(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115225719 A
(43)申请公布日 2022.10.21

(21)申请号 202211060355.0
(22)申请日 2022.08.31
(71)申请人 中建电子商务有限责任公司
地址 610000 四川省成都市青羊区腾飞大道51号18栋12层1202号
申请人 云筑信息科技(成都)有限公司
(72)发明人 杨鑫 袁海涛
(74)专利代理机构 成都春夏知识产权代理事务所(特殊普通合伙) 51317
专利代理师 夏琴
(51)Int.Cl.
H04L 67/60(2022.01)
H04L 67/02(2022.01)
H04L 67/1097(2022.01)
H04L 69/22(2022.01)
H04L 41/0246(2022.01)

权利要求书2页 说明书4页 附图1页

(54)发明名称

一种分布式定向网络数据采集解析方法

(57)摘要

本发明公开了一种分布式定向网络数据采集解析方法,包括启动分布式数据采集节点并建立其与中央控制器通信;中央控制器获取定向网络数据采集的任务;中央控制器选取任务量较少的采集节点,将任务分发到采集节点上;采集节点接收任务并向对应目标网站发送请求,同时目标网站返回响应内容;采集节点读取对应的解析配置数据;数据解析模块获取响应内容和解析配置数据,解析任务数据包和子任务;重复步骤S1至步骤S6,直至最终任务执行完成。本发明通过分布式数据采集节点,使中央控制器可实时感知数据采集节点任务执行情况,合理任务调度,充分发挥分布式计算机协同数据采集的能力;针对每个定向采集任务,无需定制化编写计算机应用程序。

CN 115225719 A

1.一种分布式定向网络数据采集解析方法,其特征在于,包括以下步骤:

步骤S1、启动分布式数据采集节点、建立分布式数据采集节点与中央控制器通信;

步骤S2、拉取待采集任务,中央控制器获取定向网络数据采集的任务T1;

步骤S3、中央控制器从分布式数据采集节点的列表中选取任务量较少的采集节点N1,将任务T1分发到采集节点N1上;

步骤S4、采集节点N1接收任务T1,并向该任务所在的目标网站发送请求,同时目标网站返回响应内容C1;

步骤S5、采集节点N1从任务缓存数据库中读取任务T1对应的解析配置数据P10;

步骤S6、数据解析模块获取响应内容C1和解析配置数据P10,对任务T1数据包和子任务执行解析;

步骤S7、重复步骤S1至步骤S6递归循环采集,直至最终任务执行完成,采集结束。

2.根据权利要求1所述的一种分布式定向网络数据采集解析方法,其特征在于,所述任务T1包含目标网址URL地址、请求方式、请求头、请求参数、HTTP响应报文所对应的任务解析器唯一标识P1、数据包D1和任务标识F1。

3.根据权利要求2所述的一种分布式定向网络数据采集解析方法,其特征在于,所述步骤S5中,采集节点N1根据任务T1的任务解析器唯一标识P1,从任务缓存数据库中读取任务解析器唯一标识P1对应的解析配置数据P10。

4.根据权利要求3所述的一种分布式定向网络数据采集解析方法,其特征在于,所述解析配置数据P10包含数据包解析配置P11、子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13、子任务HTTP请求参数解析配置P14和子任务任务标识P15,所述子任务HTTP请求URL解析配置P12包含解析具备规律URL的循环规则。

5.根据权利要求4所述的一种分布式定向网络数据采集解析方法,其特征在于,所述数据包解析配置P11、子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13和子任务HTTP请求参数解析配置P14均包含字段名P101和规则集合P102,所述规则集合P102包含但不限于常量值、Selector选择器表达式、Xpath表达式、正则表达式、字符串截取表达式、字符串替换表达式和JavaScript脚本。

6.根据权利要求5所述的一种分布式定向网络数据采集解析方法,其特征在于,数据包解析方法为:获取数据包解析配置P11,按照其内部的规则集合P102的顺序依次对任务T1数据包执行解析,响应内容C1经规则集合P102的第一个规则执行后得到响应内容C2,响应内容C2经规则集合P102的第二个规则执行后得到响应内容C3,以此类推,直到执行完规则集合P102的最后一个规则得到Cn,将规则集合P102解析的结构化数据存放至数据包D1中。

