标测量后的文本信息，基于高斯混合聚类算法，检测所述网贷网站进行网贷风险行为。

[0016] 优选地，采用VDM对匹配后的文本信息进行距离度量，距离度量表达式为：

$$VDM_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$$

；$m_{u,a}$为在属性μ上取值为a的文本样本数，$m_{u,a,i}$为在第i个样本簇中属性μ上取值为a的样本数，k为文本特征样本簇数，$VDM_p(a,b)$为 VDM度量距离。

[0017] 本发明实施例还提供了一种网贷风险行为分析检测装置，包括：

[0018] 采集单元，用于采集用户在预设时间段的上网日志，以获得日志信息；其中，所述日志信息包括预先设置的关键词信息；

[0019] 构建单元，用于根据所述关键词信息，以构建网贷风险行为分析特征；

[0020] 提取单元，用于提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配；

[0021] 检测单元，用于在匹配成功后，通过高斯混合聚类算法对所述网贷网站进行网贷风险行为分析检测。

[0022] 优选地，还包括：

[0023] 识别单元，用于基于深度包检测的应用识别算法，识别所有网站的应用标签类型；

[0024] 区分分类单元，用于根据所述网站的应用标签类型，将网站进行区分分类，以获取分类后的网贷网站。

[0025] 提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配，具体为：

[0026] 通过多变潜在语义索引文本敏感特征抽取算法对网贷网站的文本信息进行提取，并将提取后的文本信息与所述网贷风险行为分析特征进行匹配。

[0027] 优选地，检测单元，具体包括：

[0028] 距离度量以及性能指标测量模块，用于在匹配成功后，对所述匹配后的文本信息进行距离度量以及性能指标测量；

[0029] 网贷风险行为检测模块，用于将距离度量以及性能指标测量后的文本信息，基于高斯混合聚类算法，检测所述网贷网站进行网贷风险行为。

[0030] 优选地，采用VDM对匹配后的文本信息进行距离度量，距离度量表达式为：

$$VDM_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$$

；$m_{u,a}$为在属性μ上取值为a的文本样本数，$m_{u,a,i}$为在第i个样本簇中属性μ上取值为a的样本数，k为文本特征样本簇数，$VDM_p(a,b)$为μ上的a和b两个离散值之间的VDM度量距离。

[0031] 本发明第三方面还提供了一种网贷风险行为分析检测设备，包括处理器、存储器以及存储在所述存储器内的计算机程序，所述计算机程序能够被所述处理器执行以实现上述实施例所述的网贷风险行为分析检测方法。

[0032] 本发明第四方面还提供了一种计算机可读存储介质，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

设备执行如上述实施例所述的网贷风险行为分析检测方法。

[0033] 实施本发明实施例,具有如下有益技术效果:

[0034] 本发明以用户在预设时间段的上网日志,获得包括预先设置关键词的关键词信息,构建构建网贷风险行为分析特征,提取网贷网站的文本信息,并与所述网贷风险行为分析特征进行匹配,然后结合高斯混合聚类算法对网贷行为做出分析判断,提高了识别精度以及效率。

附图说明

[0035] 为了更清楚地说明本发明的技术方案,下面将对实施方式中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施方式,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0036] 图1是本发明第一实施例提供的一种网贷风险行为分析检测方法的流程示意图。

[0037] 图2是本发明实施例提供的某高校的校园网络出口部署拓扑图。

[0038] 图3是本发明第二实施例提供的一种网贷风险行为分析检测装置的结构示意图。

具体实施方式

[0039] 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做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。

[0040] 请参阅图1至图2,本发明第一实施例提供了一种网贷风险行为分析检测方法,其可由网贷风险行为分析检测设备(以下简称“分析检测设备”)来执行,特别的,由网贷风险行为分析检测设备内的一个或多个处理器来执行,并至少包括如下步骤:

