(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
2013 年 6 月 20 日 (20.06.2013)

WIPO | PCT

(10) 国际公布号
WO 2013/087012 A1

(51) 国转 利分类号：
G06F 17/30 (2006.01)

(21) 国际申请号： PCT/CN2012/086584

(22) 国际申请日： 2012 年 12 月 13 日 (13.12.2012)

(25) 申飾 言： 中文

(26) 公布语言： 中文

(30) 优先权：
201110415356.8 2011 年 12 月 13 日 (13.12.2011) CN

(71) 申请人：北大方正集团有限公司 (PEKING UNIVERSITY FOUNDER GROUP CO., LTD.) [CN/CN]; 中国北京市海淀区成府路 298 号中关村方正大厦 5 层，Beijing 100871 (CN)。北京大学 (PEKING UNIVERSITY) [CN/CN]; 中国北京市海淀区颐和园路 5 号，Beijing 100871 (CN)。北京北大方正电子有限公司 (BEIJING FOUNDER ELECTRONICS CO., LTD.) [CN/CN]; 中国北京市海淀区上地五街九号方正大厦，Beijing 100085 (CN)。

(72) 发明人：吴新丽 (WU, Xinli); 中国北京市海淀区成府路 298 号中关村方正大厦 5 层，Beijing 100871 (CN)。杨建武 (YANG, Jianwu); 中国北京市海淀区成府路 298 号中关村方正大厦 5 层，Beijing 100871 (CN)。

(74) 代理人：北京同达信恒知识产权代理有限公司 (TDIP & PARTNERS); 中国北京市西城区裕民路 18 号北环中心 A 座 2002, Beijing 100029 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, SZ, TZ, UG, ZM, ZW), 喊亚 (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,

[见续页]

(54) Title: METHOD AND SYSTEM FOR COLLECTING NETWORK DATA

(54) 发明名称：一种网络数据的采集方法和系统

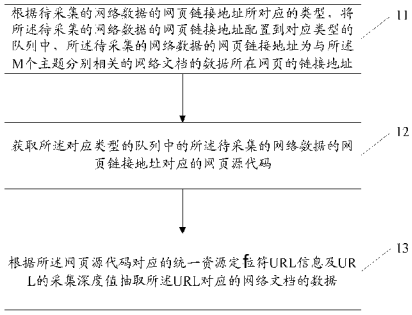

图 1 /Fig. 1

11 ACCORDING TO THE TYPE CORRESPONDING TO A WEBPAGE LINK ADDRESS OF NETWORK DATA TO BE COLLECTED, CONFIGURING THE WEBPAGE LINK ADDRESS OF NETWORK DATA TO BE COLLECTED INTO A QUEUE OF THE CORRESPONDING TYPE, THE WEBPAGE LINK ADDRESS OF NETWORK DATA TO BE COLLECTED BEING THE LINK ADDRESS OF THE WEBPAGE WHERE THE DATA OF THE NETWORK DOCUMENT WHICH IS RELEVANT TO M SUBJECTS RESPECTIVELY IS LOCATED
12 ACQUIRING THE WEBPAGE SOURCE CODES CORRESPONDING TO THE WEBPAGE LINK ADDRESS OF NETWORK DATA TO BE COLLECTED IN THE QUEUE OF THE CORRESPONDING TYPE
13 ACCORDING TO INFORMATION ABOUT A UNIFORM RESOURCE LOCATOR (URL) CORRESPONDING TO THE WEBPAGE SOURCE CODES AND A COLLECTION DEPTH VALUE OF THE URL, EXTRACTING DATA OF THE NETWORK DOCUMENT CORRESPONDING TO THE URL

(57) Abstract: Disclosed are a method and system for collecting network data. The method is used for collecting data of a network document which is released on a website and relevant to M subjects respectively, where M is a positive integer. The method includes: according to the type corresponding to a webpage link address of network data to be collected, configuring the webpage link address of network data to be collected into a queue of the corresponding type, the webpage link address of network data to be collected being the link address of the webpage where the data of the network document which is relevant to M subjects respectively is located; acquiring the webpage source codes corresponding to the webpage link address of network data to be collected in the queue of the corresponding type; and according to information about a URL corresponding to the webpage source codes and a collection depth value of the URL, extracting data of the network document corresponding to the URL.

(57) 摘要：

[见续页]

WO 2013/087012 A1 2013 08

RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, C_M, G_A, G_N, GQ, G_W, M_L, M_R, NE, S_N, T_D, TG)。

本国际公布，

– 包括国际检索报告(条约第 21 条(3))。

本发明公开了一种网络数据的采集方法和系统。该方法用于采集发布于网站上的与 M 个主题分别相关的网络文档的数据，其中M为正整数，所述方法包括：根据待采集的网络数据的网页链接地址所对应的类型，将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，所述待采集的网络数据的网页链接地址为与所述 M 个主题分别相关的网络文档的数据所在网页的链接地址；获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码；根据所述网页源代码对应的 URL 信息及所述 URL 的采集深度值抽取所述 URL 对应的网络文档的数据。

一种网络数据的采集方法和系统

本申请要求在2011年12月13日提交中国专利局、申请号为20111 041 5356.8、发明名称为"一种网络数据的采集方法和系统"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

本发明属于信息检索和数据集成技术领域，尤其涉及一种网络数据的采集方法和系统。

背景技术

随着互联网的出现及普及，互联网络为上亿网民提供了各类文学资料信息，与此同时，一种以这种新兴媒体为载体、以网民为接受对象，具有不同于传统文学特点的网络文学正悄然勃兴。

