(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109657119 A
(43)申请公布日 2019.04.19

(21)申请号 201811403690.X

(22)申请日 2018.11.23

(71)申请人 成都知道创宇信息技术有限公司
地址 610000 四川省成都市高新区天府三街219号2栋11楼

(72)发明人 仲俊霖

(74)专利代理机构 成都信博专利代理有限责任公司 51200
代理人 卓仲阳

(51)Int.Cl.
G06F 16/951(2019.01)

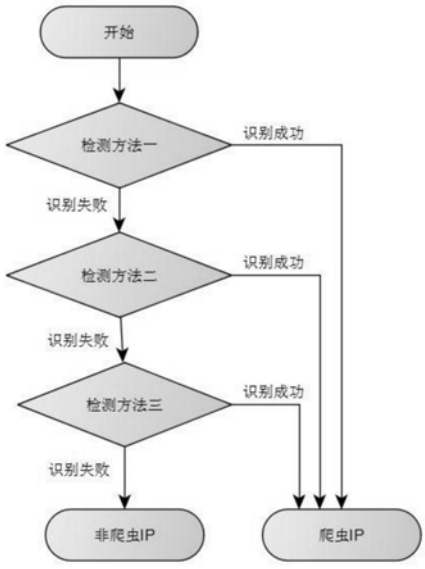
权利要求书1页 说明书4页 附图4页

(54)发明名称

一种基于访问日志IP分析的网络爬虫检测方法

(57)摘要

本发明公开了一种基于访问日志IP分析的网络爬虫检测方法，具体步骤是：使用特征检测法检测访问请求数据包中的特征来判断是否为普通爬虫；使用访问行为检测法检测IP访问静态资源和动态资源的比例来判断该IP是否为高级爬虫；使用特殊爬虫检测法检测网站接口的访问量来判断是否为爬虫；输出判定结果；本发明通过三种检测方法对IP进行识别，可以覆盖普通爬虫、高级爬虫和特殊爬虫，能够从更大范围内进行有效爬虫识别，在检测过程中还可以通过调节参数控制误报率，更加符合实际工作需要。



CN 109657119 A

1.一种基于访问日志IP分析的网络爬虫检测方法,其特征在于,包括以下步骤:

(1)、使用特征检测法检测访问请求数据包中的特征来判断是否为普通爬虫,如果识别成功则判定该IP属于网络爬虫,否则进入下一步;

(2)、使用访问行为检测法检测IP访问静态资源和动态资源的比例来判断该IP是否为高级爬虫,如果识别成功则判定该IP属于网络爬虫,否则进入下一步;

(3)、使用特殊爬虫检测法检测网站接口的访问量来判断是否为爬虫,如果识别成功则判定该IP属于网络爬虫,否则判定为非爬虫IP;

(4)、输出判定结果。

2.根据权利要求1所述的一种基于访问日志IP分析的网络爬虫检测方法,其特征在于,所述特征检测法适用于普通爬虫,首先获取访问请求中的UserAgent字段,检测UserAgent中是否包含自动化程序特征,包括python、ruby、PhantomJS、pycurl、httpunit、Wget、Java,如果检测到以上关键词特征则判定为爬虫。

3.根据权利要求1所述的一种基于访问日志IP分析的网络爬虫检测方法,其特征在于,所述访问行为检测法适用于高级爬虫,分为以下几个步骤:

(1)、将网站访问日志按照IP维度进行划分,即获取每个IP的全部网站访问日志;

(2)、在步骤(1)的基础上过滤出HTTP状态码等于200的日志,其它日志均去除掉;

(3)、在步骤(2)的基础上统计每个IP访问日志中访问静态资源和动态资源的比例;所述动、静态资源的区分标准通过访问资源的后缀名进行判断;

(4)、通过步骤(3)得到每个IP访问静态资源和动态资源的比例,如果该IP访问动态资源的比例超过静态资源的比例+预设值,则判定该IP属于爬虫IP。

4.根据权利要求3所述的一种基于访问日志IP分析的网络爬虫检测方法,其特征在于,所述静态资源文件后缀名包括但不限于:.jpg、.png、.js、.css、.gif、.ttf、.ico、.pdf、.mp3、.xls,可以根据网站的静态资源类型进行增减。

5.根据权利要求3所述的一种基于访问日志IP分析的网络爬虫检测方法,其特征在于,所述特殊爬虫检测法适用于特殊爬虫,分为以下几个步骤:

(1)、筛选出网站日志中HTTP状态码等于200的日志,其它日志均去除掉;

(2)、从步骤(1)筛选的日志中过滤出动态资源的访问日志;

(3)、统计步骤(2)结果中的日志条数和去重复后IP数,从而可以得到每个IP访问动态资源的平均次数;

(4)、在步骤(2)的基础上,统计出每个IP访问单个动态资源的次数列表;

(5)、将每个IP访问单个动态资源的次数与步骤3得到的平均访问次数进行对比,如果该IP访问某个动态资源的次数高于平均访问次数+预设值,则判定该IP为爬虫IP。

一种基于访问日志IP分析的网络爬虫检测方法

技术领域

[0001] 本发明涉及网络爬虫检测领域，具体涉及一种基于访问日志IP分析的网络爬虫检测方法。

背景技术

[0002] 随着互联网的发展，越来越多的行业开始通过网站的形式向广大网民展示其主营的各项业务和数据，而网络爬虫则可以自动的获取这些数据，从而爬虫所有者可以通过这些数据获利，例如有的人通过编写爬虫程序爬取电商网站的商品信息，从这些数据中可以获取到每个商品的价格，而作为竞争对手可以以此价格为参考，适当降低自己商场的同款商品的价格，从而保持销售优势。又或者对于一些权威信息，如企业信用信息查询，这些数据只能通过政府网站进行查询，而爬虫编写者可以通过网络爬虫批量获取数据，然后将这些数据转卖给需要这些信息的人从而获利。

[0003] 目前常用的反爬虫手段是通过访问频率来对爬虫IP进行封锁。首先设置一个访问阈值，当某个IP的访问频率超过阈值则拦截此IP的访问请求。

[0004] 这种方法对于普通爬虫来说可以起到很好的拦截效果，但是对于高级爬虫则并不能有效进行拦截。爬虫可以通过降低访问频率、增加IP数量等方式绕过检测。而且这种方法还容易产生误拦截，如公司出口IP、小区出口IP等，一个IP并不一定代表唯一一个正常用户，有时候一个IP可能有几百个用户在使用，如果单通过访问频率来识别则有可能会对正常用户进行误拦截。

[0005] 术语定义：

[0006] 网络爬虫:又称为网页蜘蛛，网络机器人，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。网络爬虫被广泛应用于搜索引擎或用来爬取特定网站内容。

[0007] 访问日志:即网站用户访问记录，访问日志详细记录了每个用户访问网站的情况，其中包含访问者的IP地址、访问的RUL、访问时间等等内容。

发明内容

[0008] 为解决上述问题，本发明分别从特征识别、访问行为识别的角度针对普通爬虫、高级爬虫和特殊爬虫进行多重检测。

[0009] 本发明提供一种基于访问日志IP分析的网络爬虫检测方法，具体包括以下步骤：

[0010] 1、使用特征检测法检测访问请求数据包中的特征来判断是否为普通爬虫，如果识别成功则判定该IP属于网络爬虫，否则进入下一步；

[0011] 2、使用访问行为检测法检测IP访问静态资源和动态资源的比例来判断该IP是否为高级爬虫，如果识别成功则判定该IP属于网络爬虫，否则进入下一步；

[0012] 3、使用特殊爬虫检测法检测网站接口的访问量来判断是否为爬虫，如果识别成功则判定该IP属于网络爬虫，否则判定为非爬虫IP；

[0013] 4、输出判定结果。

[0014] 其中三种检测方法具体如下：

[0015] 一、特征检测法

[0016] 特征检测法适用于普通爬虫，首先获取访问请求中的UserAgent字段，检测UserAgent中是否包含自动化程序特征，包括python、ruby、PhantomJS、pycurl、httpunit、Wget、Java，如果检测到以上关键词特征则判定为爬虫。

[0017] 二、访问行为检测法

[0018] 访问行为检测法适用于高级爬虫，分为以下几个步骤：

[0019] (1)、将网站访问日志按照IP维度进行划分，即获取每个IP的全部网站访问日志；

[0020] (2)、在步骤(1)的基础上过滤出HTTP状态码等于200的日志，其它日志均去除掉；

[0021] (3)、在步骤(2)的基础上统计每个IP访问日志中访问静态资源和动态资源的比例；所述动、静态资源的区分标准通过访问资源的后缀名进行判断；

[0022] (4)、通过步骤(3)得到每个IP访问静态资源和动态资源的比例，如果该IP访问动态资源的比例超过静态资源的比例+预设值，则判定该IP属于爬虫IP。

[0023] 静态资源文件后缀名包括但不限于：.jpg、.png、.js、.css、.gif、.ttf、.ico、.pdf、.mp3、.xls，可以根据网站的静态资源类型进行增减。