[0041] S101,采集用户在预设时间段的上网日志,以获得日志信息;其中,所述日志信息包括预先设置关键词的关键词信息。

[0042] 在本实施例中,所述关键词为用户搜索网贷敏感词,包括网贷敏感词特征信息以及网贷标题敏感词特征信息,其中,所述网贷敏感词特征信息包括网贷、贷款、信贷、借贷、借钱、信用贷等。所述网贷标题敏感词特征信息包括账单、订单、取现、充值、还款、个人中心、会员注册、签约、完善资料、资金、提现、交易、申请成功以及忘记密码等。

[0043] S102,根据所述关键词信息,以构建网贷风险行为分析特征。

[0044] S103,提取网贷网站的文本信息,并与所述网贷风险行为分析特征进行匹配。

[0045] S104,在匹配成功后,通过高斯混合聚类算法对所述网贷网站进行网贷风险行为分析检测。

[0046] 在本实施例中,通过多变潜在语义索引文本敏感特征抽取算法对网贷网站的文本信息进行提取,并将提取后的文本信息与所述网贷风险行为分析特征进行匹配,具体地,由于网贷网站的网站栏目中,一般都对应着一个“我要借款”“我要出借”等信息,可以用该信息作为标注,从而判断其网站类型,由于网贷网站所包含的文本信息可以被提取,其网站栏目的“敏感词”就是文本信息,根据搜索网贷敏感词特征信息以及网贷标题敏感词特征信息,用tag中的文本特征标注构成一个样本集$D=\{x_1,x_2,\cdots,x_m\}$。通过多变潜在语义索引文

本敏感特征抽取算法 $p(s)=\frac{\sum_A\sum_W\sum z p(s,\alpha,w)\,|\,p(z\,|\,a,w,s)}{\sum_S\sum_A\sum_W\sum z p(s,\alpha,w)\,|\,p(z\,|\,a,w,s)}$，构建对网贷网站敏感文字分析网站类别的特征。其中，网贷网站文本特征抽取算法如下：输入：网站网站文本集合 D＝{x_1,x_2,…,x_m}；敏感文本标注集合A＝{a_1,a_2,...a_i}。输出：抽取网贷网站的文本特征集合F＝{F_1,F_2,...F_n}。1、begin 2、网站文本预识别；3、建立语义索引文本多变参 $p(s)=\frac{\sum_A\sum_W\sum z p(s,\alpha,w)\,|\,p(z\,|\,a,w,s)}{\sum_S\sum_A\sum_W\sum z p(s,\alpha,w)\,|\,p(z\,|\,a,w,s)}$；4、设置隐主题z的个数k；5、If(ε≥σ)6、将索引文本E-M迭代求参；7、else；8、end if；9、For(i＝0,i≤n,i++)； 10、网贷敏感文本特征抽取；11、结合条件概率，生成隐主题集合Z；12、输出网贷敏感文本特征集合F；13、End。

[0047] 在本实施例中，在匹配成功后，对所述匹配后的文本信息进行距离度量以及性能指标测量；将距离度量以及性能指标测量后的文本信息，基于高斯混合聚类算法，检测所述网贷网站进行网贷风险行为。具体地，根据对网贷网站文本特征抽取，知道其属性可以划分为“离散属性”。设网贷网站的特征定义域为{我要借钱，我要出借、金融平台、贷款、借钱、信用贷…等}，通过计算每个特征对分类的距离大小来判断特征“相似度度量”，距离越大，相似度越小，反之，相似度越大。对网贷网站中的特征信息这种无序属性可采用VDM(Value Difference Metric)来度量。

[0048] 令$m_{u,a}$表示在属性μ上取值为a的文本样本数，$m_{u,a,i}$表示在第i个样本簇中属性μ上取值为a的样本数，k为文本特征样本簇数，那么μ上的a和b两个离散值之间的VDM度量距离为$VDM_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$。

