

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(43) 国际公布日
2019 年 10 月 24 日 (24.10.2019)

WIPO PCT

(10) 国际公布号
WO 2019/200784 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01) **H04L 29/06** (2006.01)

(21) 国际申请号: PCT/CN2018/100162

(22) 国际申请日: 2018 年 8 月 13 日 (13.08.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810349987.6 2018年4月18日 (18.04.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田區福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

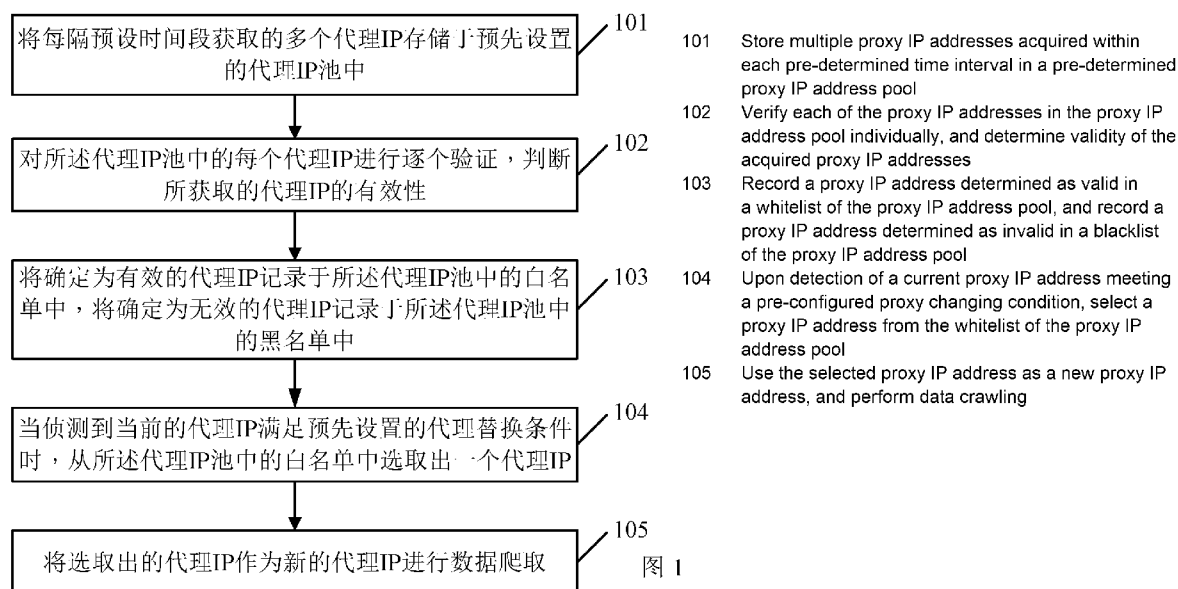
(72) 发明人: 阮晓雯(RUAN, Xiaowen); 中国广东省深圳市福田區福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 徐亮(XU, Liang); 中国广东省深圳市福田區福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 肖京(XIAO, Jing); 中国广东省深圳市福田區福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市赛恩倍吉知识产权代理有限公司(SHENZHEN SCIENBIZIP INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深圳市龙华新区龙观东路 83 号荣群大厦 9 楼, Guangdong 518109 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) Title: METHOD FOR WEB CRAWLER, TERMINAL, AND STORAGE MEDIUM

(54) 发明名称: 网络爬虫方法、终端及存储介质



(57) Abstract: A method for a web crawler, the method comprising: storing multiple proxy IP addresses acquired within each pre-determined time interval in a pre-determined proxy IP address pool; verifying each of the proxy IP addresses in the proxy IP address pool individually, and determining validity of the acquired proxy IP addresses; recording a proxy IP address determined as valid in a whitelist of the proxy IP address pool, and recording a proxy IP address determined as invalid in a blacklist of the proxy IP address pool; upon detection of a current proxy IP address meeting a pre-configured proxy changing condition, selecting a proxy IP address from the whitelist of the proxy IP address pool; and using the selected proxy IP address as a new proxy IP address, and performing data crawling. The present application further provides a terminal and a storage medium. The present application effectively resolves the issue of IP address restrictions when the same proxy IP address is used during a process of prolonged, frequent, fast, and large-volume

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

data crawling.

(57) 摘要: 一种网络爬虫方法, 包括: 将每隔预设时间段获取的多个代理IP存储于预先设置的代理IP池中; 对所述代理IP池中的每个代理IP进行逐个验证, 判断所获取的代理IP的有效性; 将确定为有效的代理IP记录于所述代理IP池中的白名单中, 将确定为无效的代理IP记录于所述代理IP池中的黑名单中; 当侦测到当前的代理IP满足预先设置的代理替换条件时, 从所述代理IP池中的所述白名单中选取出一个代理IP; 及将选取出的代理IP作为新的代理IP进行数据爬取。本申请还提供一种终端及存储介质。本申请能够有效地解决同一代理IP在长时间、多次快速、大量爬取数据的过程中的IP受限问题。

网络爬虫方法、终端及存储介质

本申请要求于 2018 年 04 月 18 日提交中国专利局，申请号为 201810349987.6、发明名称为“网络爬虫方法、终端及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及网络爬虫技术领域，具体涉及一种网络爬虫方法、终端及存储介质。

背景技术

网络爬虫是搜索引擎系统中十分重要的组成部分，它负责从互联网中搜集网页，采集信息，这些网页信息用于设置索引从而为搜索引擎提供支持，其性能的优劣直接影响着搜索引擎的效果。随着网络信息量几何级的增长，对网络爬虫页面采集的性能和效率的要求也越来越高。

我们总是希望在更短的时间内，获取更多的数据，但是，这会对网站造成非常高的负载，也带来了网络流量增加，泄露隐私数据等问题，很多网站采用爬虫检测技术，分析 Web 访问日志，当判断出爬虫时，则禁止爬虫使用地址，拒绝爬虫继续访问。尤其是在批量爬取百度指数和微博指数等时，同一账号同一 IP 的爬取次数或频率会受限。

因而为了防止爬虫被监测出来，有必要提出一种对抗反爬虫的机制。

发明内容

鉴于以上内容，有必要提出一种网络爬虫方法、终端及存储介质，结合深度信息，构造代理 IP 池，根据预先设定的选取规则或策略从代理 IP 池中挑选出代理 IP 进行爬取，有效的解决了同一代理 IP 的爬取次数或频率受限的问题。

本申请的第一方面提供一种网络爬虫方法，所述方法包括：

将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中；

对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性；

将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中，将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中；

当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP；及

将选取出的代理 IP 作为新的代理 IP 进行数据爬取。

本申请的第二方面提供一种终端，所述终端包括处理器和存储器，所述处理器用于执行所述存储器中存储的计算机可读指令时实现所述网络爬虫方法。

本申请的第三方面提供一种非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现所述网络爬虫方法。

本申请所述的网络爬虫方法、终端及存储介质，能够有效地解决同一代

理 IP 在长时间、多次快速、大量爬取数据的过程中的 IP 受限问题。尤其能够不断的获取代理 IP 并实时更新代理 IP 池，保证代理 IP 池中的优秀代理足够多；其次，根据代理 IP 的访问成功率和访问响应时间进一步确定代理 IP 是否具有第二有效性，以此确定代理 IP 的有效性程度；第三，针对确定为具有第二有效性的代理 IP 还进一步地细分了有效级别；第四，对于首次访问失败的代理 IP，还进行多次验证，进一步细分为永久无效和暂时无效，可以避免代理 IP 的浪费。不仅能够有效地解决同一代理 IP 在长时间、多次快速、大量爬取数据的过程中的 IP 受限问题，还能保证选取出最有效的代理 IP 进行数据爬取

附图说明

图 1 是本申请实施例一提供的网络爬虫方法的流程图。

图 2 是本申请实施例二提供的网络爬虫方法的流程图。

图 3 是本申请实施例三提供的网络爬虫装置的结构图。

图 4 是本申请实施例四提供的网络爬虫装置的结构图。

图 5 是本申请实施例五提供的终端的示意图。

如下具体实施方式将结合上述附图进一步说明本申请。

具体实施方式

本申请实施例的网络爬虫方法应用在一个或者多个终端中。所述网络爬虫方法也可以应用于由终端和通过网络与所述终端进行连接的服务器所构成的硬件环境中。本申请实施例的网络爬虫方法可以由服务器来执行，也可以由终端来执行；还可以是由服务器和终端共同执行。

所述对于需要进行网络爬虫方法的终端，可以直接在终端上集成本申请的方法所提供的网络爬虫功能，或者安装用于实现本申请的方法的客户端。再如，本申请所提供的方法还可以以软件开发工具包（Software Development Kit, SDK）的形式运行在服务器等设备上，以 SDK 的形式提供网络爬虫功能的接口，终端或其他设备通过提供的接口即可实现网络爬虫功能。

