(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
2020 年 1 月 23 日 (23.01.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/015192 A1

(51) 国际专利分类号:
G06F 16/00 (2019.01)

(21) 国际申请号: PCT/CN2018/108218

(22) 国际申请日: 2018 年 9 月 28 日 (28.09.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810791126.3 2018年7月18日 (18.07.2018) 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区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

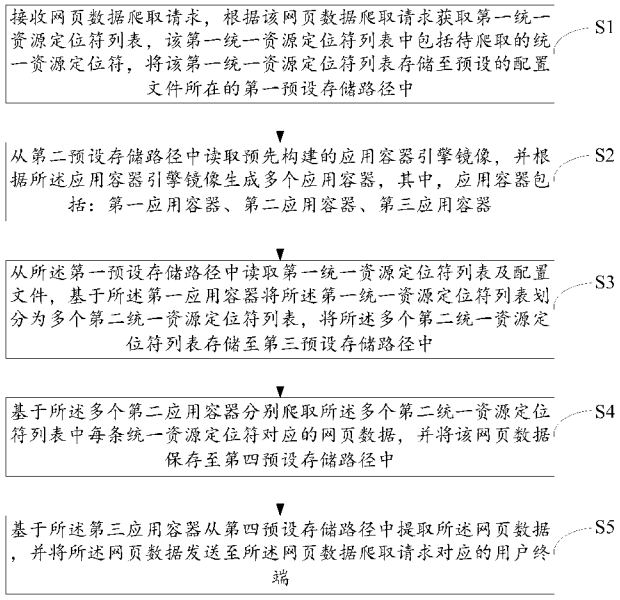
(72) 发明人: 吴壮伟 (WU, Zhuangwei); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所（普通合伙） (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路10号中浩大厦1528-1530室于志光, Guangdong 518000 (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, 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,

(54) **Title:** WEBPAGE DATA CRAWLING METHOD AND APPARATUS, AND STORAGE MEDIUM

(54) **发明名称**: 网页数据爬取方法、装置及存储介质



S1 Receive a webpage data crawling request, obtain, according to the webpage data crawling request, a first uniform resource location list that comprises a uniform resource location to be crawled, and store the first uniform resource location list in a first preset storage path where a configuration file is located

S2 Read a prebuilt application container engine mirror image from a second preset storage path, and generate a plurality of application containers according to the application container engine mirror image, wherein the application containers comprise a first application container, a second application container, and a third application container

S3 Read the first uniform resource location list and the configuration file from the first preset storage path, divide the first uniform resource location list into a plurality of second uniform resource location lists on the basis of the first application container, and store the plurality of second uniform resource location lists in a third preset storage path

S4 Respectively crawl, on the basis of the plurality of second uniform resource location lists, webpage data corresponding to each uniform resource location in the plurality of second uniform resource location lists, and store the webpage data in a fourth preset storage path

S5 Extract the webpage data from the fourth preset storage path on the basis of the third application container, and send the webpage data to a user terminal corresponding to the webpage data crawling request

图 1

(57) **Abstract:** The present application provides a webpage data crawling method. The method comprises: reading a first URL list according to a received webpage data crawling request; generating a plurality of application containers according to a prebuilt docker mirror image; dividing the read first URL list into a plurality of second URL lists, and respectively crawling webpage data corresponding to each URL in the plurality of second URL lists; and then, extracting the webpage data, and sending the webpage data to a user terminal corresponding to the webpage data crawling request. The present application also provides an electronic device and a computer storage medium. By utilizing the present application, the webpage data crawling efficiency can be improved.

WO 2020/015192 A1

[见续页]

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))。

(57) 摘要：本申请提供一种网页数据爬取方法，该方法根据接收的网页数据爬取请求读取第一URL列表，根据预先构建的docker镜像生成多个应用容器，将读取的所述第一URL列表划分为多个第二URL列表，并分别爬取所述多个第二URL列表中每条URL对应的网页数据，之后，该方法提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。本申请还提供一种电子装置及计算机存储介质。利用本申请，可以提高网页数据爬取效率。

网页数据爬取方法、装置及存储介质

本申请基于巴黎公约申明享有2018年7月18日递交的申请号为CN 201810791126.3、名称为“网页数据爬取方法、装置及存储介质”的中国专利申请的优先权，该中国专利申请的整体内容以参考的方式结合在本申请中。

技术领域

本申请涉及数据处理领域，尤其涉及一种网页数据爬取方法、电子装置及计算机可读存储介质。

背景技术

现有技术中，在一台服务器上运行多个任务的传统方式是启动多个虚拟机，然后在不同的虚拟机上运行不同的任务。传统的虚拟化技术多是采用基于VMware虚拟机，而VMware虚拟机的运行必须要运行整个操作系统，需要占用大量的系统资源。

然而，对于服务器来说，其CPU、内存、网络资源和磁盘资源等资源都是有限的。以网页数据爬取为例，目前爬虫的分布式受限于机器数量、CPU数量、线程个数和进程数量等，当服务器上启动的虚拟机过于消耗资源时，导致无法最大化利用系统资源，进而影响网页数据的爬取效率。

发明内容

鉴于以上内容，本申请提供一种网页数据爬取方法、服务器及计算机可读存储介质，其主要目的在于提高网页数据爬取效率。

为实现上述目的，本申请提供一种网页数据爬取方法，该方法包括：

S1、接收网页数据爬取请求，根据该网页数据爬取请求获取第一 URL（Uniform Resource Location，统一资源定位符）列表，该第一 URL 列表中包括待爬取的 URL，将该第一 URL 列表存储至预设的配置文件所在的第一预设存储路径中；

S2、从第二预设存储路径中读取预先构建的 docker 镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二

应用容器、第三应用容器；

S3、从所述第一预设存储路径中读取第一URL列表及配置文件，基于所述第一应用容器将所述第一URL列表划分为多个第二URL列表，将所述多个第二URL列表存储至第三预设存储路径中；

S4、基于所述多个第二应用容器分别爬取所述多个第二URL列表中每条URL对应的网页数据，并将该网页数据保存至第四预设存储路径中；及

S5、基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

此外，本申请还提供一种电子装置，其特征在于，该装置包括：存储器、处理器，所述存储器上存储有可在所述处理器上运行的网页数据爬取程序，所述网页数据爬取程序被所述处理器执行时，可实现如上所述网页数据爬取方法中的任意步骤。

此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质中包括网页数据爬取程序，所述网页数据爬取程序被处理器执行时，可实现如上所述网页数据爬取方法中的任意步骤。