[0024] 三、特殊爬虫检测法

[0025] 特殊爬虫检测法适用于特殊爬虫，分为以下几个步骤：

[0026] (1)、筛选出网站日志中HTTP状态码等于200的日志，其它日志均去除掉；

[0027] (2)、从步骤(1)筛选的日志中过滤出动态资源的访问日志；

[0028] (3)、统计步骤(2)结果中的日志条数和去重复后IP数，从而可以得到每个IP访问动态资源的平均次数；

[0029] (4)、在步骤(2)的基础上，统计出每个IP访问单个动态资源的次数列表；

[0030] (5)、将每个IP访问单个动态资源的次数与步骤3得到的平均访问次数进行对比，如果该IP访问某个动态资源的次数高于平均访问次数+预设值，则判定该IP为爬虫IP。

[0031] 本发明技术方案带来的有益效果为：

[0032] 从特征识别和访问行为识别两个角度切入，通过三种检测方法对IP进行识别，可以覆盖普通爬虫、高级爬虫和特殊爬虫，能够从更大范围内进行有效爬虫识别，在检测过程中还可以通过调节参数控制误报率，更加符合实际工作需要。

附图说明

[0033] 图1为本方案的流程图；

[0034] 图2为访问行为检测法的流程图；

[0035] 图3为特殊爬虫检测法平均参考值获取流程图；

[0036] 图4为特殊爬虫检测法的流程图。

具体实施方式

[0037] 下面结合附图对本发明进一步详细说明。

[0038] 本发明的流程如图1所示，具体为：

[0039] 1、使用特征检测法检测访问请求数据包中的特征来判断是否为普通爬虫，如果识

别成功则判定该IP属于网络爬虫。

[0040] 首先获取访问请求中的UserAgent字段，检测UserAgent中是否包含自动化程序特征，包括python、ruby、PhantomJS、pycurl、httpunit、Wget、Java，如果检测到以上关键词特征则判定为爬虫。

[0041] 注：以上的特征关键词是通过收集常见的自动化程序的UserAgent而来，在技术领域能够发起HTTP请求的工具通常是被技术人员熟知的，因此收集这些工具的特征并不困难。如果遇到新的工具出现可以将其特征添加到我们的UserAegnt特征库中。

[0042] 如果识别失败则进入下一步。

[0043] 2、使用访问行为检测法检测IP访问静态资源和动态资源的比例来判断该IP是否为高级爬虫，如果识别成功则判定该IP属于网络爬虫。

[0044] 问行为检测法如图2所示，具体分为以下几个步骤：

[0045] 2.1、将网站访问日志按照IP维度进行划分，即获取每个IP的全部网站访问日志；

[0046] 2.2、在步骤2.1的基础上过滤出HTTP状态码等于200的日志，其它日志均去除掉；

[0047] 2.3、在步骤2.2的基础上统计每个IP访问日志中访问静态资源和动态资源的比例；所述动、静态资源的区分标准通过访问资源的后缀名进行判断；

[0048] 2.4、通过步骤2.3得到每个IP访问静态资源和动态资源的比例，如果该IP访问动态资源的比例超过静态资源的比例+预设值，则判定该IP属于爬虫IP。

[0049] 注：在一般情况下，正常用户访问一个网页通常会附带多个静态资源请求，即在正常情况下网站用户访问动态资源的比例应当比访问静态资源的比例小得多，而网络爬虫通常不会去访问静态资源，它们多以获取数据为主(某些以获取静态资源的爬虫除外)，由于静态资源通常对于爬虫没有意义，因此这类爬虫多以访问动态资源为主，由此，这些爬虫具备了访问动态资源明显多于访问静态资源的行为特征。本方法也是基于这样的行为特征来判断该IP是否为网络爬虫。