网络文学，指新近产生的，以互联网为展示平台和传播媒介的，借助超文本连接和多媒体演绎等手段来表现的文学作品、类文学文本及含有一部分文学成分的网络艺术品。其中，以网络原创作品为主。网络文学可以分为三类：一类是将已发表的文学作品经过电子扫描技术或人工录入等方式形成数字资源；一类是直接在互联网络上"发表"的文学作品；还有一类是通过计算机创作或通过有关计算机软件生成的文学作品进入互联网络，以及具有互联网络开放性特点、几位作家几十位作家甚至数百位网民共同创作的"接力小说"等。其中第二类的形式居多。

伴随网络文学的发展，由此产生的版权问题、文学创作内容问题等各种问题也朴面而来。如何便捷集中的浏览网络文学的最新内容，如何实现对网络文学的检索或监管。由于没有网络文学相关数据的支撑，这些问题目前还得不到有效的解决。

发明内容

本发明提供一种网络数据采集方法和系统，能够实时采集最新的网络数据。

本发明方法一方面提供了一种网络数据采集的方法，用于采集发布于网站上的与M个主题分别相关的网络文档的数据，其中M为正整数，所述方法包括：根据待采集的网络数据的网页链接地址所对应的类型，将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，所述待采集的网络数据的网页链接地址为与所述M个主题分别相关的网络文档的数据所在网页的链接地址；获取所述对应类型的队列中的所述待采集的网络数据的网

页链接地址对应的网页源代码；根据所述网页源代码对应的统一资源定位符 URL 信息及 URL 的采集深度值抽取所述 URL 对应的网络文档的数据。

优选地，根据所述网站发布与所述 M 个主题分别相关的网络文档的更新频率设置刷新时间间隔；以及基于所述刷新时间间隔刷新所述待采集的网络数据的网页链接地址。

优选地，所述 M 个主题中每个主题为一部网络文学，所述方法还包括：根据所述网络文学的结构配置所述 URL 的采集深度值，具体为：

$$N_{Deep}=\begin{cases}\text{第一阈值，表示作品结构为“名称→卷→章节→内容”}\\ \text{第二阈值，表示作品结构为“名称→章节→内容”}\\ \text{第三阈值，表示作品结构为“章节→内容”}\end{cases}$$

优选地，所述待采集的网络数据的网页链接地址对应的类型包括主题名称页、列表页和内容页，配置所述主题名称页用于提取主题名称；配置所述列表页用于提取主题章节目录或主题章节；配置所述内容页用于提取主题正文内容。

优选地，所述将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，具体包括：将类型为所述主题名称页的链接地址加入到主题名称页队列中；将类型为所述列表页的链接地址加入到列表页队列中；将类型为所述内容页的链接地址加入到内容页队列中。

优选地，所述获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码具体为：在所述主题名称页队列中获取所述主题名称页的链接地址对应的网页源代码。

优选地，所述根据所述网页源代码对应的 URL 信息及所述 URL 的采集深度值抽取所述 URL 对应的网络文档的数据，具体为：若采集深度值为第一阈值，则抽取主题的名称及所述名称对应的 URL，并将所述名称对应的 URL 的采集深度值标记为第二阈值后加入到所述列表页队列中；若采集深度值为第二阈值，则抽取主题的名称及所述名称对应的 URL，并将所述名称对应的 URL 的采集深度值标记为第三阈值后加入到所述列表页队列中。

优选地，所述获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码具体为：在所述列表页队列中获取所述列表页的链接地址对应的网页源代码。

优选地，所述根据所述网页源代码对应的 URL 信息及所述 URL 的采集深度值抽取所述 URL 对应的网络文档的数据具体为：若采集深度值为第二阈值，则抽取主题的章节目录及所述章节目录对应的 URL，并将所述章节目录对应的 URL 的采集深度值标记为第三

阈值后加入到所述列表页队列中；若采集深度值为第三阈值，则判断所述网页源代码对应的URL是否存在上级URL: 若是，则抽取主题的章节标题及所述章节标题对应章节的URL, 并将所述章节的URL加入到所述内容页队列中；若否，则抽取主题的名称、主题的章节标题及所述章节标题对应章节的URL, 并将所述章节的URL加入到所述内容页队列中。

优选地，所述获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码具体为：在所述内容页队列中获取所述内容页的链接地址对应的网页源代码。

优选地，所述根据所述网页源代码对应的URL信息及所述URL的采集深度值抽取所述URL对应的网络文档的数据具体为：从所述网页源代码中抽取主题的章节标题、章节正文内容，并从所述网页源代码对应的URL中抽取所述章节标题对应章节的章节ID。

优选地，判断所述章节正文内容是否存在分页：若是，则提取下一页的链接地址，并同时标记当前页的页码以及下一页的页码并将下一页的链接地址加入到所述内容页队列中等待采集。

优选地，以所述章节正文内容的第一页链接为唯一键值，存放所述分页的内容，当采集到最后一页时给予结束标识。

优选地，将抽取出的所有分页的正文内容合并到一起，结合所述章节标题进行输出。

本发明另一方面提供一种网络数据采集的系统，用于采集发布于网站上的与M个主题分别相关的网络文档的数据，其中M为正整数，所述系统包括配置模块，用于根据待采集的网络数据的网页链接地址所对应的类型，将待采集的网络数据的网页链接地址配置到对应类型的队列中，所述待采集的网络数据的网页链接地址为与所述M个主题分别相关的网络文档的数据所在网页的链接地址；网页获取模块，用于获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码；数据抽取模块，用于根据所述网页源代码对应的统一资源定位符URL信息及URL的采集深度值抽取所述URL对应的网络文档的数据。