7.根据权利要求6所述的一种分布式定向网络数据采集解析方法,其特征在于,子任务解析方法为:获取子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13和子任务HTTP请求参数解析配置P14,按照其内部的规则集合P102的顺序依次对任务T1子任务待解析值的变量值执行解析并替换其变量部分,解析后得到新的子任务T2、T3、...、Tn,同时赋予每个新的子任务对应的解析器唯一标识。

8.根据权利要求7所述的一种分布式定向网络数据采集解析方法,其特征在于,子任务解析包括HTTP请求地址解析、HTTP请求头解析、HTTP请求参数解析和最终任务标识标记。

9.根据权利要求7所述的一种分布式定向网络数据采集解析方法,其特征在于,所述步

骤S7为:重复步骤S1至步骤S6递归循环采集,直至任务T1不再生成子任务Tn,采集结束。

10.根据权利要求1所述的一种分布式定向网络数据采集解析方法,其特征在于,所述步骤S4中,采集节点N1按照任务T1的任务内容,组装标准的HTTP请求报文,并向目标网站发送请求。

一种分布式定向网络数据采集解析方法

技术领域

[0001] 本发明属于计算机技术领域，具体涉及一种分布式定向网络数据采集解析方法。

背景技术

[0002] 对于一些中小型企业，迫切需要掌握行业内相关业务数据，以此来对企业业务发展方向进行判断，做出优化决策，提高自身竞争力，让业务能够得到更好、更快速的发展。但随着互联网的发展，网络数据指数级增长，企业从庞大的网络数据中筛选出有价值的数据，时间成本、人力成本巨大，因此通过一种自动化进行大规模定向网络数据采集解析势在必行。

[0003] 现有网络数据采集解析常见的方式有：没有中央控制器的单节点任务采集单节点任务采集；对于采集后的原始数据，通过定制化编写特定的解析程序进行数据提炼；搜索引擎全网采集等。其中第一种方式的单机单节点任务采集，采集效率低；第二种方式针对每个特定的定向数据采集，都需要定制化编写计算机应用程序，对于一般用户而言门槛高，同时也不具备通用性；第三种方式的搜索引擎全网采集，采集任务广，不具备行业特性，同时仅采集网页源码，提供分词搜索，无法对数据进行解析提炼，对企业自身发展无参考价值。

[0004] 因此，本发明提供了一种分布式定向网络数据采集解析方法，以至少解决上述部分技术问题。

发明内容

[0005] 本发明要解决的技术问题是：提供一种分布式定向网络数据采集解析方法，以至少解决上述部分技术问题。

[0006] 为实现上述目的，本发明采用的技术方案如下：

一种分布式定向网络数据采集解析方法包括以下步骤：

步骤S1、启动分布式数据采集节点、建立分布式数据采集节点与中央控制器通信；

步骤S2、拉取待采集任务，中央控制器获取定向网络数据采集的任务T1；

步骤S3、中央控制器从分布式数据采集节点的列表中选取任务量较少的采集节点N1，将任务T1分发到采集节点N1上；

步骤S4、采集节点N1接收任务T1，并向该任务所在的目标网站发送请求，同时目标网站返回响应内容C1；

步骤S5、采集节点N1从任务缓存数据库中读取任务T1对应的解析配置数据P10；

步骤S6、数据解析模块获取响应内容C1和解析配置数据P10，对任务T1数据包和子任务执行解析；

步骤S7、重复步骤S1至步骤S6递归循环采集，直至最终任务执行完成，采集结束。

[0007] 进一步地，所述任务T1包含目标网址URL地址、请求方式、请求头、请求参数、HTTP响应报文所对应的任务解析器唯一标识P1、数据包D1和任务标识F1。

[0008] 进一步地，所述步骤S5中，采集节点N1根据任务T1的任务解析器唯一标识P1，从任

务缓存数据库中读取任务解析器唯一标识P1对应的解析配置数据P10。

[0009] 进一步地，所述解析配置数据P10包含数据包解析配置P11、子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13、子任务HTTP请求参数解析配置P14和子任务任务标识P15，所述子任务HTTP请求URL解析配置P12包含解析具备规律URL的循环规则。