实施例一

图 1 是本申请实施例一提供的网络爬虫方法的流程图。

101：将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中。

本实施例中，在本地数据库中预先设置一个代理 IP 池，将所获取的多个代理 IP 添加到该代理 IP 池中，以供爬虫使用。可以在互联网上提供的代理 IP 的网站站点中找到代理 IP，具体的列表可以通过人工方式或另外的小爬虫自动获取。也可以通过第三方服务机构购买多个代理 IP，并将所获取的代理 IP 添加到预先设置的代理 IP 池中。

本实施例中，所述代理 IP 的代理信息可以包括，但不限于：IP 地址、名称及所述端口。

本实施例中，可以每隔预设时间段，例如，每隔一天，一周等，通过人工或另外的小爬虫自动在互联网上提供的代理 IP 的网站站点中获取代理 IP 或者通过第三方服务机构购买多个代理 IP 的方式获取代理 IP，并将所获取

的代理 IP 存储于所述代理 IP 池中，如此可以保证所述代理池中的 IP 的数量足够多，且不间断的获取代理 IP，可以更新所述代理 IP 池，供爬虫使用。

102：对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性。

本实施例中，将进行有效性验证的代理 IP 称之为待验证的代理 IP，使用待验证的代理 IP 首次访问搜索引擎（例如，谷歌、百度等）以验证是否得到搜索引擎的响应。如果首次访问得到搜索引擎的响应，表明该待验证的代理 IP 有效，如果首次访问没有得到搜索引擎的响应，表明该待验证的代理 IP 无效。所述是否能够得到搜索引擎的响应是指待验证的代理 IP 首次能够从所访问的搜索引擎中爬取到数据，即不会被所访问的搜索引擎进行首次访问限制。

103：将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中，将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中。

本实施例中，预先在所述代理 IP 池中设置一个白名单列表和一个黑名单列表。所述白名单列表用于记录所述代理 IP 池中被确定为有效的代理 IP，所述黑名单列表用于记录所述代理 IP 池中被确定为无效的代理 IP。

104：当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP。

本实施例中，预先设置代理替换条件，当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP。

本实施例中，所述当前的代理 IP 满足预先设置的代理替换条件包括以下一种或多种的组合：

1) 当前的代理 IP 的访问时间超过预先设置的访问时间阈值；

所述预先设置的访问时间阈值可以是 10 分钟，则当所述当前的代理 IP 的访问时间超过 10 分钟时，即侦测到所述当前的代理 IP 满足了所述预先设置的代理替换条件，则从所述代理 IP 池中选取出一个代理 IP，如此能够有效地解决同一代理 IP 长时间爬取数据时造成的 IP 受限问题。

2) 当前的代理 IP 的访问频率超过预先设置的访问频率阈值；

所述预先设置的访问频率阈值可以是 100 次/秒，则当所述当前的代理 IP 的访问频率超过 100 次/秒时，即侦测到所述当前的代理 IP 满足了所述预先设置的代理替换条件，则从所述代理 IP 池中选取出一个代理 IP，如此能够有效地解决同一代理 IP 多次快速爬取数据时造成的 IP 受限问题。

3) 当前的代理 IP 的访问数量超过预先设置的访问数量阈值；

所述预先设置的访问数量阈值是 200，则当所述当前的代理 IP 的访问数量超过 200 后时，即侦测到所述当前的代理 IP 满足了所述预先设置的代理替换条件，则从所述代理 IP 池中选取出一个代理 IP，如此可以有效的解决同一代理 IP 大量爬取数据时造成的 IP 受限问题。

105：将选取出的代理 IP 作为新的代理 IP 进行数据爬取。

本实施例中，将当前的代理 IP 替换为从所述代理 IP 池中的所述白名单

中选取出的代理 IP，所选取出的代理 IP 作为新的代理 IP 进行数据爬取。只要当前的代理 IP 在爬取数据的过程中满足所述预先设置的代理替换条件时，即可自动进行代理 IP 的替换，从而可以免去人工频繁替换的麻烦，爬取效率更高。

综上所述，本申请所述的网络爬虫方法，将获取的多个代理 IP 存储于预先设置的代理 IP 池中；对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性；将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中，将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中；当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP；将选取出的代理 IP 作为新的代理 IP 进行数据爬取。能够有效地解决同一代理 IP 在长时间、多次快速、大量爬取数据的过程中的 IP 受限问题。

实施例二

图 2 是本申请实施例二提供的网络爬虫方法的流程图。

201：将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中。

本实施例中的步骤 201 同实施例一中的步骤 101，在此不再详细赘述。

202：对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 是否具有第一有效性。

本实施例中，将进行第一有效性验证的代理 IP 称之为待验证的代理 IP，使用待验证的代理 IP 访问搜索引擎（例如，谷歌、百度等）以验证是否得到搜索引擎的响应。如果得到搜索引擎的响应，表明该待验证的代理 IP 具有第一有效性，如果没有得到搜索引擎的响应，表明该待验证的代理 IP 不具有第一有效性。所述是否能够得到搜索引擎的响应是指待验证的代理 IP 能够从所访问的搜索引擎中爬取到数据，即不会被所访问的搜索引擎进行访问限制。

当确定所述代理 IP 池中的代理 IP 具有所述第一有效性时，执行步骤 203；当确定所述代理 IP 池中的代理 IP 不具有所述第一有效性时，执行步骤 204。

203：根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性。

为了得到更有效的能够完成访问任务的代理 IP，可以进一步将代理 IP 的访问成功率和访问响应时间作为验证所述代理 IP 是否有效的标准。

所述根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性具体包括：

1) 使用具有所述第一有效性的代理 IP 多次访问多个搜索引擎，计算所述具有第一有效性的代理 IP 的访问成功率及访问响应时间；

所述访问成功率是指在预设的时间周期内访问成功的次数占访问总次数的比例。例如，在 1 个小时的周期内，具有第一有效性的代理 IP 访问搜索引擎的总次数为 100 次，其中，访问成功的次数为 97 次，访问失败的次数为 3 次，则计算所述具有第一有效性的代理 IP 的访问成功率为 $97/100=97\%$ 。

所述访问响应时间是指发出访问请求到接收到访问回应的的时间。例如，

具有第一有效性的代理 IP 在 9:55:54 秒发出访问请求，在 9:55:55 秒接收到访问回应，则计算所述具有第一有效性的代理 IP 的访问响应时间为 1 秒。

2) 判断所述具有第一有效性的代理 IP 的访问成功率是否大于预先设置的访问成功率阈值，同时判断所述具有第一有效性的代理 IP 的访问响应时间是否小于预先设置的访问响应时间阈值；

所述预先设置的访问成功率阈值可以是，例如，80%。所述预先设置的访问响应时间阈值可以是，例如，1 秒。

3) 当所述具有第一有效性的代理 IP 的访问成功率大于所述预先设置的访问成功率阈值且访问响应时间小于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 具有第二有效性；当所述具有第一有效性的代理 IP 的访问成功率小于或等于所述预先设置的访问成功率阈值，或访问响应时间大于或等于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 不具有第二有效性。

本实施例中，根据所述待验证的代理 IP 是否能够访问搜索引擎，在确定能访问搜索引擎时继续根据访问成功率和访问响应时间判断所述待验证的代理 IP 是否具有更高的有效性，如此不仅能够确定所述代理 IP 有效，而且还能进一步确定所述待验证的代理 IP 的质量。访问成功率越大，访问响应时间越快的代理 IP 对应的质量越好，访问成功率越小，访问响应时间越慢的代理 IP 对应的质量越差。

当确定具有所述第一有效性的代理 IP 具有第二有效性时，执行步骤 204；当确定具有所述第一有效性的代理 IP 不具有第二有效性时，执行步骤 205。

204: 将确定为具有第二有效性的代理 IP 记录于所述代理 IP 池中的白名单中。

本实施例中，预先在所述代理 IP 池中设置一个白名单列表，所述白名单列表用于记录所述代理 IP 池中被确定为具有第二有效性的代理 IP。

在所述白名单中记录代理 IP 的访问成功率及访问响应时间，目的是为了后续选取代理 IP 时，能够有针对性的进行选取。例如，优先选取访问成功率越高及/访问响应时间越快的代理 IP，能够使得爬取数据的效率更高，爬取到的数据量更多。

进一步地，所述方法还可以包括：根据代理 IP 的访问成功率及访问响应时间，对代理 IP 设置多个有效级别，将所述多个有效级别及其对应的代理 IP 记录于所述白名单中。

所述多个有效级别可以包括，但不限于：第一有效级别、第二有效级别及第三有效级别。第一有效级别对应第一访问成功率及第一响应时间，第二有效级别对应第二访问成功率及第二响应时间，第三有效级别对应第三访问成功率及第三响应时间，以此类推。所述第一有效级别的级别最高，所述第三有效级别的级别最低。本申请对设置的有效级别的数量不作具体限制，可以根据实际需求设置两个或者更多个。