优选地，所述系统还包括刷新模块，用于根据所述网站发布与所述M个主题分别相关的网络文档的更新频率，设置刷新时间间隔并基于所述刷新时间间隔刷新所述待采集的网络数据的网页链接地址。

优选地，所述待采集的网络数据的网页链接地址对应的类型包括主题名称页、列表页和内容页，所述配置模块包括网页配置模块，用于配置所述主题名称页用于提取主题名称、配置所述列表页用于提取主题章节目录或主题章节及配置所述内容页用于提取主题内容。

优选地，所述配置模块还包括队列配置模块，用于将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，所述队列分配模块包括：第一分配单元，用于将类型为所述主题名称页的链接地址分配到主题名称页队列中；第二分配单元，用于将类型为所述列表页的链接地址分配到列表页队列中；第三分配单元，用于将类型为所述内容页的链接地址分配到内容页队列中。

本发明有益效果如下：

本发明一实施例采用一网络数据采集系统采集网络数据，系统获取网络数据的链接地址然后配置链接地址的类型，并根据链接地址的类型将链接地址放入对应的队列中。从队列中获取链接地址对应的源代码，根据源代码中对应的URL信息及URL的采集深度值提取网络数据的信息，从而达到实时采集网络数据的技术效果。

进一步，还采用了内容合并模块，可以对属于同一主题的网络文档进行合并，所以可以在实时采集网络数据的基础上达到便捷集中浏览的效果。

附图说明

图1为本发明一实施例中的采集方法的流程图；

图2为本发明图1中采集方法的详细流程图；

图3为本发明第一实施例的采集系统架构图；

图4为本发明一实施例中的配置模块的架构图；

图5为本发明一实施例中的网页获取模块的架构图；

图6为本发明一实施例中的数据抽取模块的架构图；

图7为本发明第二实施例的采集系统架构图；

图8为本发明第三实施例的采集系统架构图；

图9为本发明第四实施例的采集系统架构图。

具体实施方式

为让本领域所属技术人员更清楚，更完整理解本发明，下面结合附图作详细介绍：

本发明一实施例提供了一种网络数据采集的方法，用于采集发布于一网站上的与M个主题分别相关的网络文档的数据，其中M为正整数，请参考图1，图1为本实施例中的采集方法的流程图。如图1所示，采集数据的方法包括：

步骤11：根据待采集的网络数据的网页链接地址所对应的类型，将待采集的网络数据的网页链接地址配置到对应类型的队列中，所述待采集的网络数据的网页链接地址为与所

述 M 个主题分别相关的网络文档的数据所在网页的链接地址；

步骤 12: 获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码；

步骤 13: 根据所述网页源代码对应的统一资源定位符 (Uniform Resource Locator, URL) 信息及所述 URL 的采集深度值抽取所述 URL 对应的网络文档的数据。

在步骤 11 中，网站上所发布的 M 个主题可以为 M 部网络文学作品，为方便理解本发明，以下实施例以网络文学为例，但并不限于网络文学。网络文学具有不同于例如网络新闻等主题的发布结构。一般的网络新闻都是单篇的，而网络文学作品发表在网站上一般有 2 种形式呈现。一种是类似于小说阅读网站的 "文学名称->章节目录页->具体的某一章节的网络文学内容页"，有的网络文学还会在 "章节目录页" 前存在 "卷" 的概念；另外一种则是类似普通新闻网站的内容目录网页，不同文学作品的章节会穿插在一起呈现，但会在标题中以类似 "文学作品名称 (5)„ 的形式来标明是同一个作品中的不同章节。

对不同结构的网络文学作品的网络文档的数据进行采集，应先获取网络文档的数据所在网页的链接地址。在本实施例中，根据网络文学作品在网站上发布的结构，网络文档的数据一般包括网络文档所属的网络文学作品的名称、网络文档所属的网络文学作品中卷及/或章节的名称、网络文档的正文内容。相应地，网络文档的数据所在的网页的链接地址对应的类型包括：主题名称页，用于提取网络文档所属的网络文学作品的名称；列表页，用于提取网络文学作品的章节目录链接和章节链接，其中，章节目录包括网络文学的卷目录和章目录；内容页，用于提取主题正文内容。

在本实施例中，将 M 个网络文学的数据所在的网页的链接地址根据其类型分别放入不同队列中。具体地，将类型为主题名称页的链接地址分配到主题名称页队列中；将类型为列表页的链接地址分配到列表页队列中；将类型为内容页的链接地址分配到内容页队列中。例如 A 网站上发布有三部网络文学作品，分别为 A1、A2、A3。其中，A1 在网站 A 上的发表结构为：文学名称->卷目录->章目录->具体的某一章节的网络文学内容页；A2 在网站 A 上的发表结构为：文学名称->章目录->具体的某一章节的网络文学内容页；A3 在网站 A 上的发表结构为：章名称->具体的某一章节的网络文学内容页，A3 的章名称即为 A3 的作品名称与章数的结合，例如 A3 的第一章的章名称是：A3 (一)；A3 的第五章的章名称是 A3 (五)。在针对网站 A 进行的一次采集过程起始时，将具有 A1 作品的名称的网页的链接地址 B1 放入主题名称页队列中；将具有 A2 作品的名称的网页的链接地址 B2 放入主题名称页队列中；将具有 A3 作品的章节链接的地址 B3 放入列表页队列中等待被采集。而内容页队列在采集起始时，并不会有待采集的链接地址放入。

在实际的采集过程中，由于网络文档会定时更新但更新频率不会像网络新闻和论坛信息那样快速，故可以采用定时刷新的策略，当然也可以采用自适应刷新的策略，即根据网站自身发布不同网络文学作品的频率自动调整刷新间隔。当检测到有网络文学作品到了其刷新间隔时间，则将刷新的待采集的网络数据的网页链接地址放入到其对应类型的队列中。