本申请提出的网页数据爬取方法、电子装置及计算机可读存储介质，基于docker镜像建立docker应用容器来并行的进行数据处理，docker应用容器可以节省启动操作系统所带来的资源浪费，以进程级别的消耗来提供与虚拟机类似的隔离能力，基于这个框架，用户只需要设置好配置文件、将相关程序生成镜像文件，通过建立多个docker应用容器并行地爬取网页数据，即可高效地完成网页数据爬取工作；通过对爬取到的网页数据进行数据验证，保证爬取到的网页数据的完整性；通过对爬取到的网页数据进行子URL深度挖掘，保证网页数据的全面性。

附图说明

图1为本申请网页数据爬取方法较佳实施例的流程图；

图2为本申请电子装置较佳实施例的示意图；

图3为本申请图2中网页数据爬取程序的程序模块示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供一种网页数据爬取方法。参照图 1 所示，为本申请网页数据爬取方法较佳实施例的流程图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，基于 docker 的网页数据爬取方法包括步骤 S1-S5：

S1、接收网页数据爬取请求，根据该网页数据爬取请求获取第一 URL 列表，该第一 URL 列表中包括待爬取的 URL，将该第一 URL 列表存储至预设的配置文件所在的第一预设存储路径中；

以下内容以电子装置作为执行主体对本申请方法实施例进行说明，其中，电子装置作为服务器与用户终端建立通讯连接，并接收用户终端发送的业务数据处理请求，根据请求对业务数据进行处理。该电子装置可以具有多核 CPU（Central Processing Unit，中央处理器）。

可以理解的是，在接收用户终端发送的网页数据爬取请求并对网页数据进行爬取之前，在电子装置上配置好 docker 镜像。具体地，基于 dockerfile 规则创建 docker 镜像，该 docker 镜像中包含有列表划分程序、并行处理程序、数据验证程序、数据合并程序等，将创建的 docker 镜像保存至保存至第二预设存储路径中。在创建完 docker 镜像之后，基于该 docker 镜像创建多个应用容器。每一个程序可以独立地运行在应用容器中，多个应用容器之间的运行相互独立。

此外，在步骤 S1 之前，该方法还包括步骤：

接收客户端发送的配置参数，从配置参数中获取预先配置好并行进程的数量、网页数据及各程序的指定存储路径；

根据获取的并行进程的数量、各个程序的文件路径生成配置文件，将所述配置文件存储至第一预设存储路径中。

其中，上述进程的数量根据服务器的多核 CPU 的大小、以及数据处理需要占用的 CPU 情况来进行调整。

上述网页数据爬取请求中包含的第一 URL 列表为待爬取的原始 URL 列

表，在接收到网页数据爬取请求时，根据待爬取的第一 URL 列表中每条 URL 的信息，确定每条 URL 对应的索引值，将每条 URL 对应的索引值信息更新至上述第一 URL 列表中，然后将上述更新后的第一 URL 列表存储至第一预设存储路径中，其中，第一预设存储路径可以为 Redis 数据库。

作为一种实施方式，所述“第一 URL 列表中每条 URL 对应的索引值”通过以下步骤获取：

获取并分析第一 URL 列表中每条 URL 的具体信息，确定每条 URL 的特征信息；及

根据特征信息与索引值的映射关系，为第一 URL 列表中每条 URL 匹配对应的索引值。

其中，上述特征信息可以用于表征网页的类型，索引值用于调取爬虫程序。上述特征信息与索引值的映射关系通过以下步骤获取：

获取指定 URL 的集合，确定集合中每条 URL 的特征信息，为每条 URL 标注索引值；根据索引值将集合中的指定 URL 划分至不同索引值对应的子集合中；分别统计每个子集合中不同特征信息的占比，选择占比最大的特征信息作为各子集合中指定 URL 的目标特征信息；根据各子集合中指定 URL 的目标特征信息与该子集合对应的索引值，确定特征信息与索引值的映射关系。

进一步地，该方法还包括以下步骤：

当存在无法根据特征信息匹配索引值的 URL 时，基于该 URL 生成提示信息，并接收对该 URL 匹配索引值的匹配指令。

其中，上述匹配指令中包含上述无法根据特征信息匹配索引值的 URL 对应的索引值信息。将无法匹配索引值的 URL 导出，将导出的 URL 反馈至指定终端，人为地为导出的 URL 确定对应的索引值。

S2、从第二预设存储路径中读取预先构建的 docker 镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器；

在本实施例中，应用容器包括：一个第一应用容器、多个第二应用容器、一个第三应用容器。

S3、从所述第一预设存储路径中读取第一 URL 列表及配置文件，基于所述第一应用容器将所述第一 URL 列表划分为多个第二 URL 列表，将所述多

个第二 URL 列表存储至第三预设存储路径中；

具体地，运行所述第一应用容器，从第一预设存储路径中获取第一 URL 列表，从第二预设存储路径中调用上述列表划分程序，将第一 URL 列表平均地划分为 N 个 URL 子列表，即，N 个第二 URL 列表，N 为大于 1 的整数。其中，每一个第二 URL 列表中包含多条 URL 及每条 URL 对应的索引值信息，并将每一个第二 URL 列表存储至预设的第二 URL 列表的存储路径中，即，第三预设存储路径中。

在该步骤中，第二 URL 列表的数量 N 与第二应用容器的数量一致，例如，第二应用容器的数量为 5 时，N 为 5，表示将第一 URL 列表划分为 5 各第二 URL 列表。基于该步骤，实现待爬取列表及对应的程序索引值的划分及转移。

进一步地，第一应用容器从第二预设存储路径调用配置文件，将服务器的 CPU 资源分配给下述多个第二应用容器，供多个第二应用容器并行地执行网页数据爬取操作。从该配置文件中获取数据处理参数，其中，数据处理参数包括并行进程的数量 N 以及爬取到的网页数据的存储路径。

S4、基于所述多个第二应用容器分别爬取所述多个第二 URL 列表中每条 URL 对应的网页数据，并将该网页数据保存至第四预设存储路径中；

同步地运行所述 N 个第二应用容器，一个第二应用容器对应于的一个第二 URL 列表，所述多个第二应用容器分别从所述第三预设存储路径获取一个第二 URL 列表，根据每一个第二 URL 列表中每条 URL 对应的索引值，依次调用与索引值对应的爬虫程序，进行网页数据爬取操作，并将爬取到的网页数据保存至指定的存储路径中。其中，不同的第二应用容器对应同一个第四预设存储路径，即，不同第二应用内容器爬取到的网页数据保存至同一个文件夹中。通过该步骤，实现了网页数据的爬取及汇总过程。