[0050] 如果识别失败则进入下一步；

[0051] 3、使用特殊爬虫检测法检测网站接口的访问量来判断是否为爬虫，如果识别成功则判定该IP属于网络爬虫，

[0052] 特殊爬虫检测法如图4所示，具体分为以下几个步骤：

[0053] 3.1、筛选出网站日志中HTTP状态码等于200的日志，其它日志均去除掉；

[0054] 3.2、从步骤3.1筛选的日志中过滤出动态资源的访问日志；

[0055] 3.3、统计步骤3.2结果中的日志条数和去重复后IP数，从而可以得到每个IP访问动态资源的平均次数，如图3所示；例如，整个网站的动态资源访问次数为100次，去重复后的IP有20个，那么平均每个IP访问了动态资源5次。

[0056] 3.4、在步骤3.2的基础上，统计出每个IP访问单个动态资源的次数列表；例如：

[0057]

IP	动态资源	访问次数
233.12.45.123	/data/demo1	61
	/data/getdata1	19113
	/data/getdata2	29530

[0058] 3.5、将每个IP访问单个动态资源的次数与步骤3得到的平均访问次数进行对比，如果该IP访问某个动态资源的次数高于平均访问次数+预设值，则判定该IP为爬虫IP。

[0059] 注：本方法主要是对网站接口类爬虫进行检测，这类爬虫通常只会访问网站的某几个接口，而且访问量特别大，因此它们具备了访问URL数量少，访问次数多的特点。本方法首先计算出整个网站动态资源的平均访问次数，然后再计算出每个IP访问单个动态资源的次数，注意是单个动态资源，假如某个IP在爬取网站接口数据，那么它对这些网站接口的访问量将明显高出平均访问次数，本方法以此可以识别出这类网络爬虫的IP。