在步骤 12 中，获取各个队列中的待采集的网络数据的网页链接地址对应的网页源代码具体为根据系统设定的 URL 获取策略，例如根据系统运行情况或者各队列的情况，本领域技术人员在实际操作时可根据时间需要设定 URL 的获取策略，从各个队列中获取一个待采集的链接地址，然后系统通过 Http 请求的方式获取网页源代码。在本实施例中，例如针对网站 A 上的三部网络文学作品的采集起始时，从主题名称页队列中提取的待采集的网络数据的网页链接地址 B1、B2，根据系统设定的 URL 获取策略分别获取 B1 对应的网页源代码和 B2 对应的网页源代码；从列表页队列中提取待采集的网络数据的网页链接地址 B3 并根据系统设定的 URL 获取策略获取其网页源代码。

在步骤 13 中，网页源代码对应的 URL 信息包括网络文学作品名称、章节目录及章节链接、正文内容的链接。URL 的采集深度值根据网络文学作品的结构配置，具体为：

$$N_{Deep}=\begin{cases}\text{第一阈值，表示作品结构为 "名称}\rightarrow\text{卷}\rightarrow\text{章节}\rightarrow\text{内容"}\\ \text{第二阈值，表示作品结构为 "名称}\rightarrow\text{章节}\rightarrow\text{内容"}\\ \text{第三阈值，表示作品结构为 "章节}\rightarrow\text{内容"}\end{cases}$$

在本实施例中，第一阈值为 3，第二阈值为 2，第三阈值为 1，当然本领域技术人员也可以采用其他数值或标记来标示不同的阈值，为方便说明本发明，以下以第一阈值为 3、第二阈值为 2、第三阈值为 1 进行举例说明，按照网络文学作品的结构配置的采集深度值可以结合网站 A 上发布的 A1、A2、A3 进行理解。当从主题名称页队列中获取链接地址 B1 后，根据 B1 对应的源代码获取对应的 URL（即 URL-A1），而 A1 的结构为 "文学名称->卷目录->章目录->具体的某一章节的网络文学内容页"，则 URL-A1 的采集深度值应为 3。同理，A2 的结构为 "文学名称->章目录->具体的某一章节的网络文学内容页„，则根据 B2 得到的源代码所对应的 URL（即 URL-A2）的采集深度值为 2；A3 的结构为 "章名称->具体的某一章节的网络文学内容页"，则根据 B3 得到的源代码所对应的 URL（即 URL-A3）的采集深度值为 3。

步骤 13 具体包括：(请参考图 2)

步骤 131: 根据从主题名称页队列中获取的主题名称页的链接地址对应的网页源代码所对应的 URL 信息及 URL 采集深度值，抽取 URL 对应的网络文档的数据。

步骤 132: 根据从列表页队列中获取的列表页的链接地址对应的网页源代码所对应的URL 信息及URL 采集深度值，抽取URL 对应的网络文档的数据。

步骤 133：根据从内容页队列中获取的内容页的链接地址对应的网页源代码所对应的URL, 从网页源代码中抽取主题的章节标题、章节正文内容，并从网页源代码对应的URL 中抽取所述章节标题对应章节的章节ID。

上述步骤 131、132、133 在实现时没有先后顺序的限制，只要当各个队列中有需要待采集的链接地址时，就可以对待采集的链接地址进行采集，获取待采集的网络数据的网页链接地址对应的网页源代码并根据网页源代码对应的 URL 信息及 URL 采集深度值抽取URL 对应的网络文档的数据，下面将详细说明各步骤中对网络文档数据进行抽取的过程。

在步骤 131 中，抽取 URL 对应的网络文档的数据具体为：

若 URL 的采集深度值为 3，则抽取主题的名称及该名称对应的 URL, 并将该名称对应的URL 的采集深度值标记为第二阈值后加入到列表页队列中；

若 URL 的采集深度值为 2，则抽取主题的名称及该名称对应的 URL, 并将该名称对应的URL 的采集深度标记为 1后加入到列表页队列中。

在本实施例中，从主题名称页队列中提取的链接地址为 A1 的链接地址 B1 及 A2 的链接地址 B2。因 B1 对应源代码所对应的 URL-A1 的采集深度值为 3，则应抽取 A1 的主题名称，用 "名称 A1" 表示。还应抽取 "名称 A1" 对应的URL, 用 "URL-A11" 表示，并将 "URL-A11" 的采集深度值标记为 2 后加入到列表页队列中，以便抽取 URL-A11 中属于作品 A1 的其他信息。而对于链接地址 B2, 因URL-A2 的采集深度值为 2，故应抽取 A2 的主题名称，用 "名称 A2" 表示。还应抽取 "名称 A2" 对应的URL, 用 "URL-A21" 表示，并将 "URL-A21" 的采集深度值标记为 1后加入到列表页队列中，以便抽取 URL-A21 中属于作品 A2 的其他信息。

在步骤 132 中，抽取 URL 对应的网络文档数据具体为：

若 URL 的采集深度值为 2，则抽取主题的章节目录及章节目录对应的 URL, 并将章节目录对应的URL 的采集深度值标记为 1后加入到列表页队列中；

若 URL 的采集深度值为 1，则判断网页源代码对应的URL 是否存在上级URL:

若是，则抽取主题的章节标题及章节标题对应章节的 URL, 并将章节的 URL 加入到内容页队列中；

若否，则抽取主题的名称、主题的章节标题及章节标题对应章节的URL, 并将章节的URL 加入到内容页队列中。

在本实施例中，列表页队列中在经过步骤 131 后已存放了待采集的 URL-A11 和

URL-A2L 另外，在针对网站 A1 的网络文学作品采集的起始时，已经将作品 A3 对应的链接地址 B3 放入列表页队列中。

对于 URL-A11, 其采集深度值为 2，则抽取 A1 的章节目录及章节目录对应的 URL, 用 "URL-A12" 表示。将 URL-A12 的采集深度值标记为 1 后加入到列表页队列中。

对于 URL-A21，其采集深度值为 1 且其存有上级 URL（及 URL-A21）故抽取 A2 的章节标题及章节标题对应章节的 URL, 用 "URL-A22" 表示，并将 URL-A22 加入到内容页队列中。

对于列表页队列中的 B3，因为 B3 对应源代码所对应的 URL-A3 的采集深度值为 1 且不具有上级 URL, 故抽取 A3 的名称，用 "名称 A3" 表示、章节标题，还应抽取章节标题对应章节的 URL, 用 "URL-A31" 表示并将 URL-A31 加入到内容页队列中。

在步骤 133 中，若章节正文存在分页，则需要提取下一页的链接地址，并同时标记当前页的页码以及下一页的页码并将下一页的链接地址加入到内容页队列中等待采集。

进而，以章节正文内容的第一页链接为唯一键值，存放分页的内容，当采集到最后一页时给予结束标识。

进一步地，还可以将抽取出的所有分页的正文内容合并到一起，结合章节标题进行输出。

再进一步地，将网站、主题的名称、主题的章节标题、章节 ID、章节正文内容上载到数据库中。其中，也可以将章节正文内容以附件的形式存储到文件服务器并将存放文件的路径记录到数据库中。

在本实施例中，对网络数据的采集和合并的方法可以使得网络文学以一本书的形式展现，进一步地，采用自动刷新采集数据可以实现数据的实时采集，所以本实施例可以获得实时、便捷、集中浏览网络文学作品的有益效果。

本发明第一实施例提供了一种网络数据采集的系统，用于采集发布于一网站上的与 M 个主题分别相关的网络文档的数据，其中 M 为正整数，请参考图 3，图 3 为本实施例中的采集系统的架构图。如图 3 所示，采集数据的系统包括配置模块 31、网页获取模块 32、数据抽取模块 33。配置模块 31 用于根据待采集的网络数据的网页链接地址所对应的类型，将待采集的网络数据的网页链接地址配置到对应类型的队列中，待采集的网络数据的网页链接地址为与 M 个主题分别相关的网络文档的数据所在网页的链接地址。

网页获取模块 32 用于获取对应类型的队列中的待采集的网络数据的网页链接地址对应的网页源代码。数据抽取模块 33 用于根据网页源代码对应的 URL 信息及 URL 的采集深度值抽取 URL 对应的网络文档的数据。

本实施例中，待采集的网络数据的网页链接地址对应的类型包括主题名称页、列表页和内容页。请参考图4，配置模块31包括网页配置模块311，用于配置主题名称页用于提取主题名称、配置列表页用于提取主题章节目录或主题章节及配置内容页用于提取主题内容。

请继续参考图4，配置模块31还包括队列配置模块312，用于将所述待采集的网络数据的网页链接地址配置到对应类型的队列中。队列分配模块312还包括包括：第一分配单元3121，用于将类型为主题名称页的链接地址分配到主题名称页队列中；第二分配单元3122，用于将类型为列表页的链接地址分配到列表页队列中；第三分配单元3123，用于将类型为内容页的链接地址分配到内容页队列中。

本实施例中，网页获取模块32包括：第一获取单元321，用于在主题名称页队列中获取主题名称页的链接地址对应的网页源代码。第二获取单元322，用于在列表页队列中获取列表页的链接地址对应的网页源代码。第三获取单元323，用于在内容页队列中获取内容页的链接地址对应的网页源代码。请参考图5。

本实施例中，数据抽取模块33还包括：第一抽取单元331，用于当网页源代码对应URL的采集深度值为第一阈值时，抽取主题的名称及名称对应的URL，并将名称对应URL的采集深度值标记为第二阈值后发送到第二分配单元3122中。第二抽取单元332，用于当网页源代码对应URL的采集深度值为第二阈值，抽取主题的名称及名称对应的URL，并将名称对应URL的采集深度值标记为第三阈值后发送到第二分配单元3122中。第三抽取单元333，用于当网页源代码对应URL的采集深度值为第二阈值，则抽取主题的章节目录及章节目录的URL，并将章节目录的URL的采集深度值标记为第三阈值后发送到第二分配单元3122中。第四抽取单元334，用于判断网页源代码对应的URL是否存在上级URL，并当判断结果为是时，抽取主题的章节标题及章节标题对应章节的URL，并将章节的URL发送到第三分配单元3123中，当判断结果为否时，抽取主题的名称、章节标题及章节标题对应章节的URL，并将章节的URL发送到第三分配单元3123中。第五抽取单元335，用于从网页源代码中抽取主题的章节标题、章节正文内容，并从网页源代码对应的URL中抽取章节标题对应章节的章节ID。分页判断单元336，用于判断章节正文内容是否存在分页；当章节正文内容存在分页时，第五抽取单元335还用于提取下一页的链接地址并同时标记当前页的页码以及下一页的页码并将下一页的链接地址发送到第三分配单元3123中。分页存放单元337，用于以章节正文内容的第一页链接为唯一键值，存放分页的内容，并当采集到最后一页时给予结束标识。请参考图6。