具体地，将访问成功率大于预先设置的第一访问成功率（例如，95%）及访问响应时间大于预先设置的第一访问响应时间（0.5 秒）对应的代理 IP

作为第一有效级别的代理 IP；将访问成功率小于预先设置的第一访问成功率但大于预先设置的第二访问成功率阈值（例如，90%），访问响应时间小于预先设置的第一访问响应时间但大于预先设置的第二访问响应（例如，1 秒）对应的代理 IP 作为第二有效级别的代理 IP；将访问成功率小于预先设置的第二访问成功率及访问响应时间小于预先设置的第二访问响应时间对应的代理 IP 作为第三有效级别的代理 IP。

本实施例中，根据所述访问成功率及访问响应时间确定代理 IP 的有效级别，便于后续选取代代理 IP 时，能够快速地从第一有效级别对应的多个代理 IP 中选取一个代理 IP 进行数据爬取。

更进一步的，所述方法还可以包括：所述白名单中还记录具有第二有效性的代理 IP 访问的搜索引擎的类型。

在所述白名单中记录代理 IP 对应能够访问的搜索引擎的类型，例如，当前代理 IP 只能访问百度，或者只能访问搜狗，或者当前代理 IP 能够访问任何一个搜索引擎。

本实施例中，在所述白名单中记录代理 IP 访问的搜索引擎的类型，是为了后续选取代代理 IP 时，能够根据当前的代理 IP 访问的搜索引擎的类型进行有针对性的替取代代理 IP。例如，当前的代理 IP 正在访问百度，后续需要替换当前的代理 IP 时，即可根据搜索引擎类型为百度对应的多个代理 IP 中选取一个代理 IP 进行数据爬取。

本实施例中，所述白名单列表还记录有所述具有第二有效性的代理 IP 及其对应的访问成功率、访问响应时间、访问的搜索引擎的类型、获取的时间、访问的次数等。

205: 将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中。

本实施例中，预先在所述代理 IP 池中设置一个黑名单列表，所述黑名单列表用于记录所述代理 IP 池中被确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP。

进一步地，所述将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中还可以包括：

1) 将确定为具有第一有效性但不具有第二有效性的代理 IP 记录于所述黑名单中的第一黑名单子列表中，将确定为不具有第一有效性的代理 IP 记录于所述黑名单中的第二黑名单子列表中。

2) 使用不具有第一有效性的代理 IP 多次访问多个搜索引擎，计算不具有第一有效性的代理 IP 的访问失败率。

3) 判断不具有第一有效性的代理 IP 的访问失败率是否小于预先设置的访问失败率。

所述预先设置的访问失败率阈值可以是，例如，50%。

4) 当所述不具有第一有效性的代理 IP 的访问失败率小于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为暂时无效的代理 IP，并记录于所述第一黑名单子列表中；当所述不具有第一有效性的代理

IP 的访问失败率大于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为永久无效的代理 IP，并记录于所述第二黑名单子列表中。

本实施例中，将首次进行验证时无法成功访问搜索引擎的代理 IP 确认为不具有第一有效性的代理 IP，但后续多次验证时确定访问失败率小于预先设置的失败率阈值时，认为该不具有第一有效性的代理 IP 为暂时无效的代理 IP，将其从所述第二黑名单子列表中移至所述第一黑名单子列表中。原因在于代理 IP 并不是任何时候都能很好的完成工作，代理 IP 可能不稳定，导致它某一段时间内无法使用；或者由于当前搜索引擎自身出现的问题（例如，搜索引擎性能下降）或者网速慢等原因造成的验证失败，但是以后可能依然能继续使用；或者代理 IP 因为频繁访问等因素被禁了，但是一段时间以后禁止访问权限可能会被解除。

其次，通过多次验证，以进一步验证不具有第一有效性的代理 IP 的访问失败率大于所述预先设置的访问失败率时，认为该不具有第一有效性的代理 IP 为真正无效的代理 IP，避免一次验证失败而误认为该代理 IP 永久无效，造成了代理 IP 的浪费。后续如果所述白名单中的具有第二有效性的代理 IP 全部无法使用时，可从所述第一黑名单子列表中选取代理 IP。

另外，所述第二黑名单子列表中记录不具有第一有效性的且访问失败率大于所述预先设置的访问失败率的代理 IP，还可以便于后续再在免费网站获取 IP 或者通过第三方服务机构购买代理 IP 时，能够直接与所述第二黑名单子列表中的永久无效的代理 IP 进行匹配，从而快速判断该代理 IP 是否为永久无效的代理 IP，避免对新记录于所述代理池中的代理 IP 进行逐个验证，可以节省时间。

206：当侦测到当前的代理 IP 满足预先设置的代理替换条件时，根据预先设置的代理选取规则从所述代理 IP 池中的所述白名单中选取出一个代理 IP。

本实施例中，预先设置代理选取规则，所述预先设置的代理选取规则包括以下一种或多种的组合：

1) 根据当前访问的搜索引擎类型从所述白名单中对应所述访问的搜索引擎类型的代理 IP 中选取；

例如，所述白名单中记录了各个代理 IP 的访问的搜索引擎类型，若当前访问的搜索引擎为百度时，则从所述白名单中的对应访问百度的代理 IP 中选取出一个代理 IP。

2) 根据所述白名单中记录的代理 IP 的爬取次数进行选取；

例如，所述白名单中记录了各个代理 IP 的爬取次数，则将各个代理 IP 的爬取次数进行排序，选取爬取次数少的代理 IP。

3) 根据所述白名单中记录的代理 IP 访问的搜索引擎类型的数量进行选取；

例如，所述白名单中记录了各个代理 IP 的访问的搜索引擎类型，则统计各个代理 IP 访问的搜索引擎类型的数量，将各个代理 IP 访问的搜索引擎类

型的数量进行排序，则选取访问的搜索引擎类型的数量较大的代理 IP。

4) 根据所述白名单中记录的代理 IP 的获取时间进行选取；

例如，所述白名单中记录了各个代理 IP 的获取时间，则选取最新记录入白名单中的代理 IP。

5) 延迟预设时间段后选取代理 IP。

侦测到当前的代理 IP 满足所述预先设置的代理替换条件时，延迟预设时间段后，根据上述规则 1) 至 4) 从所述白名单中选取出一个代理 IP。

207: 将选取出的代理 IP 作为新的代理 IP 进行数据爬取。

优选地，所述网络爬虫方法还可以包括：提供用户选项，供用户根据实际需要自行添加、删除或者更改。如此能更进一步地及时更新所述白名单列表，保证所述白名单列表中的代理 IP 都是可用的且有效的代理，消除代理的改变对数据爬取的影响。

优选地，所述网络爬虫方法还可以包括：将多个代理设备 useragent 进行级联且代理设备之间可相互通讯，每一个代理设备的本地存储器中均存储有预先设置的所述代理 IP 池，当某一个 useragent 的访问时间超过预设时间，替换另一个 useragent 进行数据爬取。

优选地，所述网络爬虫方法还可以包括：将所构造的代理 IP 池存储于预设设置的服务器中，通过服务器分配 IP 给代理设备进行数据爬取。所述代理设备可以包括移动设备、web 设备等。

优选地，所述网络爬虫方法还可以包括：注册多个账号，使用同一 IP 时，可通过切换不同的账号进行数据爬取。

需要说明的是，本申请所提供的从代理 IP 池中获取 IP 的过程及多个代理设置 useragent 爬取数据可同步进行，即在更换代理设备的同时可更换代理 IP。

综上所述，本申请所述的网络爬虫方法，能够不断的获取代理 IP 并实时更新代理 IP 池，保证代理 IP 池中的优秀代理足够多；其次，根据代理 IP 的访问成功率和访问响应时间进一步确定代理 IP 是否具有第二有效性，以此确定代理 IP 的有效性程度；第三，针对确定为具有第二有效性的代理 IP 还进一步地细分了有效级别；第四，对于首次访问失败的代理 IP，还进行多次验证，进一步细分为永久无效和暂时无效，可以避免代理 IP 的浪费。不仅能够有效地解决同一代理 IP 在长时间、多次快速、大量爬取数据的过程中的 IP 受限问题，还能保证选取出最有效的代理 IP 进行数据爬取。

上述对图 1、图 2 流程中各个步骤的描述，可以根据不同的需求，流程图中的执行顺序可以改变，某些步骤可以省略。

下面结合第 3 至 5 图，分别对实现上述网络爬虫方法的终端的功能模块及硬件结构进行介绍。

实施例三

图 3 为本申请网络爬虫装置较佳实施例中的功能模块图。

在一些实施例中，所述网络爬虫装置 30 运行于终端中。所述网络爬虫装置 30 可以包括多个由程序代码段所组成的功能模块。所述网络爬虫装置 30

中的各个程序段的程序代码可以存储于存储器中，并由至少一个处理器所执行，以执行（详见图 1 及其相关描述）对手部区域的跟踪。

本实施例中，所述终端的网络爬虫装置 30 根据其所执行的功能，可以被划分为多个功能模块。所述功能模块可以包括：存储模块 301、判断模块 302、记录模块 303、选取模块 304 及爬取模块 305。本申请所称的模块是指一种能够被至少一个处理器所执行并且能够完成固定功能的一系列计算机可读指令段，其存储在所述存储器中。