S5、基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

运行所述第三应用容器，从第四预设存储路径中读取上述 N 个第二应用容器爬取到的网页数据，并将网页数据发送至上述网页数据爬取请求对应的用户终端，并生成提示信息。

在其他实施例中，为了保证爬取到的网页数据的完整性，需对爬取到的网页数据进行验证。具体地，在步骤 S5 之前，该方法还包括步骤：

对所述网页数据的数量及第一 URL 列表中 URL 的数量进行比对，确定第三 URL 列表；

当第三 URL 列表不为空时，针对第三 URL 列表中的每条 URL 执行网页数据爬取操作，直至第三 URL 列表为空为止，并将所述第三 URL 列表对应的网页数据保存至第四预设存储路径中；

当第三 URL 列表为空时，继续执行步骤 S5。

具体地，基于 docker 镜像生成第四应用容器，该第四应用容器对爬取到的网页数据进行验证。其中，第三 URL 列表中包含待爬取的 URL 及待爬取 URL 对应的索引值。可以理解的是，一条 URL 对应一条网页数据，假设统计得到第一 URL 列表中 URL 的数量为 P、第四预设存储路径中网页数据的数量为 Q，当 P=Q 时，说明第三 URL 列表为空，即不存在未处理的 URL，也就是说，不存在待爬取的网页数据；当 P＞Q 时，所述第三 URL 列表中 URL 的数量为 P-Q，P-Q 条 URL 为待处理的 URL，即存在待爬取的网页数据，对于包含待处理 URL 的第三 URL 列表，运行上述第四应用容器，执行网页数据爬取操作，并将该步骤爬取到的网页数据与通过步骤 S4 爬取到的网页数据进行合并。需要说明是，当对第三 URL 列表中的 URL 执行预设次数（例如，3 次）后，某 URL 仍然存在在第三 URL 列表中时，生成预警信息。通过该步骤，可以防止网页数据的遗漏，保证网页数据的完整性。

优选地，存在一种情况，爬取到的网页数据中包含子 URL，为了保证爬取到的网页数据的全面性，还可以对第一 URL 列表进行深度挖掘，确定网页数据中包含的子 URL 对应的网页数据。具体地，在步骤 S5 之前，该方法还包括步骤：

分别从第三预设存储路径中获取第二 URL 列表及 URL 正则挖掘程序；

从第四预设存储路径中获取各第二 URL 列表对应的网页数据，对各第二 URL 列表对应的网页数据进行挖掘，确定每个第二 URL 列表对应的第四 URL 列表，将所述第四 URL 列表保存至第五预设存储路径中；

对所述第四 URL 列表执行网页数据爬取操作，继续对第四 URL 列表对应的网页数据进行子 URL 挖掘操作，抽取新的子 URL，并进行网页数据爬取操作，以此循环。

每个第二 URL 列表对应一个第四 URL 列表，对第四 URL 列表执行网页

数据爬取操作的第二应用容器与该第四 URL 列表对应的第二 URL 列表的第二应用容器为同一个。

需要理解的是，为了防止一直深度挖掘新的子 URL 的情况，预先设置一个深度阈值（表示深度挖掘子 URL 的次数），当深度挖掘 URL 列表的次数超过预设的深度阈值时，停止挖掘新的子 URL 的操作。

上述实施例提出的网页数据爬取方法，在接收到网页数据爬取请求时，根据该请求获取待处理的第一 URL 列表，将其存储至预设的配置文件所在的第一预设存储路径中，从第二预设存储路径中读取预先构建的 docker 镜像，并根据该 docker 镜像生成多个应用容器，从上述第一预设存储路径中读取配置文件和第一 URL 列表，根据多个应用容器和配置文件将第一 URL 列表分割成多个第二 URL 列表，按照多容器并行处理的方式处理多个第二 URL 列表，系统资源会被服务器分配给并行处理的多个应用容器，爬取到各第二 URL 列表对应的网页数据，并将其发送至网页数据请求对应的用户终端。本申请的方案基于 docker 镜像建立 docker 应用容器来并行的进行数据处理，docker 应用容器可以节省启动操作系统所带来的资源浪费，以进程级别的消耗来提供与虚拟机类似的隔离能力，基于这个框架，用户只需要设置好配置文件、将相关程序生成镜像文件，通过建立多个 docker 应用容器并行地爬取网页数据，即可高效地完成网页数据爬取工作；通过对爬取到的网页数据进行数据验证，保证爬取到的网页数据的完整性；通过对爬取到的网页数据进行子 URL 深度挖掘，保证网页数据的全面性。

本申请还提供一种电子装置。参照图 2 所示，为本申请电子装置 1 较佳实施例的示意图。

在本实施例中，电子装置 1 可以是智能手机、平板电脑、便携计算机、桌上型计算机等具有数据处理功能的终端设备。

该电子装置 1 包括存储器 11、处理器 12，及网络接口 13。

其中，存储器 11 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、磁性存储器、磁盘、光盘等。存储器 11 在一些实施例中可以是所述电子装置 1 的内部存储单元，例如该电子装置 1 的硬盘。存储器 11 在另一些实施例中也

可以是所述电子装置 1 的外部存储设备，例如该电子装置 1 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。进一步地，存储器 11 还可以既包括该电子装置 1 的内部存储单元也包括外部存储设备。

存储器 11 不仅可以用于存储安装于该电子装置 1 的应用软件及各类数据，例如网页数据爬取程序 10 等，还可以用于暂时地存储已经输出或者将要输出的数据。

处理器 12 在一些实施例中可以是一中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器或其他数据处理芯片，用于运行存储器 11 中存储的程序代码或处理数据，例如网页数据爬取程序 10 等。

网络接口 13 可选的可以包括标准的有线接口、无线接口（如 WI-FI 接口），通常用于在该电子装置 1 与其他电子设备之间建立通信连接。

图 2 仅示出了具有组件 11-13 的电子装置 1，本领域技术人员可以理解的是，图 2 示出的结构并不构成对电子装置 1 的限定，可以包括比图示更少或者更多的部件，或者组合某些部件，或者不同的部件布置。

可选地，该电子装置 1 还可以包括用户接口，用户接口可以包括显示器（Display）、输入单元比如键盘（Keyboard），可选的用户接口还可以包括标准的有线接口、无线接口。

可选地，在一些实施例中，显示器可以是 LED 显示器、液晶显示器、触控式液晶显示器以及有机发光二极管（Organic Light-Emitting Diode，OLED）触摸器等。其中，显示器也可以称为显示屏或显示单元，用于显示在电子装置 1 中处理的信息以及用于显示可视化的用户界面。

