

权利要求书

- 1、一种网页数据处理方法，所述方法包括：
获取第一网页的第一网页数据，查询与所述第一网页数据关联的第二网页地址；
从所述第二网页地址中获取所述第二网页对应的网站的域名，提取所述第二网页对应的网站的域名的后缀；
- 5 当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时，则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址；
根据所述第二网页的网络地址对所述第二网页进行访问，并爬取所述第二网页上的第二网页数据；及
将所述第一网页数据与所述第二网页数据分别输出至对应的类别。
- 10 2、根据权利要求1所述的方法，其特征在于，所述根据所述第二网页的网络地址对所述第二网页进行访问，并爬取所述第二网页上的第二网页数据的步骤，包括：
当所述第二网页携带有限制访问标识时，则向代理服务器发送爬取所述第二网页上的网页数据的爬取指令；
接收所述代理服务器返回的身份验证请求，并根据所述身份验证请求向所述代理服务
- 15 器发送对应的身份标识；及
当所述身份标识经所述代理服务器验证成功时，则接收所述代理服务器返回的从所述第二网页上爬取到的网页数据。
- 3、根据权利要求1所述的方法，其特征在于，所述根据所述第二网页的网络地址对所述网页进行访问，并爬取所述第二网页上的第二网页数据的步骤，包括：
- 20 当所述第二网页未携带有限制访问标识时，则根据所述第二网页地址，获取所述第二网页对应的爬取逻辑与通信协议；
根据所述第二网页对应的通信协议访问所述第二网页并遍历所述第二网页的第二网页数据；及
当遍历到与所述爬取逻辑对应的第二网页数据时，则爬取与所述爬取逻辑对应的第二
- 25 网页数据。
- 4、根据权利要求1所述的方法，其特征在于，所述将所述第一网页数据与所述第二网页数据分别输出至对应的类别的步骤，包括：
将所述第一网页数据携带的网页标识以及所述第二网页数据携带的网页标识分别与已存取的网页标识进行匹配；
- 30 当所述第一网页数据携带的网页标识以及所述第二网页数据携带的网页标识其中至少一个与已存储的网页标识不匹配时，则提取不匹配的网页数据的关键字；及
将不匹配的网页数据输出至与所述关键字对应的存储类别下。
- 5、根据权利要求4所述的方法，其特征在于，所述方法还包括：
获取预设的接收所述第一网页数据与所述第二网页数据的邮箱地址；

提取所述邮箱地址对应的部门标识，并获取与所述部门标识对应的存储类别；及
将所获取的所述存储类别下的第一网页数据与第二网页数据发送至所述邮箱地址对应的邮箱中。

6、根据权利要求 1 至 5 任一项所述的方法，其特征在于，所述根据所述第二网页的网络地址对所述网页进行访问，并爬取所述第二网页上的第二网页数据的步骤，包括：
预设爬取所述第二网页的第二网页数据的爬取时间；
当到达所述爬取时间时，则从网络地址库中随机选取可用的爬取网络地址；及
通过所述爬取网络地址访问所述第二网页，并爬取所述第二网页上的第二网页数据。

7、根据权利要求 1 至 5 任一项的方法，其特征在于，所述根据所述第二网页的网络地址对所述第二网页进行访问，并爬取所述第二网页上的第二网页数据的步骤，包括：
根据所述第二网页的网络地址访问所述第二网页并查询所述第二网页是否渲染完成；
当所述第二网页未渲染完成时，则根据所述第二网页地址获取所述第二网页对应的渲染逻辑；

根据所述第二网页对应的渲染逻辑对所述第二网页进行渲染；及
爬取渲染完成的所述第二网页上的第二网页数据。

8、一种网页数据处理装置，其特征在于，包括：
查询模块，用于获取第一网页的第一网页数据，查询与所述第一网页数据关联的第二网页地址；

提取模块，用于从所述第二网页地址中获取所述第二网页对应的网站的域名，提取所述第二网页对应的网站的域名的后缀；

获取模块，用于当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时，则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址；

爬取模块，用于根据所述第二网页的网络地址对所述第二网页进行访问，并爬取所述第二网页上的第二网页数据；及

输出模块，将所述第一网页数据与所述第二网页数据分别输出至对应的类别。

9、根据权利要求 8 所述的装置，其特征在于，所述爬取模块包括：

发送单元，用于当所述第二网页携带有限制访问标识时，则向代理服务器发送爬取所述第二网页上的网页数据的爬取指令；

第一接收单元，用于接收所述代理服务器返回的身份验证请求，并根据所述身份验证请求向所述代理服务器发送对应的身份标识；及

第二接收单元，用于当所述身份标识经所述代理服务器验证成功时，则接收所述代理服务器返回的从所述第二网页上爬取到的网页数据。

10、一种计算机设备，包括存储器和处理器，所述存储器存储有计算机可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现以下步骤：

获取第一网页的第一网页数据，查询与所述第一网页数据关联的第二网页地址；

从所述第二网页地址中获取所述第二网页对应的网站的域名,提取所述第二网页对应的网站的域名的后缀;

当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时,则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址;

5 根据所述第二网页的网络地址对所述第二网页进行访问,并爬取所述第二网页上的第二网页数据;及

将所述第一网页数据与所述第二网页数据分别输出至对应的类别。

11、根据权利要求 10 所述的计算机设备,其特征在于,所述处理器执行所述计算机可读指令时所实现的所述根据所述第二网页的网络地址对所述第二网页进行访问,并爬取
10 所述第二网页上的第二网页数据的步骤,包括:

当所述第二网页携带有限制访问标识时,则向代理服务器发送爬取所述第二网页上的网页数据的爬取指令;