[0060] 如果识别失败则判定为非爬虫IP。

[0061] 4、输出判定结果。

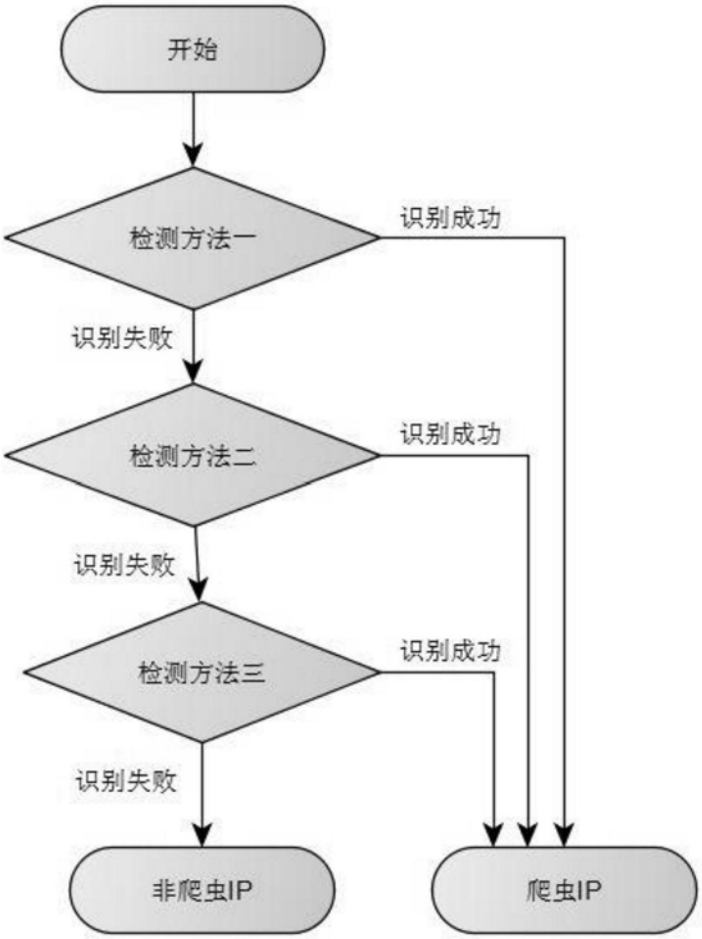

开始
检测方法一
识别成功
识别失败
检测方法二
识别成功
识别失败
检测方法三
识别成功
识别失败
非爬虫IP
爬虫IP

图1

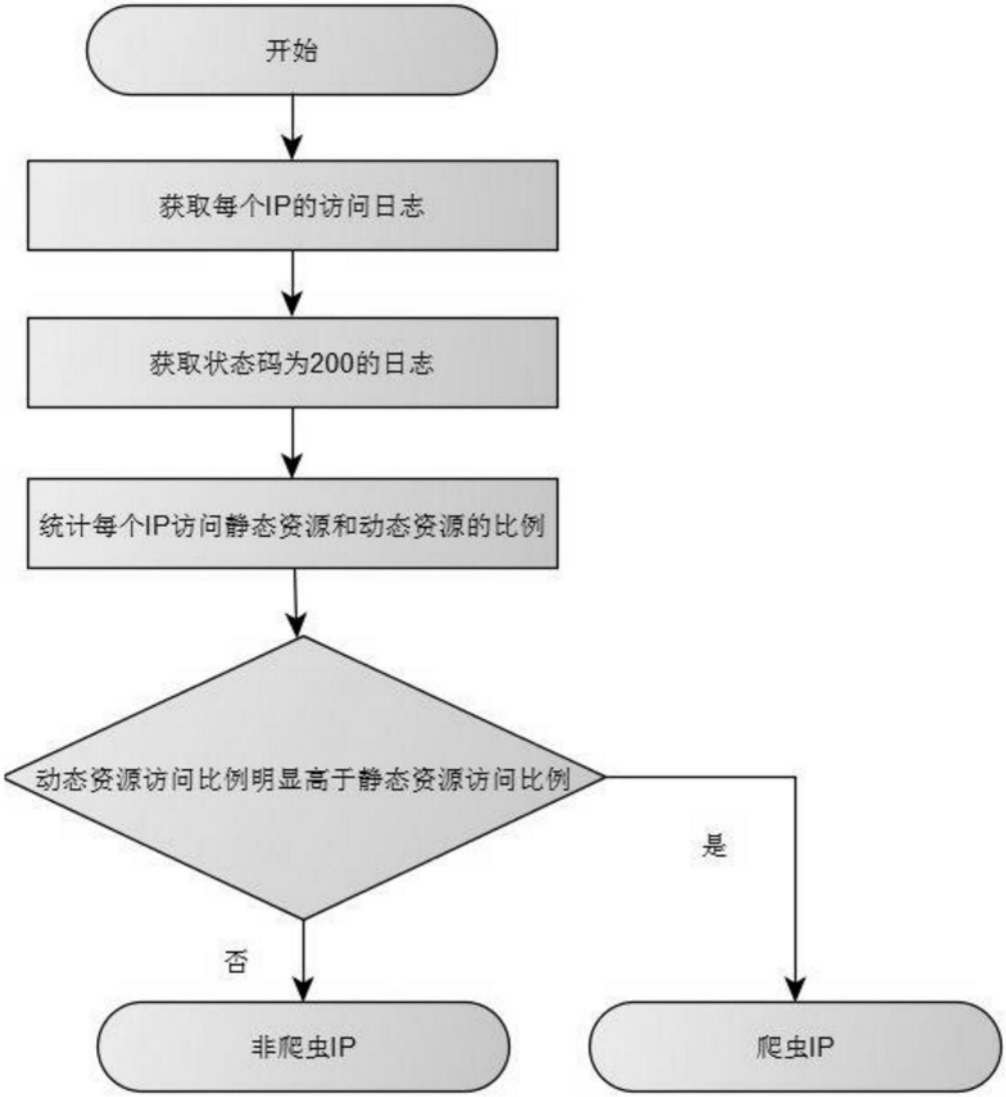

开始
获取每个IP的访问日志
获取状态码为200的日志
统计每个IP访问静态资源和动态资源的比例
动态资源访问比例明显高于静态资源访问比例
是
否
非爬虫IP
爬虫IP