在第二实施例中，与第一实施例不同的是系统还包括刷新模块34，用于根据所述网站

发布与所述 M 个主题分别相关的网络文档的更新频率 ,设置刷新时间间隔并基于所述刷新时间间隔刷新所述待采集的网络数据的网页链接地址。本实施例请参考图 7。

在第三实施例中 ,与第一、第二实施例不同的是系统还包括内容合并模块 35, 用于将抽取出的所有分页的正文内容合并到一起 ,并结合章节标题进行输出。本实施例请参考图 8。

在实施例中也可以结合第二实施例中的刷新模块进行采集工作 ,为了说明书的简洁 ,本处不再对结合使用的系统进行详细的介绍。

在第四实施例中 ,与第一、第二、第三实施例都不同的是系统还包括第一数据存储模块 36, 用于将网站、主题的名称、主题的章节标题、章节 ID、章节正文内容上载到数据库中。第二数据存储模块 37, 用于当章节正文内容可能占用较多数据库空间时 ,选择该数据库将网站、主题的名称、主题的章节标题、章节 ID、章节正文内容的存放路径上载到数据库中 ,其中 ,章节正文内容存放路径是指将章节正文内容以附件的形式存储到文件服务器的路径。本实施例请参考图 9。

在本实施例中也可以结合第二实施例中的刷新模块进行采集工作 ,为了说明书的简洁 ,本处不再对结合使用的系统进行详细的介绍。

上述第一、第二、第三及第四实施例中的系统可以根据本发明提供的一种网络数据采集方法的实施例中对方法及其各种变化形式的描述进行实施。本处为了说明书的简洁 ,所以不再详述。

本发明一实施例采用一 网络数据采集系统采集网络数据 ,系统获取网络数据的链接地址然后配置链接地址的类型 ,并根据链接地址的类型将链接地址放入对应的队列中。从队列中获取链接地址对应的源代码 ,根据源代码中对应的 URL 信息及 URL 的采集深度值提取网络数据的信息 ,从而达到实时采集网络数据的技术效果。进一步 ,还采用了内容合并模块 ,可以对属于同一主题的网络文档进行合并 ,所以可以在实时采集网络数据的基础上达到便捷集中浏览的效果。

本领域内的技术人员应明白 ,本发明的实施例可提供为方法、系统、或计算机程序产品。因此 ,本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且 ,本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介盾 (包括但不限于磁盘存储器、CD-ROM 、光学存储器等)上实施的计算机程序产品的形式。

本发明是参照根据本发明实施例的方法、设备 (系统) 和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机程序指令实现流程图和 / 或方框图中的每一流

程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

显然，本领域的技术人员可以对本发明实施例进行各种改动和变型而不脱离本发明实施例的精神和范围。这样，倘若本发明实施例的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

1、一种网络数据采集的方法，用于采集发布于网站上的与 M 个主题分别相关的网络文档的数据，其中M 为正整数，其特征在于，所述方法包括：

根据待采集的网络数据的网页链接地址所对应的类型，将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，所述待采集的网络数据的网页链接地址为与所述 M 个主题分别相关的网络文档的数据所在网页的链接地址；

获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码；

根据所述网页源代码对应的统一资源定位符 URL 信息及 URL 的采集深度值抽取所述 URL 对应的网络文档的数据。

2、如权利要求 1所述的方法，其特征在于，所述方法还包括：

根据所述网站发布与所述 M 个主题分别相关的网络文档的更新频率，设置刷新时间间隔；以及

基于所述刷新时间间隔刷新所述待采集的网络数据的网页链接地址。

3、如权利要求 1所述的方法，其特征在于，所述 M 个主题中每个主题为一部网络文学，所述方法还包括：根据所述网络文学的结构配置所述 URL 的采集深度值，具体为：

$$N_{Deep} = \begin{cases} \text{第一阈值，表示作品结构为 "名称} \rightarrow \text{卷} \rightarrow \text{章节} \rightarrow \text{内容"} \\ \text{第二阈值，表示作品结构为 "名称} \rightarrow \text{章节} \rightarrow \text{内容"} \\ \text{第三阈值，表示作品结构为 "章节} \rightarrow \text{内容"} \end{cases}。$$

4、如权利要求 1所述的方法，其特征在于，所述待采集的网络数据的网页链接地址对应的类型包括主题名称页、列表页和内容页，配置所述主题名称页用于提取主题名称；配置所述列表页用于提取主题章节目录或主题章节；配置所述内容页用于提取主题正文内容。

5、如权利要求 4所述的方法，其特征在于，所述将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，具体包括：

将类型为所述主题名称页的链接地址分配到主题名称页队列中；

将类型为所述列表页的链接地址分配到列表页队列中；

将类型为所述内容页的链接地址分配到内容页队列中。

6、如权利要求 5所述的方法，其特征在于，所述获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码具体为：

在所述主题名称页队列中获取所述主题名称页的链接地址对应的网页源代码。

7、如权利要求 6 所述的方法，其特征在于，所述根据所述网页源代码对应的 URL 信息及所述 URL 的采集深度值抽取所述 URL 对应的网络文档的数据，具体为：

若采集深度值为第一阈值，则抽取主题的名称及所述名称对应的 URL，并将所述名称对应的 URL 的采集深度值标记为第二阈值后加入到所述列表页队列中；

若采集深度值为第二阈值，则抽取主题的名称及所述名称对应的 URL，并将所述名称对应的 URL 的采集深度值标记为第三阈值后加入到所述列表页队列中。

8、如权利要求 5 所述的方法，其特征在于，所述获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码具体为：

在所述列表页队列中获取所述列表页的链接地址对应的网页源代码。

9、如权利要求 8 所述的方法，其特征在于，所述根据所述网页源代码对应的 URL 信息及所述 URL 的采集深度值抽取所述 URL 对应的网络文档的数据具体为：

若采集深度值为第二阈值，则抽取主题的章节目录及所述章节目录对应的 URL，并将所述章节目录对应的 URL 的采集深度值标记为第三阈值后加入到所述列表页队列中；

若采集深度值为第三阈值，则判断所述网页源代码对应的 URL 是否存在上级 URL:

若是，则抽取主题的章节标题及所述章节标题对应章节的 URL，并将所述章节的 URL 加入到所述内容页队列中；

若否，则抽取主题的名称、主题的章节标题及所述章节标题对应章节的 URL，并将所述章节的 URL 加入到所述内容页队列中。

10、如权利要求 5 所述的方法，其特征在于，所述获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码具体为：

在所述内容页队列中获取所述内容页的链接地址对应的网页源代码。

11、如权利要求 10 所述的方法，其特征在于，所述根据所述网页源代码对应的 URL 信息及所述 URL 的采集深度值抽取所述 URL 对应的网络文档的数据具体为：

从所述网页源代码中抽取主题的章节标题、章节正文内容，并从所述网页源代码对应的 URL 中抽取所述章节标题对应章节的章节 ID。

12、如权利要求 11 所述的方法，其特征在于，所述方法还包括：

当所述章节正文内容存在分页时，提取下一页的链接地址，并同时标记当前页的页码以及下一页的页码并将下一页的链接地址加入到所述内容页队列中等待采集。

13、如权利要求 12 所述的方法，其特征在于，所述方法还包括：

以所述章节正文内容的第一页链接为唯一键值，存放所述分页的内容，当采集到最后

页时给予结束标识。

14、如权利要求 13 所述的方法，其特征在于，所述方法还包括：

将抽取出的所有分页的正文内容合并到一起，结合所述章节标题进行输出。

15、一种网络数据采集的系统，用于采集发布于网站上的与 M 个主题分别相关的网络文档的数据，其中 M 为正整数，其特征在于，所述系统包括：

配置模块，用于根据待采集的网络数据的网页链接地址所对应的类型，将待采集的网络数据的网页链接地址配置到对应类型的队列中，所述待采集的网络数据的网页链接地址为与所述 M 个主题分别相关的网络文档的数据所在网页的链接地址；

网页获取模块，用于获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码；

数据抽取模块，用于根据所述网页源代码对应的统一资源定位符 URL 信息及 URL 的采集深度值抽取所述 URL 对应的网络文档的数据。

16、如权利要求 15 所述的系统，其特征在于，所述系统还包括：刷新模块，用于根据所述网站发布与所述 M 个主题分别相关的网络文档的更新频率，设置刷新时间间隔并基于所述刷新时间间隔刷新所述待采集的网络数据的网页链接地址。

17、如权利要求 15 所述的系统，其特征在于，所述待采集的网络数据的网页链接地址对应的类型包括主题名称页、列表页和内容页，所述配置模块包括：网页配置模块，用于配置所述主题名称页用于提取主题名称、配置所述列表页用于提取主题章节目录或主题章节及配置所述内容页用于提取主题内容。

18、如权利要求 17 所述的系统，其特征在于，所述配置模块还包括：队列配置模块，用于将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，所述队列分配模块包括：

第一分配单元，用于将类型为所述主题名称页的链接地址分配到主题名称页队列中；

第二分配单元，用于将类型为所述列表页的链接地址分配到列表页队列中；

第三分配单元，用于将类型为所述内容页的链接地址分配到内容页队列中。

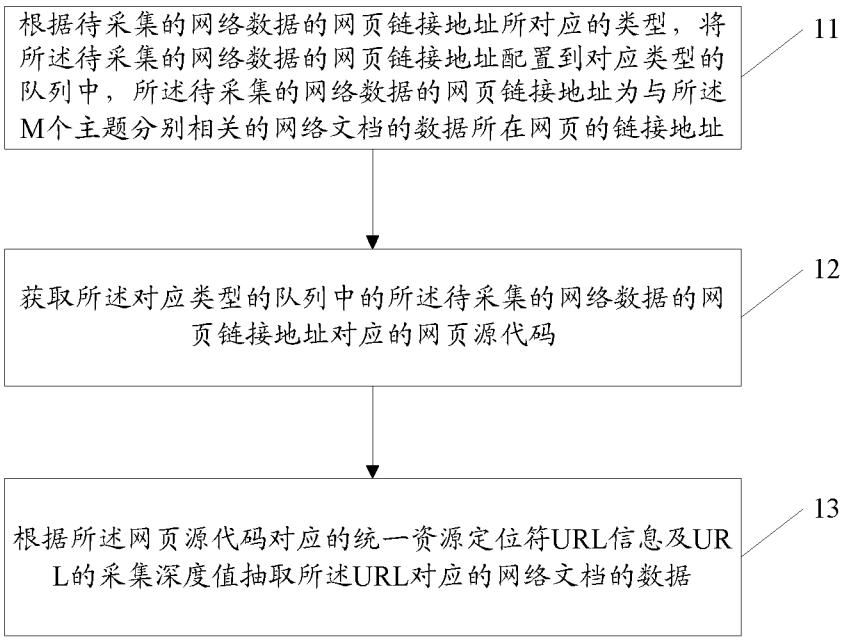
根据待采集的网络数据的网页链接地址所对应的类型，将所述待采集的网络数据的网页链接地址配置到对应类型的队列中，所述待采集的网络数据的网页链接地址为与所述M个主题分别相关的网络文档的数据所在网页的链接地址
11
获取所述对应类型的队列中的所述待采集的网络数据的网页链接地址对应的网页源代码
12
根据所述网页源代码对应的统一资源定位符URL信息及URL的采集深度值抽取所述URL对应的网络文档的数据
13