[0049] 在本实施例中，网贷网站中包含的文本信息相对固定且包含的文本经常使用大量的金融名词信息，通常有用户自己选择或者网站所打的标签，同时该网站会含有大量的会员注册信息。由于这些标签在一定程度上代表着文本的网站类别，把这些文本看成观测集D，并对其进行簇划分，具体如下：给定文本观测集D＝{x_1,x_2,…,x_m}，K-均值算法针对聚类所得簇划分C＝{C_1,C_2,…C_k}最小化平方误差 $E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|_2^2$ （4）；其中 $\mu_i=\frac{1}{|C_i|}$，x是簇C_i的均值向量。E在一定程度上刻画了簇内样本围绕簇均值向量的紧密程度，E值越小则簇样本相似度越高。

[0050] 其中，在对网站文本和标签识别过程中，由于网站中会包含大量图像信息，并不能保证采集到的图像中文本信息，对于传统的K-均值聚类算法往往因参数选择的不合理而导致算法收敛速度慢，检测效果不理想。为此，给出一种改进的高斯混合聚类模型检测方法对其做了进一步改进和优化。

[0051] 对于一个网站来说，文本的特征在一段时间内都不会发生变化，那么可以确定在一段时间内，文本特征服从高斯分布。为了得到高斯混合模型的初始参数值，可以选择图像文本作为训练序列，将高斯混合算法在线训练序列中对其特定的文本的信息值进行聚类，同时更新对应的特征文本向量均值、方差值和样本值。最后根据每个对应于每个文本的聚类的数量确定构造文本类型所需的高斯分布的数量。通过对应于每个聚类的文本向量的均

值、方差和样本数来初始化相应高斯混合分布的权重值和方差等。

[0052] 具体地，对n维样本空间x中的随机向量x若x服从高斯分布，其概率密度函数为 $p(x)=\frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ （5）；其中μ表示n维均值向量，Σ表示n×n 的协方差矩阵。由式(5)可以看出，μ和Σ这两个参数决定了高斯分布的概率。将高斯混合分布定义为 $p(x)=\sum_{i=1}^{k}\alpha_i\cdot p(x|\mu_i,\Sigma_i)$ （6）；在式(6)中p(x)表示有k 个混合成分组成，每个文本混合成分对应一个高斯分布。而μ_i，Σ_i表示第i个高斯混合成分的参数。p(x|μ，Σ)表示概率密度函数。

[0053] 如果新的文本样本的生成过程满足高斯混合分布：定义文本特征α_1，α_2，…，α_k的先验分布符合高斯混合成分，α_i为第i个文本混合成分的先验概率。令特征文本样本观测集D＝{x_1，x_2，…，x_m}，其随机变量$z_j\in\{1,2...,k\}$表示产生新样本x_j的高斯混合成分，且为未知变量。同时，z_j的先验概率p(z_j＝i)对应于α_i(i＝1，2，...，k)。最后z_j的后验分布式为：$p(z_j=i|x_j)=\frac{p(z_j=i)\cdot p(x_j|z_j=i)}{p(x_j)}=\frac{\alpha_i\cdot p(x_j|\mu_i,\Sigma_i)}{\sum_{i=1}^{k}\alpha_l\cdot p(x_j||\mu_l,\Sigma_l)}$ （7）；其中，p(z_j＝i|x_j)中的x_j表示为第i个高斯混合成分生成的后验概率。可将其简化为γ_{ji}(i＝1，2，...，k)。根据式(6)，把样本集D划分为k个簇C＝{C_1，C_2，…C_k}，每个样本x_j的簇标记λ_j：λ_j＝argmaxγ_{ji}(i∈{1，2，...k)（8）；针对网贷敏感文本样本集D，采用极大似然估计，即 $LL(D)=\ln(\prod_{j=1}^{m}p(x_j))=\sum_{j=1}^{m}\ln(\sum_{i=1}^{m}\alpha_i\cdot p(x_j|\mu_i,\Sigma_i))$ (9)；将EM算法进行迭代优化求解，若参数{(a_i，μ_i，Σ_i|1≤i≤k}能使式(9)最大化，则由$\frac{\partial LL(D)}{\partial\mu_i}=0$有