存储模块 301，用于将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中。

本实施例中，在本地数据库中预先设置一个代理 IP 池，将所获取的多个代理 IP 添加到该代理 IP 池中，以供爬虫使用。可以在互联网上提供的代理 IP 的网站站点中找到代理 IP，具体的列表可以通过人工方式或另外的小爬虫自动获取。也可以通过第三方服务机构购买多个代理 IP，并将所获取的代理 IP 添加到预先设置的代理 IP 池中。

本实施例中，所述代理 IP 的代理信息可以包括，但不限于：IP 地址、名称及所述端口。

本实施例中，可以每隔预设时间段，例如，每隔一天，一周等，通过人工或另外的小爬虫自动在互联网上提供的代理 IP 的网站站点中获取代理 IP 或者通过第三方服务机构购买多个代理 IP 的方式获取代理 IP，并将所获取的代理 IP 存储于所述代理 IP 池中，如此可以保证所述代理池中的 IP 的数量足够多，且不间断的获取代理 IP，可以更新所述代理 IP 池，供爬虫使用。

判断模块 302，用于对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性。

本实施例中，将进行有效性验证的代理 IP 称之为待验证的代理 IP，使用待验证的代理 IP 首次访问搜索引擎（例如谷歌、百度等）以验证是否得到搜索引擎的响应。如果首次访问得到搜索引擎的响应，表明该待验证的代理 IP 有效，如果首次访问没有得到搜索引擎的响应，表明该待验证的代理 IP 无效。所述是否能够得到搜索引擎的响应是指待验证的代理 IP 首次能够从所访问的搜索引擎中爬取到数据，即不会被所访问的搜索引擎进行首次访问限制。

记录模块 303，用于将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中，将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中。

本实施例中，预先在所述代理 IP 池中设置一个白名单列表和一个黑名单列表。所述白名单列表用于记录所述代理 IP 池中被确定为有效的代理 IP，所述黑名单列表用于记录所述代理 IP 池中被确定为无效的代理 IP。

选取模块 304，用于当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP。

本实施例中，预先设置代理替换条件，当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP。

本实施例中,所述当前的代理 IP 满足预先设置的代理替换条件包括以下一种或多种的组合:

1) 当前的代理 IP 的访问时间超过预先设置的访问时间阈值;

所述预先设置的访问时间阈值可以是 10 分钟,则当所述当前的代理 IP 的访问时间超过 10 分钟时,即侦测到所述当前的代理 IP 满足了所述预先设置的代理替换条件,则从所述代理 IP 池中选取出一个代理 IP,如此能够有效地解决同一代理 IP 长时间爬取数据时造成的 IP 受限问题。

2) 当前的代理 IP 的访问频率超过预先设置的访问频率阈值;

所述预先设置的访问频率阈值可以是 100 次/秒,则当所述当前的代理 IP 的访问频率超过 100 次/秒时,即侦测到所述当前的代理 IP 满足了所述预先设置的代理替换条件,则从所述代理 IP 池中选取出一个代理 IP,如此能够有效地解决同一代理 IP 多次快速爬取数据时造成的 IP 受限问题。

3) 当前的代理 IP 的访问数量超过预先设置的访问数量阈值;

所述预先设置的访问数量阈值是 200,则当所述当前的代理 IP 的访问数量超过 200 后时,即侦测到所述当前的代理 IP 满足了所述预先设置的代理替换条件,则从所述代理 IP 池中选取出一个代理 IP,如此可以有效的解决同一代理 IP 大量爬取数据时造成的 IP 受限问题。

爬取模块 305,用于将选取出的代理 IP 作为新的代理 IP 进行数据爬取。

本实施例中,将当前的代理 IP 替换为从所述代理 IP 池中的所述白名单中选取出的代理 IP,所选取出的代理 IP 作为新的代理 IP 进行数据爬取。只要当前的代理 IP 在爬取数据的过程中满足所述预先设置的代理替换条件时,即可自动进行代理 IP 的替换,从而可以免去人工频繁替换的麻烦,爬取效率更高。

综上所述,本申请所述的网络爬虫装置 30,存储模块 301 将获取的多个代理 IP 存储于预先设置的代理 IP 池中;判断模块 302 对所述代理 IP 池中的每个代理 IP 进行逐个验证,判断所获取的代理 IP 的有效性;记录模块 303 将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中,将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中;选取模块 304 侦测到当前的代理 IP 满足预先设置的代理替换条件时,从所述代理 IP 池中的所述白名单中选取出一个代理 IP;爬取模块 305 将选取出的代理 IP 作为新的代理 IP 进行数据爬取。能够有效地解决同一代理 IP 在长时间、多次快速、大量爬取数据的过程中的 IP 受限问题。

实施例四

图 4 为本申请网络爬虫装置的较佳实施例中的功能模块图。

在一些实施例中,所述网络爬虫装置 40 运行于终端中。所述网络爬虫装置 40 可以包括多个由程序代码段所组成的功能模块。所述网络爬虫装置 40 中的各个程序段的程序代码可以存储于存储器中,并由至少一个处理器所执行,以执行(详见图 2 及其相关描述)对网络爬虫的防止。

本实施例中,所述终端的网络爬虫装置 40 根据其所执行的功能,可以被划分为多个功能模块。所述功能模块可以包括:存储模块 401、第一判断模

块 402、第二判断模块 403、第一记录模块 404、第二记录模块 405、选取模块 406 及爬取模块 407。

存储模块 401，用于将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中。

本实施例中，可以每隔预设时间段，例如，每隔一天，一周等，通过人工或另外的小爬虫自动在互联网上提供的代理 IP 的网站站点中获取代理 IP 或者通过第三方服务机构购买多个代理 IP 的方式获取代理 IP，并将所获取的代理 IP 存储于所述代理 IP 池中，如此可以保证所述代理池中的 IP 的数量足够多，且不间断的获取代理 IP，可以更新所述代理 IP 池，供爬虫使用。

第一判断模块 402，用于对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 是否具有第一有效性。

本实施例中，将进行第一有效性验证的代理 IP 称之为待验证的代理 IP，使用待验证的代理 IP 访问搜索引擎（例如，谷歌、百度等）以验证是否得到搜索引擎的响应。如果得到搜索引擎的响应，表明该待验证的代理 IP 具有第一有效性，如果没有得到搜索引擎的响应，表明该待验证的代理 IP 不具有第一有效性。所述是否能够得到搜索引擎的响应是指待验证的代理 IP 能够从所访问的搜索引擎中爬取到数据，即不会被所访问的搜索引擎进行访问限制。

第二判断模块 403，用于当所述第一判断模块 402 确定所述代理 IP 池中的代理 IP 具有所述第一有效性时，根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性。

为了得到更有效的能够完成访问任务的代理 IP，可以进一步将代理 IP 的访问成功率和访问响应时间作为验证所述代理 IP 是否有效的标准。

所述第二判断模块 403 根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性具体包括：

1) 使用具有所述第一有效性的代理 IP 多次访问多个搜索引擎，计算所述具有第一有效性的代理 IP 的访问成功率及访问响应时间；

所述访问成功率是指在预设的时间周期内访问成功的次数占访问总次数的比例。例如，在 1 个小时的周期内，具有第一有效性的代理 IP 访问搜索引擎的总次数为 100 次，其中，访问成功的次数为 97 次，访问失败的次数为 3 次，则计算所述具有第一有效性的代理 IP 的访问成功率为 $97/100=97\%$ 。

所述访问响应时间是指发出访问请求到接收到访问回应的的时间。例如，具有第一有效性的代理 IP 在 9:55:54 秒发出访问请求，在 9:55:55 秒接收到访问回应，则计算所述具有第一有效性的代理 IP 的访问响应时间为 1 秒。

2) 判断所述具有第一有效性的代理 IP 的访问成功率是否大于预先设置的访问成功率阈值，同时判断所述具有第一有效性的代理 IP 的访问响应时间是否小于预先设置的访问响应时间阈值；

所述预先设置的访问成功率阈值可以是，例如，80%。所述预先设置的访问响应时间阈值可以是，例如，1 秒。

3) 当所述具有第一有效性的代理 IP 的访问成功率大于所述预先设置的访问成功率阈值且访问响应时间小于所述预先设置的访问响应时间阈值时，

确定所述具有第一有效性的代理 IP 具有第二有效性；当所述具有第一有效性的代理 IP 的访问成功率小于或等于所述预先设置的访问成功率阈值，或访问响应时间大于或等于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 不具有第二有效性。

本实施例中，根据所述待验证的代理 IP 是否能够访问搜索引擎，在确定能访问搜索引擎时继续根据访问成功率和访问响应时间判断所述待验证的代理 IP 是否具有更高的有效性，如此不仅能够确定所述代理 IP 有效，而且还能进一步确定所述待验证的代理 IP 的质量。访问成功率越大，访问响应时间越快的代理 IP 对应的质量越好，访问成功率越小，访问响应时间越慢的代理 IP 对应的质量越差。