在图 2 所示的电子装置 1 实施例中，作为一种计算机存储介质的存储器 11 中存储有网页数据爬取程序 10，处理器 12 执行存储器 11 中存储的网页数据爬取程序 10 时，实现如下步骤：

A1、接收网页数据爬取请求，根据该网页数据爬取请求获取第一 URL（Uniform Resource Location，统一资源定位符）列表，该第一 URL 列表中包括待爬取的 URL，将该第一 URL 列表存储至预设的配置文件所在的第一预设存储路径中；

以下内容以电子装置作为执行主体对本申请方法实施例进行说明，其中，电子装置作为服务器与用户终端建立通讯连接，并接收用户终端发送的业务数据处理请求，根据请求对业务数据进行处理。该电子装置可以具有多核 CPU（Central Processing Unit ，中央处理器）。

可以理解的是，在接收用户终端发送的网页数据爬取请求并对网页数据进行爬取之前，在电子装置上配置好 docker 镜像。具体地，基于 dockerfile 规则创建 docker 镜像，该 docker 镜像中包含有列表划分程序、并行处理程序、数据验证程序、数据合并程序等，将创建的 docker 镜像保存至保存至第二预设存储路径中。在创建完 docker 镜像之后，基于该 docker 镜像创建多个应用容器。每一个程序可以独立地运行在应用容器中，多个应用容器之间的运行相互独立。

此外，在步骤 A1 之前，该方法还包括步骤：

接收客户端发送的配置参数，从配置参数中获取预先配置好并行进程的数量、网页数据及各程序的指定存储路径；

根据获取的并行进程的数量、各个程序的文件路径生成配置文件，将所述配置文件存储至第一预设存储路径中。

其中，上述进程的数量根据服务器的多核 CPU 的大小、以及数据处理需要占用的 CPU 情况来进行调整。

上述网页数据爬取请求中包含的第一 URL 列表为待爬取的原始 URL 列表，在接收到网页数据爬取请求时，根据待爬取的第一 URL 列表中每条 URL 的信息，确定每条 URL 对应的索引值，将每条 URL 对应的索引值信息更新至上述第一 URL 列表中，然后将上述更新后的第一 URL 列表存储至第一预设存储路径中，其中，第一预设存储路径可以为 Redis 数据库。

作为一种实施方式，所述“第一 URL 列表中每条 URL 对应的索引值”通过以下步骤获取：

获取并分析第一 URL 列表中每条 URL 的具体信息，确定每条 URL 的特征信息；及

根据特征信息与索引值的映射关系，为第一 URL 列表中每条 URL 匹配对应的索引值。

其中，上述特征信息可以用于表征网页的类型，索引值用于调取爬虫程

序。上述特征信息与索引值的映射关系通过以下步骤获取：

获取指定 URL 的集合，确定集合中每条 URL 的特征信息，为每条 URL 标注索引值；根据索引值将集合中的指定 URL 划分至不同索引值对应的子集合中；分别统计每个子集合中不同特征信息的占比，选择占比最大的特征信息作为各子集合中指定 URL 的目标特征信息；根据各子集合中指定 URL 的目标特征信息与该子集合对应的索引值，确定特征信息与索引值的映射关系。

进一步地，该方法还包括以下步骤：

当存在无法根据特征信息匹配索引值的 URL 时，基于该 URL 生成提示信息，并接收对该 URL 匹配索引值的匹配指令。

其中，上述匹配指令中包含上述无法根据特征信息匹配索引值的 URL 对应的索引值信息。将无法匹配索引值的 URL 导出，将导出的 URL 反馈至指定终端，人为地为导出的 URL 确定对应的索引值。

A2、从第二预设存储路径中读取预先构建的 docker 镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器；

在本实施例中，应用容器包括：一个第一应用容器、多个第二应用容器、一个第三应用容器。

A3、从所述第一预设存储路径中读取第一 URL 列表及配置文件，基于所述第一应用容器将所述第一 URL 列表划分为多个第二 URL 列表，将所述多个第二 URL 列表存储至第三预设存储路径中；

具体地，运行所述第一应用容器，从第一预设存储路径中获取第一 URL 列表，从第二预设存储路径中调用上述列表划分程序，将第一 URL 列表平均地划分为 N 个 URL 子列表，即，N 个第二 URL 列表，N 为大于 1 的整数。其中，每一个第二 URL 列表中包含多条 URL 及每条 URL 对应的索引值信息，并将每一个第二 URL 列表存储至预设的第二 URL 列表的存储路径中，即，第三预设存储路径中。

在该步骤中，第二 URL 列表的数量与第二应用容器的数量一致，例如，第二应用容器的数量为 5 时，N 为 5，表示将第一 URL 列表划分为 5 各第二 URL 列表。基于该步骤，实现待爬取列表及对应的程序索引值的划分及转移。

进一步地，第一应用容器从第二预设存储路径调用配置文件，将服务器

的CPU资源分配给下述多个第二应用容器，供多个第二应用容器并行地执行网页数据爬取操作。从该配置文件中获取数据处理参数，其中，数据处理参数包括并行进程的数量N以及爬取到的网页数据的存储路径。

A4、基于所述多个第二应用容器分别爬取所述多个第二URL列表中每条URL对应的网页数据，并将该网页数据保存至第四预设存储路径中；

同步地运行所述N个第二应用容器，一个第二应用容器对应于的一个第二URL列表，所述多个第二应用容器分别从所述第三预设存储路径获取一个第二URL列表，根据每一个第二URL列表中每条URL对应的索引值，依次调用与索引值对应的爬虫程序，进行网页数据爬取操作，并将爬取到的网页数据保存至指定的存储路径中。其中，不同的第二应用容器对应同一个第四预设存储路径，即，不同第二应用内容器爬取到的网页数据保存至同一个文件夹中。通过该步骤，实现了网页数据的爬取及汇总过程。

A5、基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

运行所述第三应用容器，从第四预设存储路径中读取上述N个第二应用容器爬取到的网页数据，并将网页数据发送至上述网页数据爬取请求对应的用户终端，并生成提示信息。

在其他实施例中，为了保证爬取到的网页数据的完整性，需对爬取到的网页数据进行验证。具体地，所述网页数据爬取程序10被所述处理器执行时，在步骤A5之前，还实现如下步骤：

对所述网页数据的数量及第一URL列表中URL的数量进行比对，确定第三URL列表；

当第三URL列表不为空时，针对第三URL列表中的每条URL执行网页数据爬取操作，直至第三URL列表为空为止，并将所述第三URL列表对应的网页数据保存至第四预设存储路径中；