$\sum_{j=1}^{m}\frac{\alpha_i\cdot p(x_j|\mu_i,\Sigma_i)}{\alpha_l\cdot p(x_j||\mu_l,\Sigma_l)}(x_j-\mu_i)=0$ （10）；由式(7)以及γ_{ji}＝p(z_j＝i|x_j)，有$\mu_i=\frac{\sum_{j=1}^{m}\gamma_{ji}x_j}{\sum_{j=1}^{m}\gamma_{ji}}$（11）；

由$\frac{\partial LL(D)}{\partial\Sigma_i}=0$可得：$\Sigma_i=\frac{\sum_{j=1}^{m}\gamma_{ji}(x_j-\mu_i)(x_j-\mu_i)^T}{\sum_{j=1}^{m}\gamma_{ji}}$ (12)；参数α_i除了要最大化 LL(D)，且$\alpha_i\geq0$，$\sum_{i=1}^{k}\alpha_i=1$。将LL(D)转换成拉格朗日公式：$\partial LL(D)+\lambda(\sum_{i=1}^{k}\alpha_i-1)$（13）；其中λ为拉格朗日乘子，当$\alpha_i$＝0，有$\sum_{j=1}^{m}\frac{p(x_j|\mu_i,\Sigma_i)}{\sum_{l=1}^{k}\alpha_l\cdot p(x_j||\mu_l,\Sigma_l)}+\lambda=0$ （14）；两边同乘以α_i，对所有混合成分求合

可知λ＝-m,有$\alpha_i = \frac{1}{m}\sum_{j-1}^{m}\gamma_{ji}$ （15）;由以上列式可获得高斯混合模型的EM算法:通过迭代,算出每个样本属于每个高斯成分的后验概率γ_{ji}(E步),然后通过式(11)、式(12)、式(15)更新模型参数 $\{(a_i,\mu_i,\Sigma_i|1\leqslant i\leqslant k\}$(M步)。

[0054] 综上,本发明以用户在预设时间段的上网日志,获得包括预先设置关键词的关键词信息,构建构建网贷风险行为分析特征,提取网贷网站的文本信息,并与所述网贷风险行为分析特征进行匹配,然后结合高斯混合聚类算法对网贷行为做出分析判断,提高了识别精度以及效率。

[0055] 在上述实施例的基础上,本发明一优选实施例中,为快速识别出网贷网站所属的类型,提高检测模型效率,基于深度包检测的应用识别算法,识别所有网站的应用标签类型,根据所述网站的应用标签类型,将网站进行区分分类,以获取分类后的网贷网站。具体地,

[0056] 基于深度包检测的应用识别算法对网站进行应用标签分类,通过对“指纹”技术(变动位置的特征、固定位置特征字和状态特征匹配三种类型)将其要识别的文本进行匹配。深度包检测技术通过对“指纹”的升级有很强的扩展功能,能实现对绝大部份网站协议的检测,从而实现网站的分类。

[0057] 其中,基于深度包检测是一种基于应用层的流量检测和控制技术,网络应用层中不同的应用对应的协议也不用,每个协议中都含有一个不同的“指纹”,在中,将网站中的“特征字”的通过比对数据报文中的“指纹”信息,以此来检测其业务流所对应的业务。某些业务的控制流和业务流是分离的,业务流没有任何特征。控制流是双方建立连接,协商信息发送的,所以包含了业务的特征信息,同时其数据内容包含协商出的数据流的五元组信息。首先识别出控制流,然后从控制流中解析出数据流的五元组信息,最后将数据流的五元组信息加入关联表中,使用关联表识别后续的数据流流量。

[0058] 为了便于说明,以下以实际应用场景为例进行说明:

[0059] 为了验证该方法的实用性和可靠性,采用现有的数据,共获取了6月份总计10G多的原始日志数据。

[0060] 实验平台的具体配置如下,CPU是Intel(R)Core(TM)i7-9700F,内存是16GB,硬盘容量为SSD512G,操作系统是Windows10。为了获取的日志样本的调用序列,所有日志样本运行在一台主机上,具体配置如下,CPU是Intel(R)Core(TM) i5 2.50GHz,内存是8GB,硬盘容量为SSD256G,操作系统是Windows 10。实验框架如图3所示,共分为三大模块:网贷网站为分析模块和网贷网站识别算法模块、网贷网站和校园网用户访问关系构建模块。

[0061] 从某NAT出口设备上采集了2019年6月份期间的用户上网日志(30天),日志信息为“用户日志数据集.CSV”文件,日志存储字段如表1所示。

[0062] 表1:

[0063]

字段名	表示
ID	序列号
Username	用户姓名
SourceIP	源 IP
DestinationIP	目的 IP
Web-Classify	网站分类
Title	访问标题

[0064]

urlname	访问域名
url-address	url 地址
time	访问时间

[0065] 根据选取的文本特征集对训练访问集合进行特征提取，假定聚类簇数k＝3，算法开始时抽取三个特征样本x_1,x_2,x_3作为初始均值向量，将高斯混合分布的模型参数初始化为$\alpha1=\alpha2=\alpha3=\frac{1}{3}$；$\mu_1=x_1,\mu_2=x_3,\mu_3=x_3$；$\Sigma_1=\Sigma_2=\Sigma_3=\begin{pmatrix}0.1 & 0.0\\ 0.0 & 0.1\end{pmatrix}$。在第一轮迭代中，先计算样本由各混合成分生成的后验概率，以x_4为例，由式(7) 算出后验概率$\gamma_{11}=0.00136$，$\gamma_{12}=0.00308$，$\gamma_{13}=0.00306$.所有样本的后验概率算完后，得到如下新的模型能数：$\alpha_1'=0.00342$，$\alpha_2'=0.00318$，$\alpha_3'=0.00306$；$\mu_1'=$ (0.00471;0.00232)，$\mu_2'=$(0.00563;0.00273)，$\mu_3'=$0.00514;0.00238)；$\Sigma_1'=\begin{pmatrix}0.00023 & 0.00003\\ 0.00003 & 0.00015\end{pmatrix}$，$\Sigma_2=\begin{pmatrix}0.00027 & 0.00003\\ 0.00003 & 0.00016\end{pmatrix}$，$\Sigma_3=\begin{pmatrix}0.00026 & 0.00005\\ 0.00005 & 0.00014\end{pmatrix}$；模式参数更新后，不断重复上述过程，不同轮数之后的聚类结果共得到163个校园贷访问信息，可得校园贷访问Rc＝0.00025％。