第一记录模块 404，用于将确定为具有第二有效性的代理 IP 记录于所述代理 IP 池中的白名单中。

本实施例中，预先在所述代理 IP 池中设置一个白名单列表，所述白名单列表用于记录所述代理 IP 池中被确定为具有第二有效性的代理 IP。

在所述白名单中记录代理 IP 的访问成功率及访问响应时间，目的是为了后续选取代理 IP 时，能够有针对性的进行选取。例如，优先选取访问成功率越高及/访问响应时间越快的代理 IP，能够使得爬取数据的效率更高，爬取到的数据量更多。

进一步地，所述第一记录模块 404 还用于根据代理 IP 的访问成功率及访问响应时间，对代理 IP 设置多个有效级别，将所述多个有效级别及其对应的代理 IP 记录于所述白名单中。

所述多个有效级别可以包括，但不限于：第一有效级别、第二有效级别及第三有效级别。第一有效级别对应第一访问成功率及第一响应时间，第二有效级别对应第二访问成功率及第二响应时间，第三有效级别对应第三访问成功率及第三响应时间，以此类推。所述第一有效级别的级别最高，所述第三有效级别的级别最低。本申请对设置的有效级别的数量不作具体限制，可以根据实际需求设置两个或者更多个。

具体地，将访问成功率大于预先设置的第一访问成功率（例如，95%）及访问响应时间大于预先设置的第一访问响应时间（0.5 秒）对应的代理 IP 作为第一有效级别的代理 IP；将访问成功率小于预先设置的第一访问成功率但大于预先设置的第二访问成功率阈值（例如，90%），访问响应时间小于预先设置的第一访问响应时间但大于预先设置的第二访问响应（例如，1 秒）对应的代理 IP 作为第二有效级别的代理 IP；将访问成功率小于预先设置的第二访问成功率及访问响应时间小于预先设置的第二访问响应时间对应的代理 IP 作为第三有效级别的代理 IP。

本实施例中，根据所述访问成功率及访问响应时间确定代理 IP 的有效级别，便于后续选取代理 IP 时，能够快速地从第一有效级别对应的多个代理 IP 中选取一个代理 IP 进行数据爬取。

更进一步的，所述第一记录模块 404 还用于将具有第二有效性的代理 IP 访问的搜索引擎的类型记录于所述白名单中。

在所述白名单中记录代理 IP 对应能够访问的搜索引擎的类型，例如，当前代理 IP 只能访问百度，或者只能访问搜狗，或者当前代理 IP 能够访问任何一个搜索引擎。

本实施例中，在所述白名单中记录代理 IP 访问的搜索引擎的类型，是为了后续选取代理 IP 时，能够根据当前的代理 IP 访问的搜索引擎的类型进行有针对性的替换代理 IP。例如，当前的代理 IP 正在访问百度，后续需要替换当前的代理 IP 时，即可根据搜索引擎类型为百度对应的多个代理 IP 中选取一个代理 IP 进行数据爬取。

本实施例中，所述白名单列表还记录有所述具有第二有效性的代理 IP 及其对应的访问成功率、访问响应时间、访问的搜索引擎的类型、获取的时间、访问的次数等。

第二记录模块 405，用于将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中。

本实施例中，预先在所述代理 IP 池中设置一个黑名单列表，所述黑名单列表用于记录所述代理 IP 池中被确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP。

进一步地，所述第二记录模块 405 将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中还可以包括：

1) 将确定为具有第一有效性但不具有第二有效性的代理 IP 记录于所述黑名单中的第一黑名单子列表中，将确定为不具有第一有效性的代理 IP 记录于所述黑名单中的第二黑名单子列表中。

2) 使用不具有第一有效性的代理 IP 多次访问多个搜索引擎，计算不具有第一有效性的代理 IP 的访问失败率。

3) 判断不具有第一有效性的代理 IP 的访问失败率是否小于预先设置的访问失败率。

所述预先设置的访问失败率阈值可以是，例如，50%。

4) 当所述不具有第一有效性的代理 IP 的访问失败率小于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为暂时无效的代理 IP，并记录于所述第一黑名单子列表中；当所述不具有第一有效性的代理 IP 的访问失败率大于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为永久无效的代理 IP，并记录于所述第二黑名单子列表中。

本实施例中，将首次进行验证时无法成功访问搜索引擎的代理 IP 确认为不具有第一有效性的代理 IP，但后续多次验证时确定访问失败率小于预先设置的失败率阈值时，认为该不具有第一有效性的代理 IP 为暂时无效的代理 IP，将其从所述第二黑名单子列表中移至所述第一黑名单子列表中。原因在于代理 IP 并不是任何时候都能很好的完成工作，代理 IP 可能不稳定，导致它某一段时间内无法使用；或者由于当前搜索引擎自身出现的问题（例如，搜索引擎性能下降）或者网速慢等原因造成的验证失败，但是以后可能依然

能继续使用；或者代理 IP 因为频繁访问等因素被禁了，但是一段时间以后禁止访问权限可能会被解除。

其次，通过多次验证，以进一步验证不具有第一有效性的代理 IP 的访问失败率大于所述预先设置的访问失败率时，认为该不具有第一有效性的代理 IP 为真正无效的代理 IP，避免一次验证失败而误认为该代理 IP 永久无效，造成了代理 IP 的浪费。后续如果所述白名单中的具有第二有效性的代理 IP 全部无法使用时，可从所述第一黑名单子列表中选取代理 IP。

另外，所述第二黑名单子列表中记录不具有第一有效性的且访问失败率大于所述预先设置的访问失败率的代理 IP，还可以便于后续再在免费网站获取 IP 或者通过第三方服务机构购买代理 IP 时，能够直接与所述第二黑名单子列表中的永久无效的代理 IP 进行匹配，从而快速判断该代理 IP 是否为永久无效的代理 IP，避免对新记录于所述代理池中的代理 IP 进行逐个验证，可以节省时间。

选取模块 406，用于当侦测到当前的代理 IP 满足预先设置的代理替换条件时，根据预先设置的代理选取规则从所述代理 IP 池中的所述白名单中选取出一个代理 IP。

本实施例中，预先设置代理选取规则，所述预先设置的代理选取规则包括以下一种或多种的组合：

1) 根据当前访问的搜索引擎类型从所述白名单中对应所述访问的搜索引擎类型的代理 IP 中选取；

例如，所述白名单中记录了各个代理 IP 的访问的搜索引擎类型，若当前访问的搜索引擎为百度时，则从所述白名单中的对应访问百度的代理 IP 中选取出一个代理 IP。

2) 根据所述白名单中记录的代理 IP 的爬取次数进行选取；

例如，所述白名单中记录了各个代理 IP 的爬取次数，则将各个代理 IP 的爬取次数进行排序，选取爬取次数少的代理 IP。

3) 根据所述白名单中记录的代理 IP 访问的搜索引擎类型的数量进行选取；

例如，所述白名单中记录了各个代理 IP 的访问的搜索引擎类型，则统计各个代理 IP 访问的搜索引擎类型的数量，将各个代理 IP 访问的搜索引擎类型的数量进行排序，则选取访问的搜索引擎类型的数量较大的代理 IP。

4) 根据所述白名单中记录的代理 IP 的获取时间进行选取；

例如，所述白名单中记录了各个代理 IP 的获取时间，则选取最新记录入白名单中的代理 IP。

5) 延迟预设时间段后选取代理 IP。

侦测到当前的代理 IP 满足所述预先设置的代理替换条件时，延迟预设时间段后，根据上述规则 1) 至 4) 从所述白名单中选取出一个代理 IP。

爬取模块 407，用于将选取出代理 IP 作为新的代理 IP 进行数据爬取。

优选地，所述白名单列表还可提供用户选项，供用户根据实际需要自行添加、删除或者更改，如此能更进一步地及时更新所述白名单列表，保证所

述白名单列表中的代理 IP 都是可用的且有效的代理，消除代理的改变对数据爬取的影响。

综上所述，本申请所述的网络爬虫装置 40，能够不断的获取代理 IP 并实时更新代理 IP 池，保证代理 IP 池中的优秀代理足够多；其次，根据代理 IP 的访问成功率和访问响应时间进一步确定代理 IP 是否具有第二有效性，以此确定代理 IP 的有效性程度；第三，针对确定为具有第二有效性的代理 IP 还进一步地细分了有效级别；第四，对于首次访问失败的代理 IP，还进行多次验证，进一步细分为永久无效和暂时无效，可以避免代理 IP 的浪费。

实施例五

图 5 为本申请实施例五提供的终端的示意图。

所述终端 5 包括：存储器 51、至少一个处理器 52、存储在所述存储器 51 中并可在所述至少一个处理器 52 上运行的计算机可读指令 53 及至少一条通讯总线 54。