接收所述代理服务器返回的身份验证请求,并根据所述身份验证请求向所述代理服务器发送对应的身份标识;及

15 当所述身份标识经所述代理服务器验证成功时,则接收所述代理服务器返回的从所述第二网页上爬取到的网页数据。

12、根据权利要求 10 所述的计算机设备,其特征在于,所述处理器执行所述计算机可读指令时所实现的所述根据所述第二网页的网络地址对所述网页进行访问,并爬取所述
第二网页上的第二网页数据的步骤,包括:

20 当所述第二网页未携带有限制访问标识时,则根据所述第二网页地址,获取所述第二网页对应的爬取逻辑与通信协议;

根据所述第二网页对应的通信协议访问所述第二网页并遍历所述第二网页的第二网页数据;及

25 当遍历到与所述爬取逻辑对应的第二网页数据时,则爬取与所述爬取逻辑对应的第二网页数据。

13、根据权利要求 10 所述的计算机设备,其特征在于,所述处理器执行所述计算机可读指令时所实现的所述将所述第一网页数据与所述第二网页数据分别输出至对应的类别的步骤,包括:将所述第一网页数据携带的网页标识以及所述第二网页数据携带的网页标识分别与已存取的网页标识进行匹配;

30 当所述第一网页数据携带的网页标识以及所述第二网页数据携带的网页标识其中至少一个与已存储的网页标识不匹配时,则提取不匹配的网页数据的关键字;及

将不匹配的网页数据输出至与所述关键字对应的存储类别下。

14、根据权利要求 10 所述的计算机设备,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以下步骤:

35 获取预设的接收所述第一网页数据与所述第二网页数据的邮箱地址;

提取所述邮箱地址对应的部门标识, 并获取与所述部门标识对应的存储类别; 及
将所获取的所述存储类别下的第一网页数据与第二网页数据发送至所述邮箱地址对应的邮箱中。

15 15、根据权利要求 10 至 15 任一项所述的计算机设备, 其特征在于, 所述处理器执行所述计算机可读指令时所实现的所述根据所述第二网页的网络地址对所述网页进行访问, 并爬取所述第二网页上的第二网页数据的步骤, 包括:

预设爬取所述第二网页的第二网页数据的爬取时间;

当到达所述爬取时间时, 则从网络地址库中随机选取可用的爬取网络地址; 及

通过所述爬取网络地址访问所述第二网页, 并爬取所述第二网页上的第二网页数据。

10 16、根据权利要求 10 至 15 任一项所述的计算机设备, 其特征在于, 所述处理器执行所述计算机可读指令时所实现的所述根据所述第二网页的网络地址对所述第二网页进行访问, 并爬取所述第二网页上的第二网页数据的步骤, 包括:

根据所述第二网页的网络地址访问所述第二网页并查询所述第二网页是否渲染完成;

15 当所述第二网页未渲染完成时, 则根据所述第二网页地址获取所述第二网页对应的渲染逻辑;

根据所述第二网页对应的渲染逻辑对所述第二网页进行渲染; 及

爬取渲染完成的所述第二网页上的第二网页数据。

17、一个或多个存储有计算机可读指令的非易失性计算机可读存储介质, 其特征在于, 所述计算机可读指令被一个或多个处理器执行时, 使得一个或多个处理器执行以下步骤:

20 获取第一网页的第一网页数据, 查询与所述第一网页数据关联的第二网页地址;

从所述第二网页地址中获取所述第二网页对应的网站的域名, 提取所述第二网页对应的网站的域名的后缀;

当所述第二网页对应的网站的域名的后缀与预存的标准域名的后缀相同时, 则获取与所述标准域名对应的网络地址作为所述第二网页的网络地址;

25 根据所述第二网页的网络地址对所述第二网页进行访问, 并爬取所述第二网页上的第二网页数据; 及

将所述第一网页数据与所述第二网页数据分别输出至对应的类别。

18、根据权利要求 17 所述的存储介质, 其特征在于, 所述计算机可读指令被一个或多个处理器执行时, 使得一个或多个处理器执行所述根据所述第二网页的网络地址对所述第二网页进行访问, 并爬取所述第二网页上的第二网页数据的步骤, 包括:

当所述第二网页携带有限制访问标识时, 则向代理服务器发送爬取所述第二网页上的网页数据的爬取指令;

接收所述代理服务器返回的身份验证请求, 并根据所述身份验证请求向所述代理服务器发送对应的身份标识; 及

35 当所述身份标识经所述代理服务器验证成功时, 则接收所述代理服务器返回的从所述

第二网页上爬取到的网页数据。

19、根据权利要求 17 所述的存储介质，其特征在于，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行所述根据所述第二网页的网络地址对所述网页进行访问，并爬取所述第二网页上的第二网页数据的步骤，包括：

- 5 当所述第二网页未携带有限制访问标识时，则根据所述第二网页地址，获取所述第二网页对应的爬取逻辑与通信协议；

 根据所述第二网页对应的通信协议访问所述第二网页并遍历所述第二网页的第二网页数据；及

- 10 当遍历到与所述爬取逻辑对应的第二网页数据时，则爬取与所述爬取逻辑对应的第二网页数据。

20、根据权利要求 17 所述的存储介质，其特征在于，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行所述将所述第一网页数据与所述第二网页数据分别输出至对应的类别的步骤，包括：

- 15 将所述第一网页数据携带的网页标识以及所述第二网页数据携带的网页标识分别与已存储的网页标识进行匹配；

 当所述第一网页数据携带的网页标识以及所述第二网页数据携带的网页标识其中至少一个与已存储的网页标识不匹配时，则提取不匹配的网页数据的关键字；及

 将不匹配的网页数据输出至与所述关键字对应的存储类别下。