[0066] 参见图3，本发明第二实施例还提供了一种网贷风险行为分析检测装置，包括：

[0067] 采集单元100，用于采集用户在预设时间段的上网日志，以获得日志信息；其中，所述日志信息包括预先设置的关键词信息；

[0068] 构建单元200，用于根据所述关键词信息，以构建网贷风险行为分析特征；

[0069] 提取单元300，用于提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配；

[0070] 检测单元400，用于在匹配成功后，通过高斯混合聚类算法对所述网贷网站进行网贷风险行为分析检测。

[0071] 优选地，还包括：

[0072] 识别单元，用于基于深度包检测的应用识别算法，识别所有网站的应用标签类型；

[0073] 区分分类单元，用于根据所述网站的应用标签类型，将网站进行区分分类，以获取分类后的网贷网站。

[0074] 提取网贷网站的文本信息，并与所述网贷风险行为分析特征进行匹配，具体为：

[0075] 通过多变潜在语义索引文本敏感特征抽取算法对网贷网站的文本信息进行提取，并将提取后的文本信息与所述网贷风险行为分析特征进行匹配。

[0076] 优选地，检测单元400，具体包括：

[0077] 距离度量以及性能指标测量模块，用于在匹配成功后，对所述匹配后的文本信息进行距离度量以及性能指标测量；

[0078] 网贷风险行为检测模块，用于将距离度量以及性能指标测量后的文本信息，基于高斯混合聚类算法，检测所述网贷网站进行网贷风险行为。

[0079] 优选地，采用VDM对匹配后的文本信息进行距离度量，距离度量表达式为：$VDM_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$；$m_{u,a}$为在属性μ上取值为a的文本样本数，$m_{u,a,i}$为在第i个样本簇中属性μ上取值为a的样本数，k为文本特征样本簇数，$VDM_p(a,b)$为μ上的a和b两个离散值之间的VDM度量距离。

[0080] 本发明第三实施例：

[0081] 本发明第三实施例还提供了一种网贷风险行为分析检测设备，包括处理器、存储器以及存储在所述存储器内的计算机程序，所述计算机程序能够被所述处理器执行以实现如上述实施例所述的网贷风险行为分析检测方法。

[0082] 本发明第四实施例：

[0083] 本发明第四实施例提供了一种计算机可读存储介质，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上述的网贷风险行为分析检测方法。

[0084] 示例性的，所述计算机程序可以被分割成一个或多个单元，所述一个或者多个单元被存储在所述存储器中，并由所述处理器执行，以完成本发明。所述一个或多个单元可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述所述计算机程序在网贷风险行为分析检测设备中的执行过程。

[0085] 所述网贷风险行为分析检测设备可包括但不仅限于处理器、存储器。本领域技术人员可以理解，所述示意图仅仅是网贷风险行为分析检测设备的示例，并不构成对网贷风险行为分析检测设备的限定，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，例如所述网贷风险行为分析检测设备还可以包括输入输出设备、网络接入设备、总线等。

[0086] 所称处理器可以是中央处理单元(Central Processing Unit，CPU)，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,所述网贷风险行为分析检测设备的控制中心,利用各种接口和线路连接整个网贷风险行为分析检测设备的各个部分。

[0087] 所述存储器可用于存储所述计算机程序和/或模块,所述处理器通过运行或执行存储在所述存储器内的计算机程序和/或模块,以及调用存储在存储器内的数据,实现所述网贷风险行为分析检测设备的各种功能。所述存储器可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等;存储数据区可存储根据手机的使用所创建的数据(比如音频数据、电话本等)等。此外,存储器可以包括高速随机存取存储器,还可以包括非易失性存储器,例如硬盘、内存、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)、至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0088] 其中,所述网贷风险行为分析检测设备集成的单元如果以软件功能单元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,本发明实现上述实施例方法中的全部或部分流程,也可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一计算机可读存储介质中,该计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。其中,所述计算机程序包括计算机程序代码,所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM, Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是,所述计算机可读介质包含的内容可以根据专利实践的要求进行适当的增减,例如在某些专利实践的要求下,计算机可读介质不包括电载波信号和电信信号。

[0089] 需说明的是,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单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,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模块来实现本实施例方案的目的。另外,本发明提供的装置实施例附图中,模块之间的连接关系表示它们之间具有通信连接,具体可以实现为一条或多条通信总线或信号线。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0090] 以上所述是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也视为本发明的保护范围。