所述至少一个处理器 52 执行所述计算机可读指令 53 时实现上述网络爬虫方法实施例中的步骤。

所述终端 5 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。本领域技术人员可以理解，所述示意图 5 仅仅是终端 5 的示例，并不构成对终端 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，例如所述终端 5 还可以包括输入输出设备、网络接入设备、总线等。

所述至少一个处理器 52 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。该处理器 52 可以是微处理器或者该处理器 52 也可以是任何常规的处理器等，所述处理器 52 是所述终端 5 的控制中心，利用各种接口和线路连接整个终端 5 的各个部分。

所述存储器 51 可用于存储所述计算机可读指令 53 和/或模块/单元，所述处理器 52 通过运行或执行存储在所述存储器 51 内的计算机可读指令和/或模块/单元，以及调用存储在存储器 51 内的数据，实现所述终端 5 的各种功能。所述存储器 51 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终端 5 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 51 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。

所述终端 5 集成的模块/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，可以存储在一个非易失性可读存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计

计算机可读指令来指令相关的硬件来完成。其中，所述计算机可读指令包括计算机可读指令代码，所述计算机可读指令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。所述非易失性可读介质可以包括：能够携带所述计算机可读指令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，所述非易失性可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，非易失性可读介质不包括电载波信号和电信信号。

另外，在本申请各个实施例中的各功能单元可以集成在相同处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在相同单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

此外，显然“包括”一词不排除其他单元或，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神范围。

权利要求书

1. 一种网络爬虫方法，其特征在于，所述方法包括：
将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中；
对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性；
将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中，将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中；
当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP；及
将选取出的代理 IP 作为新的代理 IP 进行数据爬取。
2. 如权利要求 1 所述的方法，其特征在于，所述对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性包括：
对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 是否具有第一有效性；
根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性；
将确定为具有第二有效性的代理 IP 记录于所述代理 IP 池中的白名单中；
将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中。
3. 如权利要求 2 所述的方法，其特征在于，所述根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性包括：
使用具有所述第一有效性的代理 IP 多次访问多个搜索引擎，计算所述具有第一有效性的代理 IP 的访问成功率及访问响应时间；
判断所述具有第一有效性的代理 IP 的访问成功率是否大于预先设置的访问成功率阈值，同时判断所述具有第一有效性的代理 IP 的访问响应时间是否小于预先设置的访问响应时间阈值；
当所述具有第一有效性的代理 IP 的访问成功率大于所述预先设置的访问成功率阈值且访问响应时间小于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 具有第二有效性；
当所述具有第一有效性的代理 IP 的访问成功率小于或等于所述预先设置的访问成功率阈值，或访问响应时间大于或等于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 不具有第二有效性。
4. 如权利要求 2 所述的方法，其特征在于，所述将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中包括：
将确定为具有第一有效性但不具有第二有效性的代理 IP 记录于所述黑名单中的第一黑名单子列表中，将确定为不具有第一有效性的代理 IP 记录于所述黑名单中的第二黑名单子列表中；
使用不具有第一有效性的代理 IP 多次访问多个搜索引擎，计算不具有第一

有效性的代理 IP 的访问失败率；

判断不具有第一有效性的代理 IP 的访问失败率是否小于预先设置的访问失败率；

当所述不具有第一有效性的代理 IP 的访问失败率小于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为暂时无效的代理 IP，并记录于所述第一黑名单子列表中；或

当所述不具有第一有效性的代理 IP 的访问失败率大于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为永久无效的代理 IP，并记录于所述第二黑名单子列表中。

5. 如权利要求 1 至 4 任意一项所述的方法，其特征在于，所述方法还包括：根据代理 IP 的访问成功率及访问响应时间，对代理 IP 设置多个有效级别，将所述多个有效级别及其对应的代理 IP 记录于所述白名单中。

6. 如权利要求 5 所述的方法，其特征在于，所述对代理 IP 设置多个有效级别包括：

将访问成功率大于预先设置的第一访问成功率及访问响应时间大于预先设置的第一访问响应时间对应的代理 IP 作为第一有效级别的代理 IP；

将访问成功率小于预先设置的第一访问成功率但大于预先设置的第二访问成功率阈值，访问响应时间小于预先设置的第一访问响应时间但大于预先设置的第二访问响应对应的代理 IP 作为第二有效级别的代理 IP；

将访问成功率小于预先设置的第二访问成功率及访问响应时间小于预先设置的第二访问响应时间对应的代理 IP 作为第三有效级别的代理 IP。

7. 如权利要求 1 所述的方法，其特征在于，所述从所述代理 IP 池中的所述白名单中选取出一个代理 IP 是根据预先设置的代理选取规则进行选取，所述预先设置的代理选取规则包括以下一种或多种的组合：

根据当前访问的搜索引擎类型从所述白名单中对应所述访问的搜索引擎类型的代理 IP 中选取；根据所述白名单中记录的代理 IP 的爬取次数进行选取；根据所述白名单中记录的代理 IP 访问的搜索引擎类型的数量进行选取；根据所述白名单中记录的代理 IP 的获取时间进行选取；延迟预设时间段后选取代理 IP。

8. 如权利要求 1 所述的方法，其特征在于，所述方法还包括：

对所述白名单列表提供用户选项，根据用户的添加、删除或更改操作更新所述白名单列表。

9. 一种终端，其特征在于，所述终端包括处理器和存储器，所述处理器用于执行所述存储器中存储的计算机可读指令时实现如下步骤：

将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中；

对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性；

将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中，将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中；

当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP

池中的所述白名单中选取出一个代理 IP；及

将选取出的代理 IP 作为新的代理 IP 进行数据爬取。

10. 如权利要求 9 所述的终端，其特征在于，所述对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性包括：

对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 是否具有第一有效性；

根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性；

将确定为具有第二有效性的代理 IP 记录于所述代理 IP 池中的白名单中；

将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中。

11. 如权利要求 10 所述的终端，其特征在于，所述根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性包括：

使用具有所述第一有效性的代理 IP 多次访问多个搜索引擎，计算所述具有第一有效性的代理 IP 的访问成功率及访问响应时间；

判断所述具有第一有效性的代理 IP 的访问成功率是否大于预先设置的访问成功率阈值，同时判断所述具有第一有效性的代理 IP 的访问响应时间是否小于预先设置的访问响应时间阈值；

当所述具有第一有效性的代理 IP 的访问成功率大于所述预先设置的访问成功率阈值且访问响应时间小于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 具有第二有效性；

当所述具有第一有效性的代理 IP 的访问成功率小于或等于所述预先设置的访问成功率阈值，或访问响应时间大于或等于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 不具有第二有效性。

12. 如权利要求 10 所述的终端，其特征在于，所述将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中包括：

将确定为具有第一有效性但不具有第二有效性的代理 IP 记录于所述黑名单中的第一黑名单子列表中，将确定为不具有第一有效性的代理 IP 记录于所述黑名单中的第二黑名单子列表中；

使用不具有第一有效性的代理 IP 多次访问多个搜索引擎，计算不具有第一有效性的代理 IP 的访问失败率；

判断不具有第一有效性的代理 IP 的访问失败率是否小于预先设置的访问失败率；

当所述不具有第一有效性的代理 IP 的访问失败率小于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为暂时无效的代理 IP，并记录于所述第一黑名单子列表中；或

当所述不具有第一有效性的代理 IP 的访问失败率大于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为永久无效的代理 IP，并

记录于所述第二黑名单子列表中。

13. 如权利要求 9 至 12 任意一项所述的终端，其特征在于，所述处理器执行所述计算机可读指令时还实现如下步骤：

根据代理 IP 的访问成功率及访问响应时间，对代理 IP 设置多个有效级别，将所述多个有效级别及其对应的代理 IP 记录于所述白名单中。

14. 如权利要求 13 所述的终端，其特征在于，所述对代理 IP 设置多个有效级别包括：

将访问成功率大于预先设置的第一访问成功率及访问响应时间大于预先设置的第一访问响应时间对应的代理 IP 作为第一有效级别的代理 IP；

将访问成功率小于预先设置的第一访问成功率但大于预先设置的第二访问成功率阈值，访问响应时间小于预先设置的第一访问响应时间但大于预先设置的第二访问响应对应的代理 IP 作为第二有效级别的代理 IP；

将访问成功率小于预先设置的第二访问成功率及访问响应时间小于预先设置的第二访问响应时间对应的代理 IP 作为第三有效级别的代理 IP。

15. 一种非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，其特征在于，所述计算机可读指令被处理器执行时实现如下步骤：

将每隔预设时间段获取的多个代理 IP 存储于预先设置的代理 IP 池中；

对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性；

将确定为有效的代理 IP 记录于所述代理 IP 池中的白名单中，将确定为无效的代理 IP 记录于所述代理 IP 池中的黑名单中；

当侦测到当前的代理 IP 满足预先设置的代理替换条件时，从所述代理 IP 池中的所述白名单中选取出一个代理 IP；及

将选取出的代理 IP 作为新的代理 IP 进行数据爬取。

16. 如权利要求 15 所述的存储介质，其特征在于，所述对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 的有效性包括：