图2

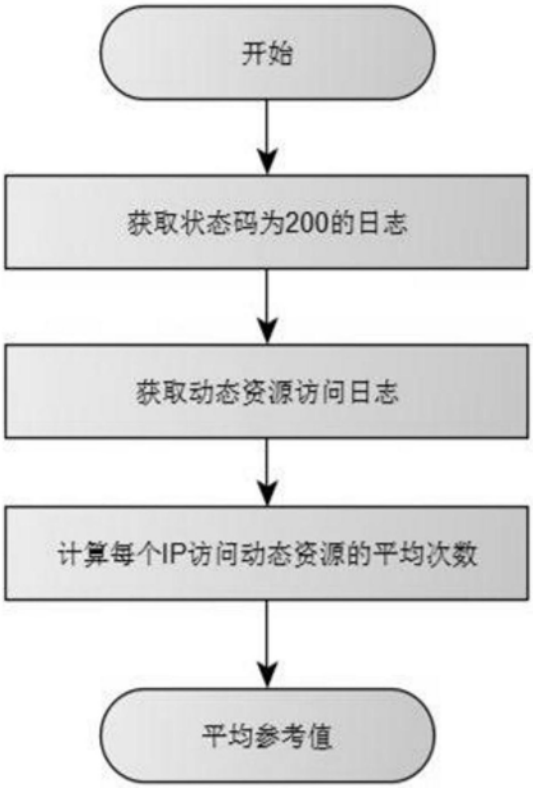


图3

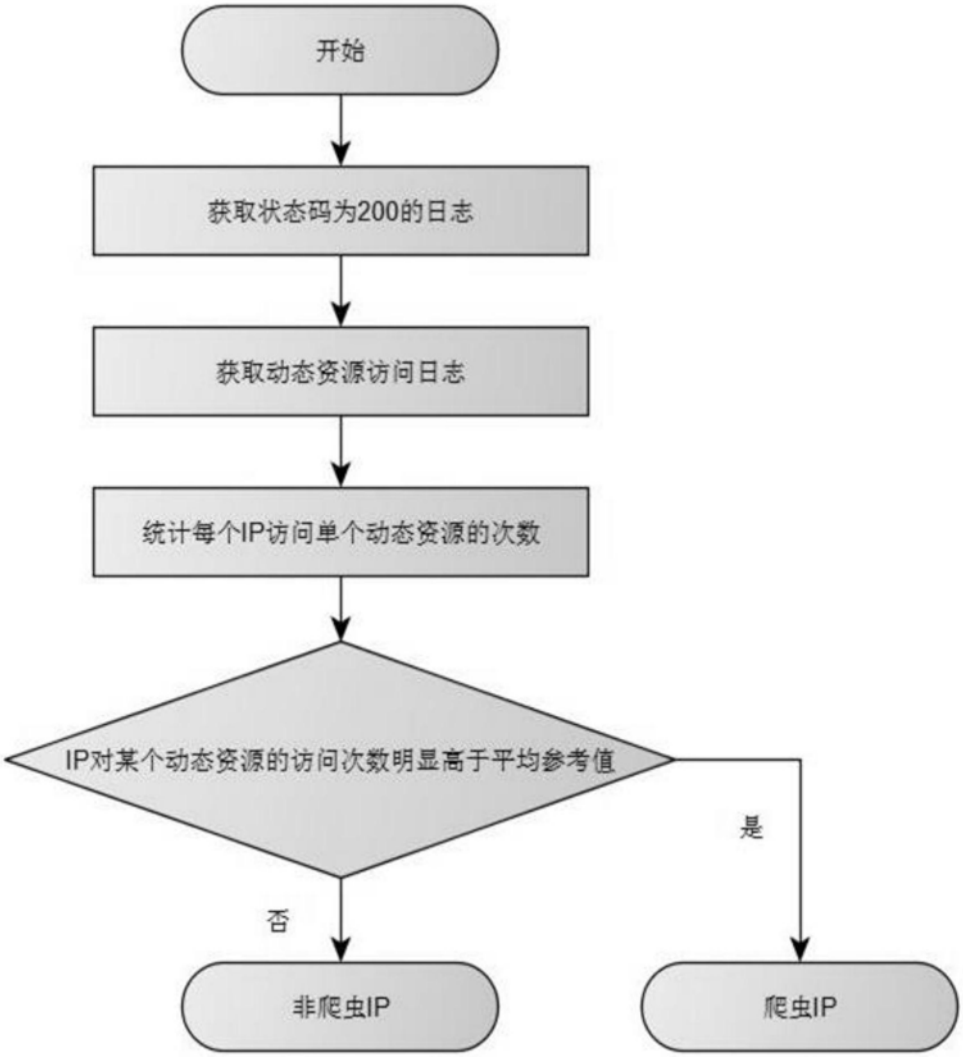

开始
获取状态码为200的日志
获取动态资源访问日志
统计每个IP访问单个动态资源的次数
IP对某个动态资源的访问次数明显高于平均参考值
是
否
非爬虫IP
爬虫IP

图4