[0010] 对于一些POST方式请求，请求URL地址相同，而分页页码在请求参数中，因此子任务HTTP请求参数解析配置P14也包含解析具备规律URL的循环规则。

[0011] 进一步地，所述数据包解析配置P11、子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13和子任务HTTP请求参数解析配置P14均包含字段名P101和规则集合P102，所述规则集合P102包含但不限于常量值、Selector选择器表达式、Xpath表达式、正则表达式、字符串截取表达式、字符串替换表达式和JavaScript脚本。

[0012] 进一步地，数据包解析方法为：获取数据包解析配置P11，按照其内部的规则集合P102的顺序依次对任务T1数据包执行解析，响应内容C1经规则集合P102的第一个规则执行后得到响应内容C2，响应内容C2经规则集合P102的第二个规则执行后得到响应内容C3，以此类推，直到执行完规则集合P102的最后一个规则得到Cn，将规则集合P102解析的结构化数据存放至数据包D1中。

[0013] 进一步地，子任务解析方法为：获取子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13和子任务HTTP请求参数解析配置P14，按照其内部的规则集合P102的顺序依次对任务T1子任务待解析值的变量值执行解析并替换其变量部分，解析后得到新的子任务T2、T3、...、Tn，同时赋予每个新的子任务对应的解析器唯一标识。

[0014] 进一步地，子任务解析包括HTTP请求地址解析、HTTP请求头解析、HTTP请求参数解析和最终任务标识标记。

[0015] 进一步地，所述步骤S7为：重复步骤S1至步骤S6递归循环采集，直至任务T1不再生成子任务Tn，采集结束。

[0016] 进一步地，所述步骤S4中，采集节点N1按照任务T1的任务内容，组装标准的HTTP请求报文，并向目标网站发送请求。

[0017] 与现有技术相比，本发明具有以下有益效果：

本发明具备协同性，通过分布式数据采集节点，使中央控制器可实时感知数据采集节点的任务执行情况，合理进行任务调度，充分发挥分布式计算机协同数据采集的能力；本发明具备通用性，针对每个定向采集任务，无需定制化编写计算机应用程序，即使不具备相关专业能力的用户，也能通过配置完成数据采集、数据提炼，提高效率。

附图说明

[0018] 图1为本发明方法流程图。

具体实施方式

[0019] 术语解释：

URL为统一资源定位系统；

HTTP为超文本传输协议；

JSON为是一种轻量级的数据交换格式；

DOM(Document Object Model)为文档对象模型。

[0020] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图，对本发明进一步详细说明。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。

[0021] 如图1所示，本发明提供的一种分布式定向网络数据采集解析方法，包括以下步骤：

步骤S1、启动分布式数据采集节点、建立分布式数据采集节点与中央控制器通信；

步骤S2、拉取待采集任务，中央控制器获取定向网络数据采集的任务T1；

步骤S3、中央控制器从分布式数据采集节点的列表中选取任务量较少的采集节点N1，将任务T1分发到采集节点N1上；

步骤S4、采集节点N1接收任务T1，并向该任务所在的目标网站发送请求，同时目标网站返回响应内容C1；

步骤S5、采集节点N1从任务缓存数据库中读取任务T1对应的解析配置数据P10；

步骤S6、数据解析模块获取响应内容C1和解析配置数据P10，对任务T1数据包和子任务执行解析；

步骤S7、重复步骤S1至步骤S6递归循环采集，直至最终任务执行完成，采集结束。

[0022] 本发明具备协同性，通过分布式数据采集节点，使中央控制器可实时感知数据采集节点的任务执行情况，合理进行任务调度，充分发挥分布式计算机协同数据采集的能力；本发明具备通用性，针对每个定向采集任务，无需定制化编写计算机应用程序，即使不具备相关专业能力的用户，也能通过配置完成数据采集、数据提炼，提高效率。

[0023] 所述步骤S1中，分布式数据采集节点启动后，与对应的中央控制器建立通信，如此，分布式数据采集节点自身任务的执行情况会同步给中央控制器，便于中央控制器做出任务分发决策。