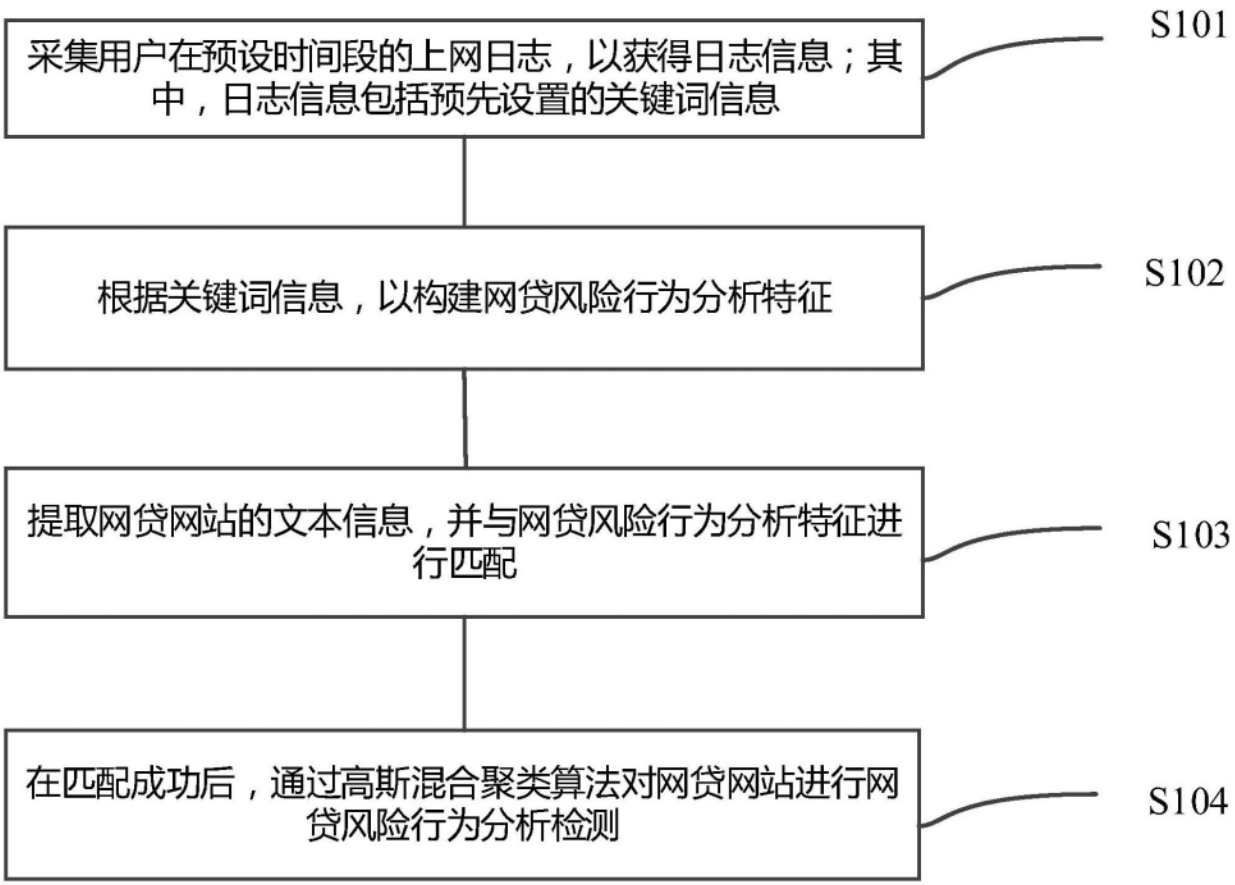

采集用户在预设时间段的上网日志，以获得日志信息；其中，日志信息包括预先设置的关键词信息
S101
根据关键词信息，以构建网贷风险行为分析特征
S102
提取网贷网站的文本信息，并与网贷风险行为分析特征进行匹配
S103
在匹配成功后，通过高斯混合聚类算法对网贷网站进行网贷风险行为分析检测
S104