对所述代理 IP 池中的每个代理 IP 进行逐个验证，判断所获取的代理 IP 是否具有第一有效性；

根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性；

将确定为具有第二有效性的代理 IP 记录于所述代理 IP 池中的白名单中；

将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中。

17. 如权利要求 16 所述的存储介质，其特征在于，所述根据代理 IP 的访问成功率和访问响应时间判断具有所述第一有效性的代理 IP 是否具有第二有效性包括：

使用具有所述第一有效性的代理 IP 多次访问多个搜索引擎，计算所述具有第一有效性的代理 IP 的访问成功率及访问响应时间；

判断所述具有第一有效性的代理 IP 的访问成功率是否大于预先设置的访问成功率阈值，同时判断所述具有第一有效性的代理 IP 的访问响应时间是

否小于预先设置的访问响应时间阈值；

当所述具有第一有效性的代理 IP 的访问成功率大于所述预先设置的访问成功率阈值且访问响应时间小于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 具有第二有效性；

当所述具有第一有效性的代理 IP 的访问成功率小于或等于所述预先设置的访问成功率阈值，或访问响应时间大于或等于所述预先设置的访问响应时间阈值时，确定所述具有第一有效性的代理 IP 不具有第二有效性。

18. 如权利要求 16 所述的存储介质，其特征在于，所述将确定为不具有第一有效性及具有第一有效性但不具有第二有效性的代理 IP 记录于所述代理 IP 池中的黑名单中包括：

将确定为具有第一有效性但不具有第二有效性的代理 IP 记录于所述黑名单中的第一黑名单子列表中，将确定为不具有第一有效性的代理 IP 记录于所述黑名单中的第二黑名单子列表中；

使用不具有第一有效性的代理 IP 多次访问多个搜索引擎，计算不具有第一有效性的代理 IP 的访问失败率；

判断不具有第一有效性的代理 IP 的访问失败率是否小于预先设置的访问失败率；

当所述不具有第一有效性的代理 IP 的访问失败率小于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为暂时无效的代理 IP，并记录于所述第一黑名单子列表中；或

当所述不具有第一有效性的代理 IP 的访问失败率大于所述预先设置的访问失败率时，将对应的不具有第一有效性的代理 IP 确认为永久无效的代理 IP，并记录于所述第二黑名单子列表中。

19. 如权利要求 15 至 18 任意一项所述的存储介质，其特征在于，所述计算机可读指令被处理器执行时还实现如下步骤：

根据代理 IP 的访问成功率及访问响应时间，对代理 IP 设置多个有效级别，将所述多个有效级别及其对应的代理 IP 记录于所述白名单中。

20. 如权利要求 19 所述的存储介质，其特征在于，所述对代理 IP 设置多个有效级别包括：

将访问成功率大于预先设置的第一访问成功率及访问响应时间大于预先设置的第一访问响应时间对应的代理 IP 作为第一有效级别的代理 IP；

将访问成功率小于预先设置的第一访问成功率但大于预先设置的第二访问成功率阈值，访问响应时间小于预先设置的第一访问响应时间但大于预先设置的第二访问响应对应的代理 IP 作为第二有效级别的代理 IP；