[0024] 所述步骤S2中，任务仓储层的定时任务按照特定频率拉取待采集任务，中央中控制器获取到定向网络数据采集的任务T1。所述任务T1包含目标网址URL地址、请求方式、请求头、请求参数、HTTP响应报文所对应的任务解析器唯一标识P1、数据包D1和任务标识F1。所述数据包D1是由键(key)和值(value)组合而成的结构化数据，由后续解析操作的每层任务解析得到，并将随着任务向下层任务传递，直到最终任务解析完成后，将数据包D1写入非关系型数据库中。

[0025] 所述步骤S3中，中央控制器从本地缓存的分布式数据采集节点的列表中，根据每个节点任务执行情况，选取出任务量较少的采集节点N1，将任务T1分发到采集节点N1上。

[0026] 所述步骤S4中，采集节点N1接收到任务T1，按照任务内容，组装标准的HTTP请求报文，向该任务所在的目标网站发送请求，同时目标网站返回响应内容C1。

[0027] 所述步骤S5中，采集节点N1根据该任务T1的HTTP响应报文所对应的任务解析器唯一标识P1，从任务缓存数据库中读取任务T1对应的解析配置数据P10。所述解析配置数据P10是具备解析提炼同一类型网页源码的JSON格式数据，同一类型网页源码具备结构相同的DOM树，填充不同的特性内容，如商品详情页，对于不同的商品网页结构DOM树相同，商品内容不同，因此同一个解析配置，能够解析出不同的商品数据。所述解析配置可根据某一特

定的网页数据，基于DOM树解析技术，从网页中选出需要提取的字段，自动分析出特定字段所处DOM树中的位置。所述解析配置数据P10包含数据包解析配置P11、子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13、子任务HTTP请求参数解析配置P14和子任务任务标识P15，所述子任务HTTP请求URL解析配置P12包含解析具备规律URL的循环规则。所述数据包解析配置P11、子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13和子任务HTTP请求参数解析配置P14均包含字段名P101和规则集合P102；所述规则集合P102按一定顺序组合而成，包含但不限于常量值、Selector选择器表达式、Xpath表达式、正则表达式、字符串截取表达式、字符串替换表达式和JavaScript脚本。

[0028] 所述步骤S6中，数据解析模块获取响应内容C1和解析配置数据P10，对任务T1数据包和子任务执行解析。

[0029] 所述数据包解析方法为：获取数据包解析配置P11，按照其内部的规则集合P102的顺序依次对任务T1数据包执行解析，响应内容C1经规则集合P102的第一个规则执行后得到响应内容C2，响应内容C2经规则集合P102的第二个规则执行后得到响应内容C3，以此类推，直到执行完规则集合P102的最后一个规则得到Cn，将规则集合P102解析的结构化数据存放至数据包D1中。其中逐层解析出的键(key)为字段名P101、逐层解析出的值(value)为Cn。根据实际业务场景，若存在多个需要解析的数据，重复以上步骤，若任务标识F1为最终任务，数据包中所存放的数据即为由每一层任务追加生成得到的结构化数据，存放于非关系型数据库中，供消费者使用。

[0030] 所述子任务解析方法为：获取子任务HTTP请求URL解析配置P12、子任务HTTP请求头解析配置P13和子任务HTTP请求参数解析配置P14，按照其内部的规则集合P102的顺序依次对任务T1子任务待解析值的变量值执行解析并替换其变量部分，例如变量用${key}的方式表示，其中key表示变量名，解析后得到新的子任务T2、T3、...、Tn，同时赋予每个新的子任务对应的解析器唯一标识。若待解析值包含多个变量，按照数据包解析方法执行解析，解析完所有变量值并替换变量部分，得到完整的期望值。其中子任务HTTP请求URL解析配置P12包含解析具备规律URL的循环规则，可用于解析具备一定规律的URL，由此可批量生成多个URL地址。子任务解析包括HTTP请求地址解析、HTTP请求头解析、HTTP请求参数解析和最终任务标识标记。完成四个部分解析后，组合得到新的子任务T2、T3、...、Tn，同时赋予每个任务对应的解析器唯一标识。