图1

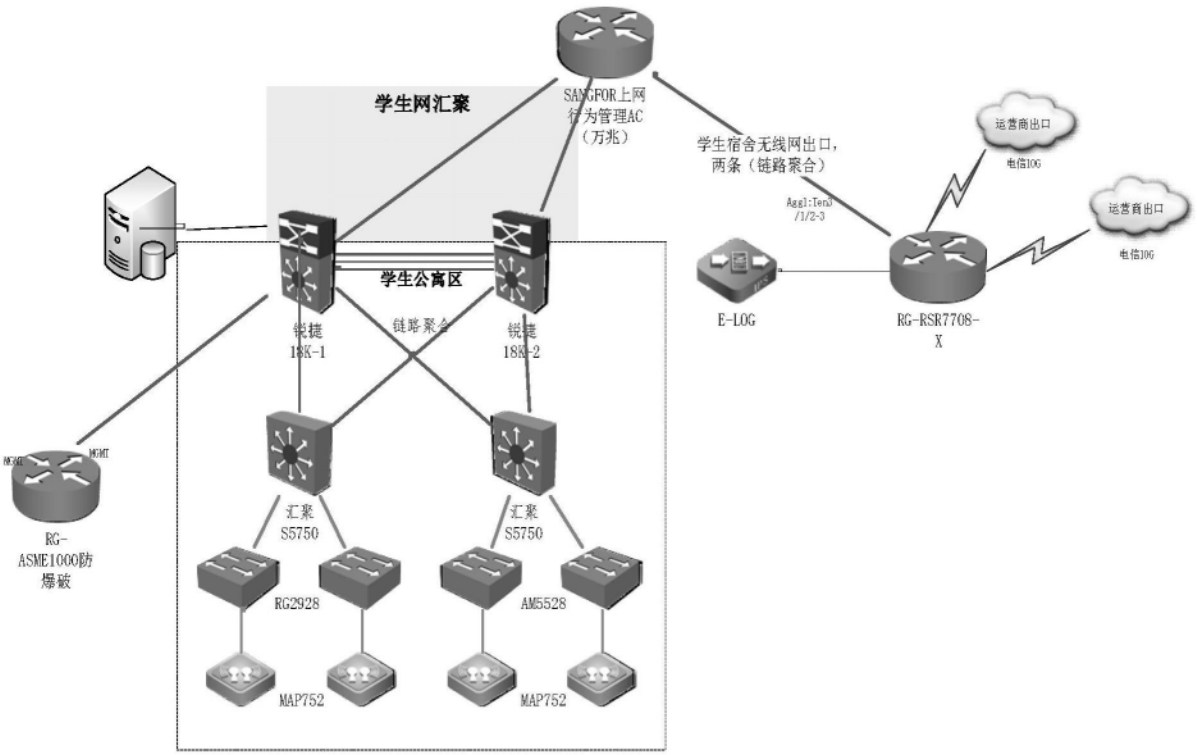

学生网汇聚
SANGFOR上网
行为管理AC
（万兆）
学生宿舍无线网出口，
两条（链路聚合）
运营商出口
电信10G
运营商出口
电信10G
Agg1:Ten3
/1/2-3
E-LOG
RG-RSR7708-
X
学生公寓区
锐捷
18K-1
链路聚合
锐捷
18K-2
MGMT
MGMT
RG-
ASME1000防
爆破
汇聚
S5750
汇聚
S5750
RG2928
AM5528
MAP752
MAP752

图2

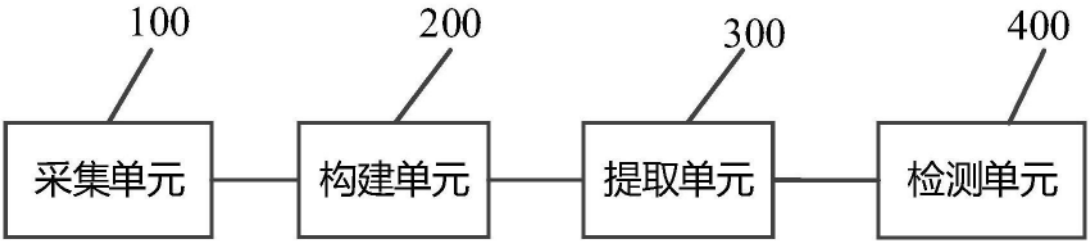

100
200
300
400
采集单元
构建单元
提取单元
检测单元

图3