将访问成功率小于预先设置的第二访问成功率及访问响应时间小于预先设置的第二访问响应时间对应的代理 IP 作为第三有效级别的代理 IP。

1/3

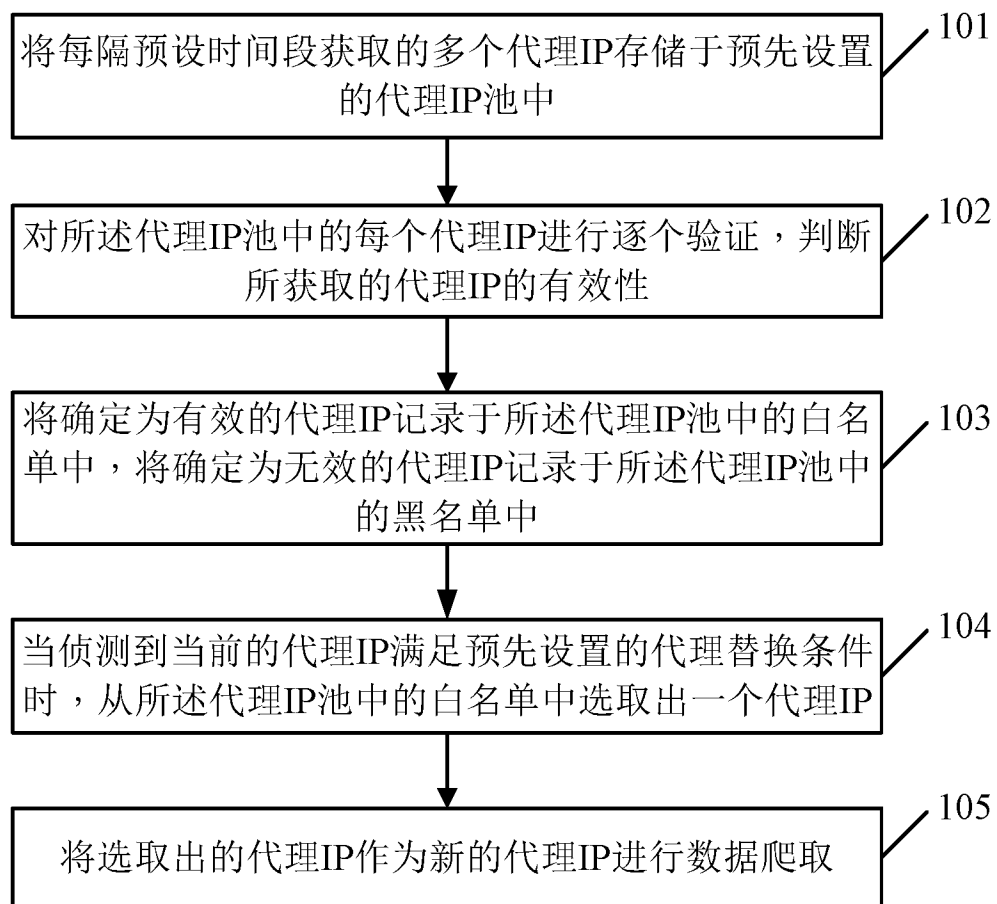


图 1

2/3

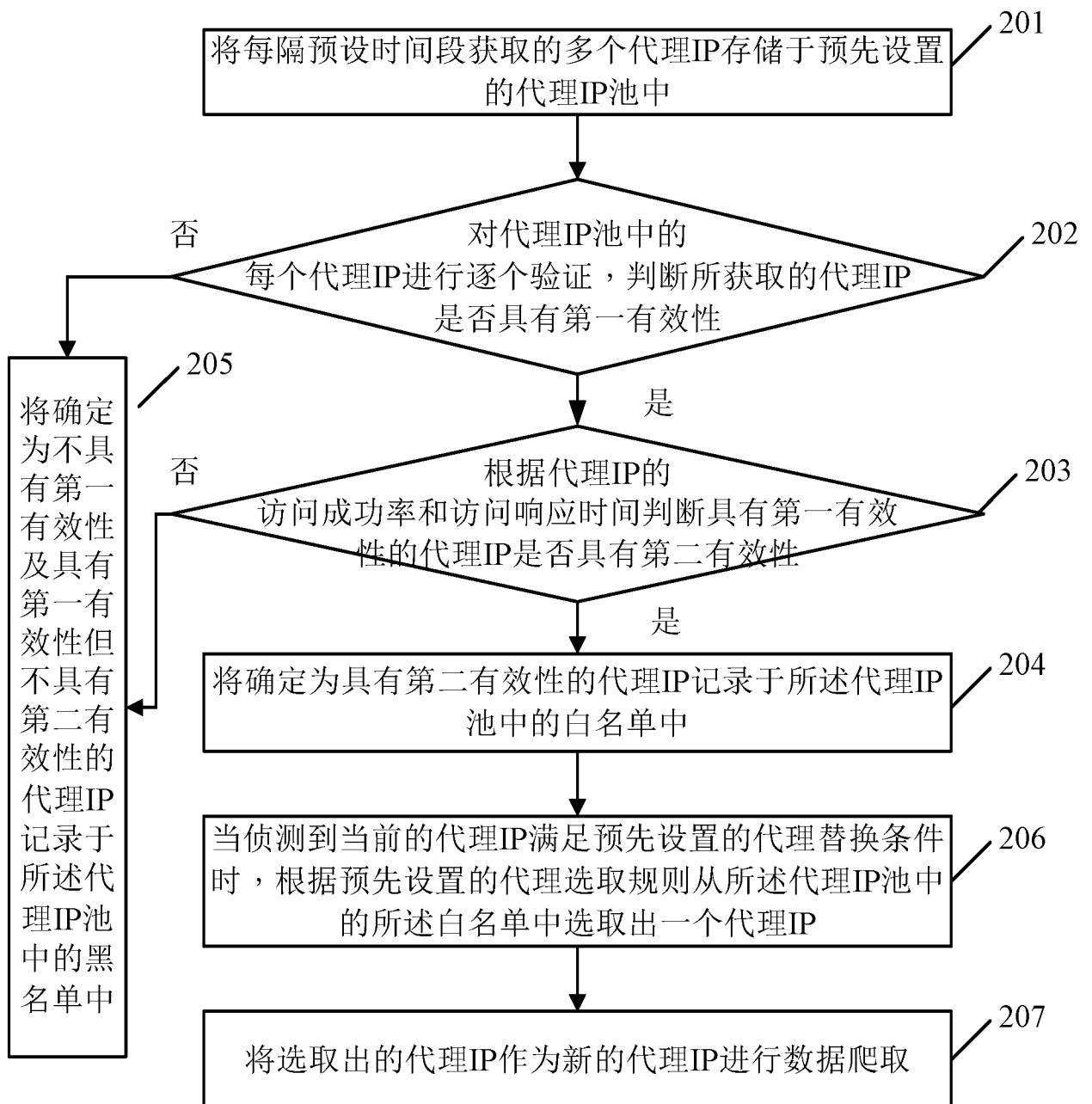


图 2

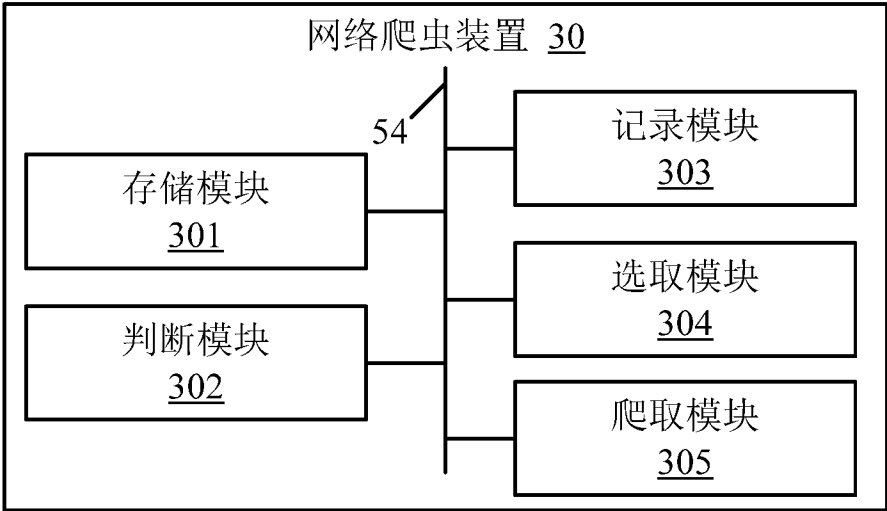


图 3

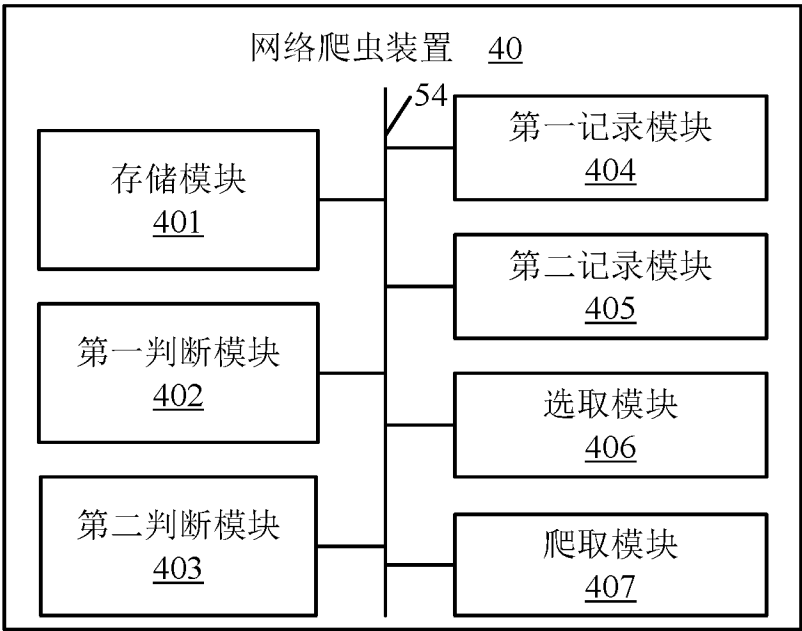


图 4

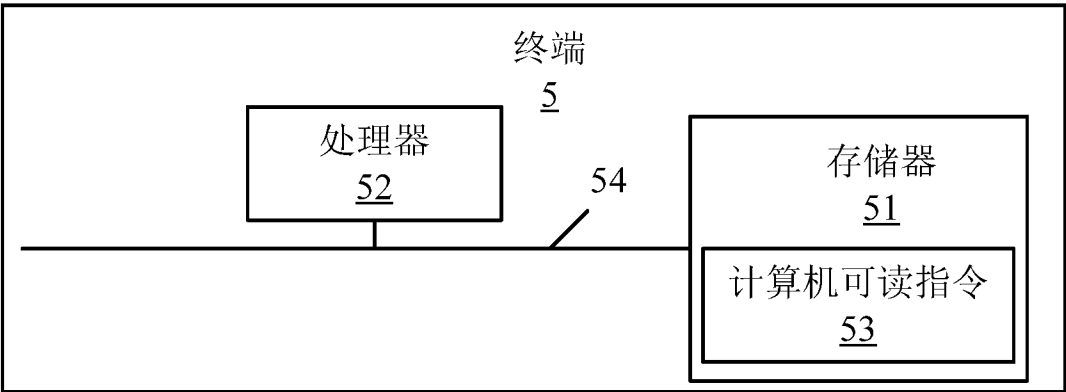


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/100162

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)i; H04L 29/06(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 爬虫, IP地址, 代理, 池, 集合, 黑名单, 白名单, 有效, 无效, 成功率, 响应, 时间, 时长, 时延, crawler, web spider, spider, web robot, ant, automatic indexer, search engine, IP address, proxy, pool, set, assembly, black list, white list, valid, invalid, success, rate, ratio, response, time, duration, delay

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106547793 A (BEIJING GRIDSUM TECHNOLOGY CO., LTD.) 29 March 2017 (2017-03-29) description, paragraphs 17-38	1, 7-9, 15
A	CN 106547793 A (BEIJING GRIDSUM TECHNOLOGY CO., LTD.) 29 March 2017 (2017-03-29) description, paragraphs 17-38	2-6, 10-14, 16-20
A	CN 105426415 A (TCL CORPORATION) 23 March 2016 (2016-03-23) entire document	1-20
A	CN 105827619 A (WUXI ZHONGKE FUNONG INTERNET OF THINGS TECHNOLOGY CO., LTD. ET AL.) 03 August 2016 (2016-08-03) entire document	1-20
A	US 2008183889 A1 (ANDREEV, D. ET AL.) 31 July 2008 (2008-07-31) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

10 December 2018

Date of mailing of the international search report

03 January 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/100162

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	106547793	A	29 March 2017	None	
CN	105426415	A	23 March 2016	None	
CN	105827619	A	03 August 2016	None	
US	2008183889	A1	31 July 2008	None	

国际检索报告

国际申请号

PCT/CN2018/100162

A. 主题的分类

G06F 17/30(2006.01)i; H04L 29/06(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F; H04L

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, CNKI, WPI, EPODOC; 爬虫, IP地址, 代理, 池, 集合, 黑名单, 白名单, 有效, 无效, 成功率, 响应, 时间, 时长, 时延, crawler, web spider, spider, web robot, ant, automatic indexer, search engine, IP address, proxy, pool, set, assembly, black list, white list, valid, invalid, success, rate, ratio, response, time, duration, delay

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 106547793 A (北京国双科技有限公司) 2017年 3月 29日 (2017 - 03 - 29) 说明书第17-38段	1, 7-9, 15
A	CN 106547793 A (北京国双科技有限公司) 2017年 3月 29日 (2017 - 03 - 29) 说明书第17-38段	2-6, 10-14, 16-20
A	CN 105426415 A (TCL集团股份有限公司) 2016年 3月 23日 (2016 - 03 - 23) 全文	1-20
A	CN 105827619 A (无锡中科富农物联网科技有限公司 等) 2016年 8月 3日 (2016 - 08 - 03) 全文	1-20
A	US 2008183889 A1 (ANDREEV, DMITRY 等) 2008年 7月 31日 (2008 - 07 - 31) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2018年 12月 10日

国际检索报告邮寄日期

2019年 1月 3日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

李普昕

电话号码 86-(10)-53961653

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/100162

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106547793	A	2017年 3月 29日	无	
CN	105426415	A	2016年 3月 23日	无	
CN	105827619	A	2016年 8月 3日	无	
US	2008183889	A1	2008年 7月 31日	无	