[0031] 所述步骤S7中，子任务重复步骤S1至步骤S6递归循环采集，直至任务T1不再生成子任务Tn，即最终任务执行完成后，采集结束。

[0032] 最后应说明的是：以上各实施例仅仅为本发明的较优实施例用以说明本发明的技术方案，而非对其限制，当然更不是限制本发明的专利范围；尽管参照前述各实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的范围；也就是说，但凡在本发明的主体设计思想和精神上作出的毫无实质意义的改动或润色，其所解决的技术问题仍然与本发明一致的，均应当包含在本发明的保护范围之内；另外，将本发明的技术方案直接或间接的运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

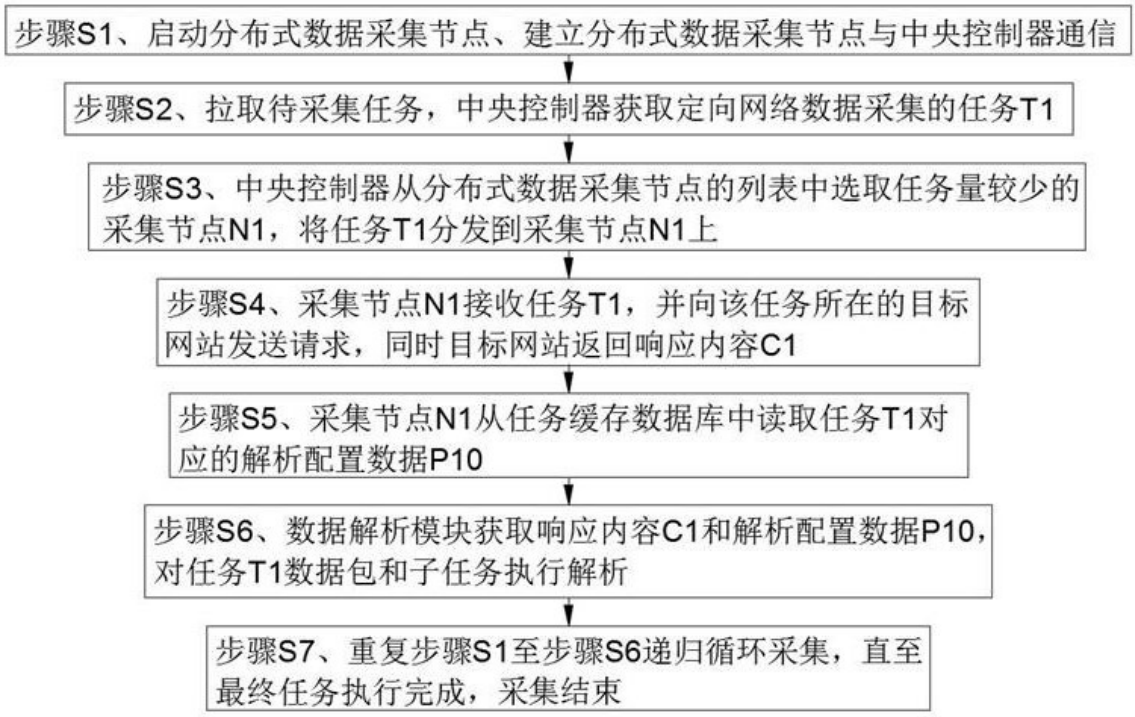
步骤S1、启动分布式数据采集节点、建立分布式数据采集节点与中央控制器通信
步骤S2、拉取待采集任务，中央控制器获取定向网络数据采集的任务T1
步骤S3、中央控制器从分布式数据采集节点的列表中选取任务量较少的采集节点N1，将任务T1分发到采集节点N1上
步骤S4、采集节点N1接收任务T1，并向该任务所在的目标网站发送请求，同时目标网站返回响应内容C1
步骤S5、采集节点N1从任务缓存数据库中读取任务T1对应的解析配置数据P10
步骤S6、数据解析模块获取响应内容C1和解析配置数据P10，对任务T1数据包和子任务执行解析
步骤S7、重复步骤S1至步骤S6递归循环采集，直至最终任务执行完成，采集结束

图1