当第三URL列表为空时，继续执行步骤A5。

具体地，基于docker镜像生成第四应用容器，该第四应用容器对爬取到的网页数据进行验证。其中，第三URL列表中包含待爬取的URL及待爬取URL对应的索引值。可以理解的是，一条URL对应一条网页数据，假设统计得到第一URL列表中URL的数量为P、第四预设存储路径中网页数据的数量

为 Q，当 P=Q 时，说明第三 URL 列表为空，即不存在未处理的 URL，也就是说，不存在待爬取的网页数据；当 P＞Q 时，所述第三 URL 列表中 URL 的数量为 P-Q，P-Q 条 URL 为待处理的 URL，即存在待爬取的网页数据，对于包含待处理 URL 的第三 URL 列表，运行上述第四应用容器，执行网页数据爬取操作，并将该步骤爬取到的网页数据与通过步骤 A4 爬取到的网页数据进行合并。需要说明是，当对第三 URL 列表中的 URL 执行预设次数（例如，3 次）后，某 URL 仍然存在在第三 URL 列表中时，生成预警信息。通过该步骤，可以防止网页数据的遗漏，保证网页数据的完整性。

优选地，存在一种情况，爬取到的网页数据中包含子 URL，为了保证爬取到的网页数据的全面性，还可以对第一 URL 列表进行深度挖掘，确定网页数据中包含的子 URL 对应的网页数据。具体地，具体地，所述网页数据爬取程序 10 被所述处理器执行时，在步骤 A5 之前，还实现如下步骤：

分别从第三预设存储路径中获取第二 URL 列表及 URL 正则挖掘程序；

从第四预设存储路径中获取各第二 URL 列表对应的网页数据，对各第二 URL 列表对应的网页数据进行挖掘，确定每个第二 URL 列表对应的第四 URL 列表，将所述第四 URL 列表保存至第五预设存储路径中；

对所述第四 URL 列表执行网页数据爬取操作，继续对第四 URL 列表对应的网页数据进行子 URL 挖掘操作，抽取新的子 URL，并进行网页数据爬取操作，以此循环。

每个第二 URL 列表对应一个第四 URL 列表，对第四 URL 列表执行网页数据爬取操作的第二应用容器与该第四 URL 列表对应的第二 URL 列表的第二应用容器为同一个。

需要理解的是，为了防止一直深度挖掘新的子 URL 的情况，预先设置一个深度阈值（表示深度挖掘子 URL 的次数），当深度挖掘 URL 列表的次数超过预设的深度阈值时，停止挖掘新的子 URL 的操作。

上述实施例提出的电子装置 1，基于 docker 镜像建立 docker 应用容器来并行的进行数据处理，docker 应用容器可以节省启动操作系统所带来的资源浪费，以进程级别的消耗来提供与虚拟机类似的隔离能力，基于这个框架，用户只需要设置好配置文件、将相关程序生成镜像文件，通过建立多个 docker 应用容器并行地爬取网页数据，即可高效地完成网页数据爬取工作；通过对

爬取到的网页数据进行数据验证，保证爬取到的网页数据的完整性；通过对爬取到的网页数据进行子 URL 深度挖掘，保证网页数据的全面性。

可选地，在其他的实施例中，网页数据爬取程序 10 还可以被分割为一个或者多个模块，一个或者多个模块被存储于存储器 11 中，并由一个或多个处理器（本实施例为处理器 12）所执行，以完成本申请，本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段。例如，参照图 3 所示，为图 2 中网页数据爬取程序 10 的模块示意图，该实施例中，网页数据爬取程序 10 可以被分割为接收模块 110、容器生成模块 120、列表划分模块 130、数据爬取模块 140 及数据发送模块 150，所述模块 110-150 所实现的功能或操作步骤均与上文类似，此处不再详述，示例性地，例如其中：

接收模块 110，用于接收网页数据爬取请求，根据该网页数据爬取请求获取第一 URL 列表，该第一 URL 列表中包括待爬取的 URL，将该第一 URL 列表存储至预设的配置文件所在的第一预设存储路径中；

容器生成模块 120，用于从第二预设存储路径中读取预先构建的 docker 镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器；及

列表划分模块 130，用于从所述第一预设存储路径中读取第一 URL 列表及配置文件，基于所述第一应用容器将所述第一 URL 列表划分为多个第二 URL 列表，将所述多个第二 URL 列表存储至第三预设存储路径中；

数据爬取模块140，用于基于所述多个第二应用容器分别爬取所述多个第二URL列表中每条URL对应的网页数据，并将该网页数据保存至第四预设存储路径中；及

数据发送模块150，用于基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

此外，本申请实施例还提出一种计算机可读存储介质，所述计算机可读存储介质中包括网页数据爬取程序 10，所述网页数据爬取程序 10 被处理器执行时实现如下操作：

A1、接收网页数据爬取请求，根据该网页数据爬取请求获取第一URL列表，该第一URL列表中包括待爬取的URL，将该第一URL列表存储至预设的配置文件所在的第一预设存储路径中；

A2、从第二预设存储路径中读取预先构建的 docker 镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器；

A3、从所述第一预设存储路径中读取第一URL列表及配置文件，基于所述第一应用容器将所述第一URL列表划分为多个第二URL列表，将所述多个第二URL列表存储至第三预设存储路径中；

A4、基于所述多个第二应用容器分别爬取所述多个第二URL列表中每条URL对应的网页数据，并将该网页数据保存至第四预设存储路径中；及

A5、基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

本申请之计算机可读存储介质的具体实施方式与上述网页数据爬取方法的具体实施方式大致相同，在此不再赘述。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质（如ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，或者网络设备等）执行本申请各个实施例所述的方法。

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其它相关的技术领域，均同理包括在本申请的专利保护范围内。

5

权 利 要 求 书

1、一种网页数据爬取方法，应用于电子装置，其特征在于，所述方法包括：

S1、接收网页数据爬取请求，根据该网页数据爬取请求获取第一 URL（Uniform Resource Location，统一资源定位符）列表，该第一 URL 列表中包括待爬取的 URL，将该第一 URL 列表存储至预设的配置文件所在的第一预设存储路径中；

S2、从第二预设存储路径中读取预先构建的应用容器引擎镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器；

S3、从所述第一预设存储路径中读取第一 URL 列表及配置文件，基于所述第一应用容器将所述第一 URL 列表划分为多个第二 URL 列表，将所述多个第二 URL 列表存储至第三预设存储路径中；

S4、基于所述多个第二应用容器分别爬取所述多个第二 URL 列表中每条 URL 对应的网页数据，并将该网页数据保存至第四预设存储路径中；及

S5、基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

2、根据权利要求 1 所述的网页数据爬取方法，其特征在于，在步骤 S1 之前，该方法还包括：

接收客户端发送的配置参数，从配置参数中获取预先配置好并行进程的数量、网页数据及各程序的指定存储路径；及

根据获取的并行进程的数量、各个程序的文件路径生成配置文件，将所述配置文件存储至第一预设存储路径中。

3、根据权利要求 2 所述的网页数据爬取方法，其特征在于，在步骤 S5 之前，该方法还包括以下步骤：

对所述网页数据的数量及第一 URL 列表中 URL 的数量进行比对，确定第三 URL 列表；

当第三 URL 列表不为空时，针对第三 URL 列表中的每条 URL 执行网页数据爬取操作，直至第三 URL 列表为空为止，并将所述第三 URL 列表对应的网页数据保存至第四预设存储路径中；及

当第三 URL 列表为空时，继续执行步骤 S5。

4、根据权利要求 1 所述的网页数据爬取方法，其特征在于，所述第一 URL 列表中还包括每条 URL 对应的索引值信息，其中，所述“每条 URL 对应的索引值信息”通过以下步骤获取：

获取并分析第一 URL 列表中每条 URL 的具体信息，确定每条 URL 的特征信息；及

根据特征信息与索引值的映射关系，为第一 URL 列表中每条 URL 匹配对应的索引值。

5、根据权利要求 4 所述的网页数据爬取方法，其特征在于，该方法还包括以下步骤：

当存在无法根据特征信息匹配索引值的 URL 时，基于该 URL 生成提示信息，并接收对该 URL 匹配索引值的匹配指令。

6、根据权利要求 5 所述的网页数据爬取方法，其特征在于，在步骤 S5 之前，该方法还包括以下步骤：

对所述网页数据的数量及第一 URL 列表中 URL 的数量进行比对，确定第三 URL 列表；

当第三 URL 列表不为空时，针对第三 URL 列表中的每条 URL 执行网页数据爬取操作，直至第三 URL 列表为空为止，并将所述第三 URL 列表对应的网页数据保存至第四预设存储路径中；及

当第三 URL 列表为空时，继续执行步骤 S5。

7、根据权利要求 6 所述的网页数据爬取方法，其特征在于，在步骤 S5 之前，该方法还包括以下步骤：

分别从第三预设存储路径中获取第二 URL 列表及 URL 正则挖掘程序；

从第四预设存储路径中获取各第二 URL 列表对应的网页数据，对各第二 URL 列表对应的网页数据进行挖掘，确定每个第二 URL 列表对应的第四 URL 列表，将所述第四 URL 列表保存至第五预设存储路径中；及

对所述第四 URL 列表执行网页数据爬取操作，并对第四 URL 列表对应的网页数据进行子 URL 挖掘操作，抽取新的子 URL，并进行网页数据爬取操作，以此循环。

8、一种电子装置，其特征在于，该装置包括：存储器、处理器，所述存储器上存储有可在所述处理器上运行的网页数据爬取程序，所述网页数据爬

取程序被所述处理器执行时，可实现如下步骤：

A1、接收网页数据爬取请求，根据该网页数据爬取请求获取第一 URL 列表，该第一 URL 列表中包括待爬取的 URL，将该第一 URL 列表存储至预设的配置文件所在的第一预设存储路径中；

A2、从第二预设存储路径中读取预先构建的 docker 镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器；

A3、从所述第一预设存储路径中读取第一 URL 列表及配置文件，基于所述第一应用容器将所述第一 URL 列表划分为多个第二 URL 列表，将所述多个第二 URL 列表存储至第三预设存储路径中；

A4、基于所述多个第二应用容器分别爬取所述多个第二 URL 列表中每条 URL 对应的网页数据，并将该网页数据保存至第四预设存储路径中；及

A5、基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

9、根据权利要求 8 所述的电子装置，其特征在于，所述网页数据爬取程序被所述处理器执行时，在步骤 A1 之前，还实现如下步骤：

接收客户端发送的配置参数，从配置参数中获取预先配置好并行进程的数量、网页数据及各程序的指定存储路径；及

根据获取的并行进程的数量、各个程序的文件路径生成配置文件，将所述配置文件存储至第一预设存储路径中。

10、根据权利要求 9 所述的电子装置，其特征在于，所述网页数据爬取程序被所述处理器执行时，在步骤 A5 之前，还实现如下步骤：

对所述网页数据的数量及第一 URL 列表中 URL 的数量进行比对，确定第三 URL 列表；

当第三 URL 列表不为空时，针对第三 URL 列表中的每条 URL 执行网页数据爬取操作，直至第三 URL 列表为空为止，并将所述第三 URL 列表对应的网页数据保存至第四预设存储路径中；及

当第三 URL 列表为空时，继续执行步骤 A5。

11、根据权利要求 8 所述的电子装置，其特征在于，所述第一 URL 列表中还包括每条 URL 对应的索引值信息，其中，所述“每条 URL 对应的索引值

信息”通过以下步骤获取：

获取并分析第一 URL 列表中每条 URL 的具体信息，确定每条 URL 的特征信息；及

根据特征信息与索引值的映射关系，为第一 URL 列表中每条 URL 匹配对应的索引值。

12、根据权利要求 11 所述的电子装置，其特征在于，所述网页数据爬取程序被所述处理器执行时，还实现如下步骤：

当存在无法根据特征信息匹配索引值的 URL 时，基于该 URL 生成提示信息，并接收对该 URL 匹配索引值的匹配指令。

13、根据权利要求 12 所述的电子装置，其特征在于，所述网页数据爬取程序被所述处理器执行时，在步骤 A5 之前，还实现如下步骤：

对所述网页数据的数量及第一 URL 列表中 URL 的数量进行比对，确定第三 URL 列表；

当第三 URL 列表不为空时，针对第三 URL 列表中的每条 URL 执行网页数据爬取操作，直至第三 URL 列表为空为止，并将所述第三 URL 列表对应的网页数据保存至第四预设存储路径中；及

当第三 URL 列表为空时，继续执行步骤 S5。

14、根据权利要求 13 所述的电子装置，其特征在于，所述网页数据爬取程序被所述处理器执行时，在步骤 A5 之前，还实现如下步骤：

分别从第三预设存储路径中获取第二 URL 列表及 URL 正则挖掘程序；

从第四预设存储路径中获取各第二 URL 列表对应的网页数据，对各第二 URL 列表对应的网页数据进行挖掘，确定每个第二 URL 列表对应的第四 URL 列表，将所述第四 URL 列表保存至第五预设存储路径中；

对所述第四 URL 列表执行网页数据爬取操作，继续对第四 URL 列表对应的网页数据进行子 URL 挖掘操作，抽取新的子 URL，并进行网页数据爬取操作，以此循环。

15、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中包括网页数据爬取程序，所述网页数据爬取程序被处理器执行时，可实现如下步骤：

A1、接收网页数据爬取请求，根据该网页数据爬取请求获取第一 URL 列

表，该第一 URL 列表中包括待爬取的 URL，将该第一 URL 列表存储至预设的配置文件所在的第一预设存储路径中；

A2、从第二预设存储路径中读取预先构建的 docker 镜像，并根据所述 docker 镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器；

A3、从所述第一预设存储路径中读取第一 URL 列表及配置文件，基于所述第一应用容器将所述第一 URL 列表划分为多个第二 URL 列表，将所述多个第二 URL 列表存储至第三预设存储路径中；

A4、基于所述多个第二应用容器分别爬取所述多个第二 URL 列表中每条 URL 对应的网页数据，并将该网页数据保存至第四预设存储路径中；及

A5、基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端。

16、根据权利要求 15 所述的计算机可读存储介质，其特征在于，所述网页数据爬取程序被所述处理器执行时，在步骤 A1 之前，还实现如下步骤：

接收客户端发送的配置参数，从配置参数中获取预先配置好并行进程的数量、网页数据及各程序的指定存储路径；及

根据获取的并行进程的数量、各个程序的文件路径生成配置文件，将所述配置文件存储至第一预设存储路径中。

17、根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述第一 URL 列表中还包括每条 URL 对应的索引值信息，其中，所述“每条 URL 对应的索引值信息”通过以下步骤获取：

获取并分析第一 URL 列表中每条 URL 的具体信息，确定每条 URL 的特征信息；及

根据特征信息与索引值的映射关系，为第一 URL 列表中每条 URL 匹配对应的索引值。

18、根据权利要求 17 所述的计算机可读存储介质，其特征在于，所述网页数据爬取程序被所述处理器执行时，还实现如下步骤：

当存在无法根据特征信息匹配索引值的 URL 时，基于该 URL 生成提示信息，并接收对该 URL 匹配索引值的匹配指令。

19、根据权利要求 18 所述的计算机可读存储介质，其特征在于，所述网

页数据爬取程序被所述处理器执行时，在步骤 A5 之前，还实现如下步骤：

对所述网页数据的数量及第一 URL 列表中 URL 的数量进行比对，确定第三 URL 列表；

当第三 URL 列表不为空时，针对第三 URL 列表中的每条 URL 执行网页数据爬取操作，直至第三 URL 列表为空为止，并将所述第三 URL 列表对应的网页数据保存至第四预设存储路径中；及

当第三 URL 列表为空时，继续执行步骤 S5。

20、根据权利要求 19 所述的计算机可读存储介质，其特征在于，所述网页数据爬取程序被所述处理器执行时，在步骤 A5 之前，还实现如下步骤：

分别从第三预设存储路径中获取第二 URL 列表及 URL 正则挖掘程序；

从第四预设存储路径中获取各第二 URL 列表对应的网页数据，对各第二 URL 列表对应的网页数据进行挖掘，确定每个第二 URL 列表对应的第四 URL 列表，将所述第四 URL 列表保存至第五预设存储路径中；

对所述第四 URL 列表执行网页数据爬取操作，继续对第四 URL 列表对应的网页数据进行子 URL 挖掘操作，抽取新的子 URL，并进行网页数据爬取操作，以此循环。

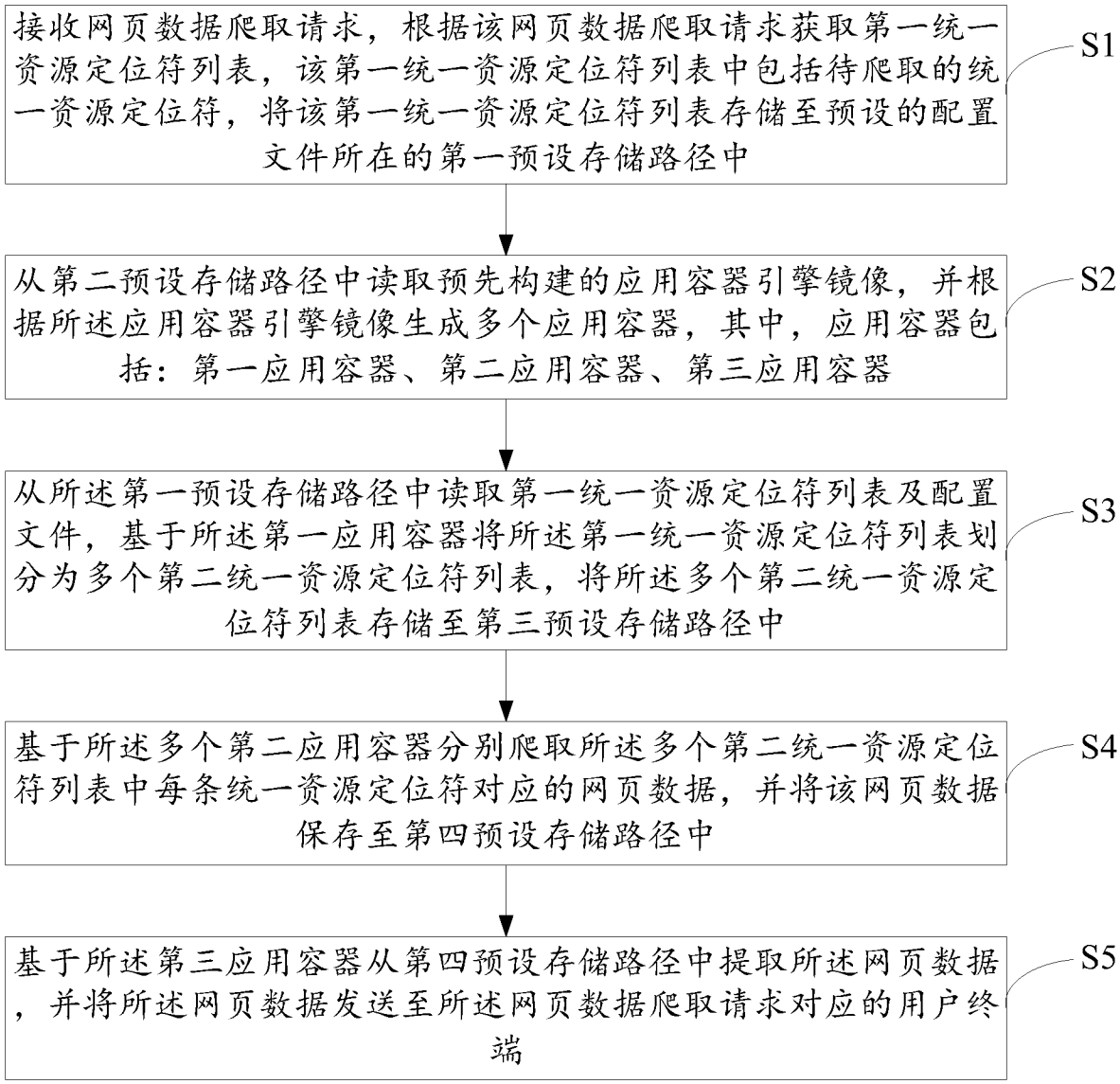

接收网页数据爬取请求，根据该网页数据爬取请求获取第一统一资源定位符列表，该第一统一资源定位符列表中包括待爬取的统一资源定位符，将该第一统一资源定位符列表存储至预设的配置文件所在的第一预设存储路径中
S1
从第二预设存储路径中读取预先构建的应用容器引擎镜像，并根据所述应用容器引擎镜像生成多个应用容器，其中，应用容器包括：第一应用容器、第二应用容器、第三应用容器
S2
从所述第一预设存储路径中读取第一统一资源定位符列表及配置文件，基于所述第一应用容器将所述第一统一资源定位符列表划分为多个第二统一资源定位符列表，将所述多个第二统一资源定位符列表存储至第三预设存储路径中
S3
基于所述多个第二应用容器分别爬取所述多个第二统一资源定位符列表中每条统一资源定位符对应的网页数据，并将该网页数据保存至第四预设存储路径中
S4
基于所述第三应用容器从第四预设存储路径中提取所述网页数据，并将所述网页数据发送至所述网页数据爬取请求对应的用户终端
S5

图 1

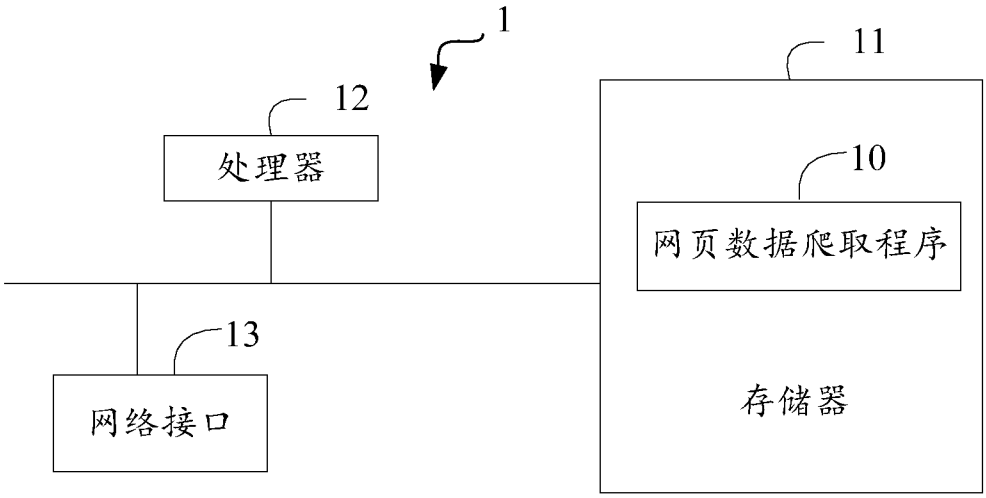

1
11
12
处理器
10
网页数据爬取程序
13
网络接口
存储器

图 2


网页数据爬取程序 10
接收模块 110
容器生成模块 120
列表划分模块 130
数据爬取模块 140
数据发送模块 150

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/108218

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/00(2019.01)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

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)

WPI, EPODOC, CNKI, CNPAT: 爬取, 爬虫, 镜像, 容器, 并行, 网页, 引擎, docker, image, container, crawling, web, URL, parallel

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106101176 A (CHENGDU SOBEY DIGITAL TECHNOLOGY CO., LTD.) 09 November 2016 (2016-11-09) description, paragraphs [0018]-[0027]	1-20
A	CN 108062413 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 22 May 2018 (2018-05-22) entire document	1-20
A	CN 106484886 A (KINGDEE SOFTWARE (CHINA) CO., LTD.) 08 March 2017 (2017-03-08) entire document	1-20
A	CN 108197633 A (BAINIAN JINHAI SCIENCE & TECHNOLOGY CO., LTD.) 22 June 2018 (2018-06-22) entire document	1-20
A	赵乐乐 (ZHAO, Lele). "基于Docker技术的全文搜索引擎的研究与应用 (Research and Application of Full-Text Search Engine Based on Docker Technology)" *中国优秀硕士学位论文全文数据库 信息科技辑 (Chinese Master's Theses Full-text Database, Information Science and Technology),* 15 February 2018 (2018-02-15), 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	Date of mailing of the international search report
02 April 2019	**22 April 2019**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **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/CN2018/108218

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 106101176 A	09 November 2016	None	
CN 108062413 A	22 May 2018	None	
CN 106484886 A	08 March 2017	None	
CN 108197633 A	22 June 2018	None	

国际检索报告

国际申请号
PCT/CN2018/108218

A. 主题的分类

G06F 16/00(2019.01)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

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））

WPI, EPODOC, CNKI, CNPAT: 爬取，爬虫，镜像，容器，并行，网页，引擎，docker, image, container, crawling, web, URL, parallel

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
A	CN 106101176 A (成都索贝数码科技股份有限公司) 2016年 11月 9日 (2016-11-09) 说明书第[0018]-[0027]段	1-20
A	CN 108062413 A (平安科技深圳有限公司) 2018年 5月 22日 (2018-05-22) 全文	1-20
A	CN 106484886 A (金蝶软件中国有限公司) 2017年 3月 8日 (2017-03-08) 全文	1-20
A	CN 108197633 A (百年金海科技有限公司) 2018年 6月 22日 (2018-06-22) 全文	1-20
A	赵乐乐. "基于Docker技术的全文搜索引擎的研究与应用" 中国优秀硕士学位论文全文数据库 信息科技辑, 2018年 2月 15日 (2018-02-15), 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	国际检索报告邮寄日期
2019年 4月 2日	2019年 4月 22日

ISA/CN的名称和邮寄地址	受权官员
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	苏宁
传真号 (86-10)62019451	电话号码 86-(10)-53961759

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/108218

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 106101176 A	2016年 11月 9日	无	
CN 108062413 A	2018年 5月 22日	无	
CN 106484886 A	2017年 3月 8日	无	
CN 108197633 A	2018年 6月 22日	无	