图 1

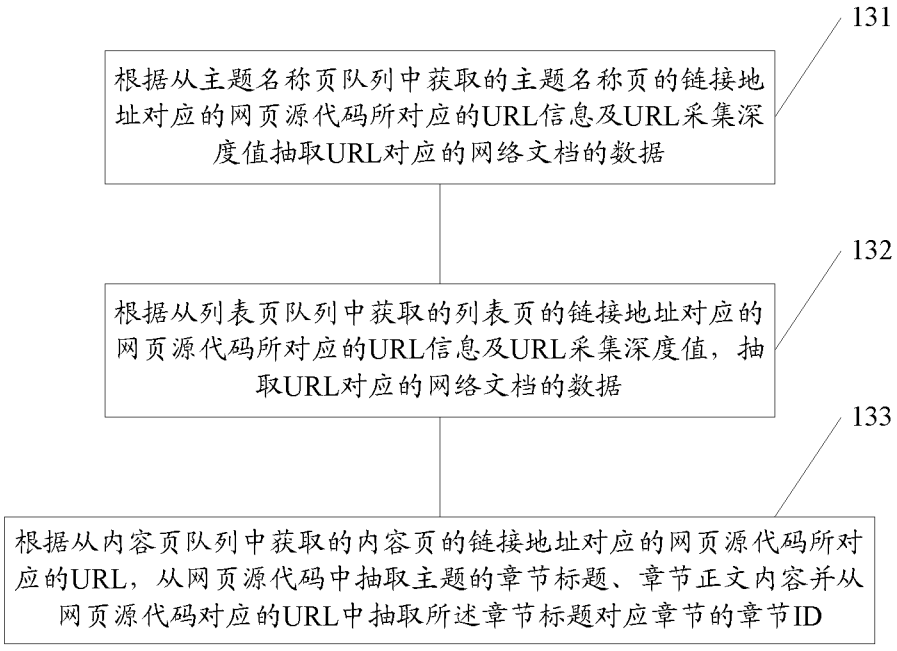
131
根据从主题名称页队列中获取的主题名称页的链接地址对应的网页源代码所对应的URL信息及URL采集深度值抽取URL对应的网络文档的数据
132
根据从列表页队列中获取的列表页的链接地址对应的网页源代码所对应的URL信息及URL采集深度值，抽取URL对应的网络文档的数据
133
根据从内容页队列中获取的内容页的链接地址对应的网页源代码所对应的URL，从网页源代码中抽取主题的章节标题、章节正文内容并从网页源代码对应的URL中抽取所述章节标题对应章节的章节ID

图 2

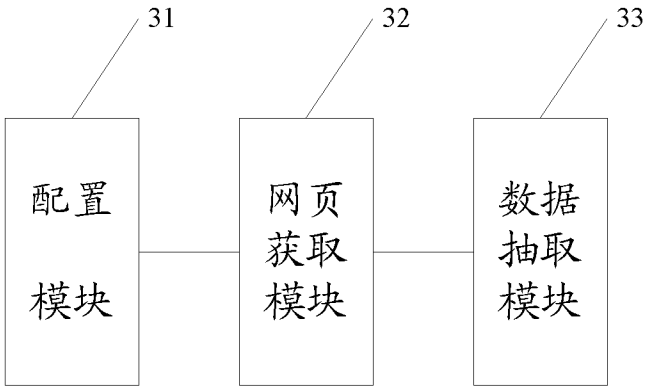
31
32
33
配置模块
网页获取模块
数据抽取模块

图 3

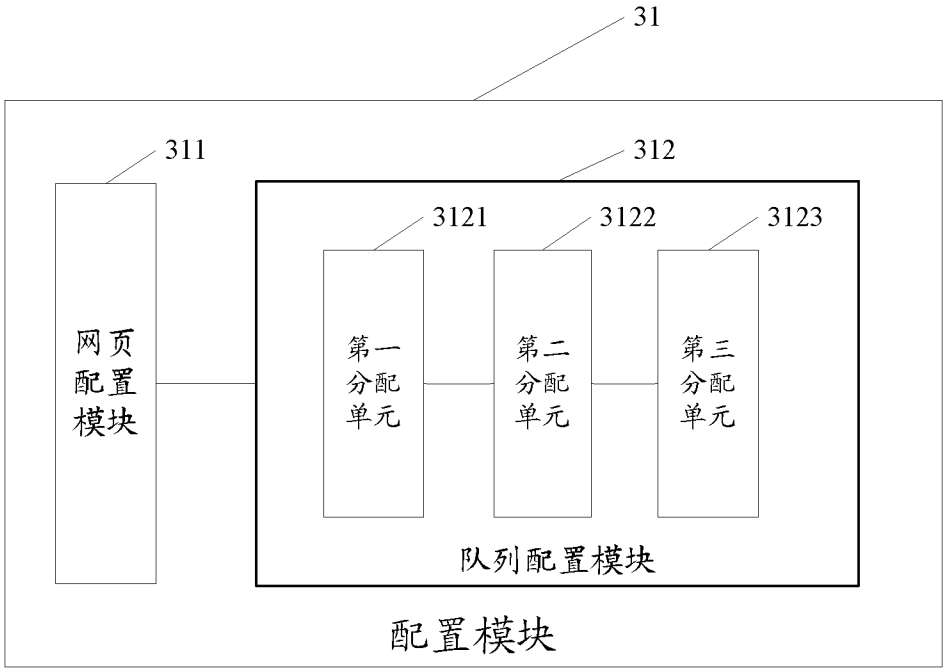
31
311
312
3121
3122
3123
网页配置模块
第一分配单元
第二分配单元
第三分配单元
队列配置模块
配置模块

图 4

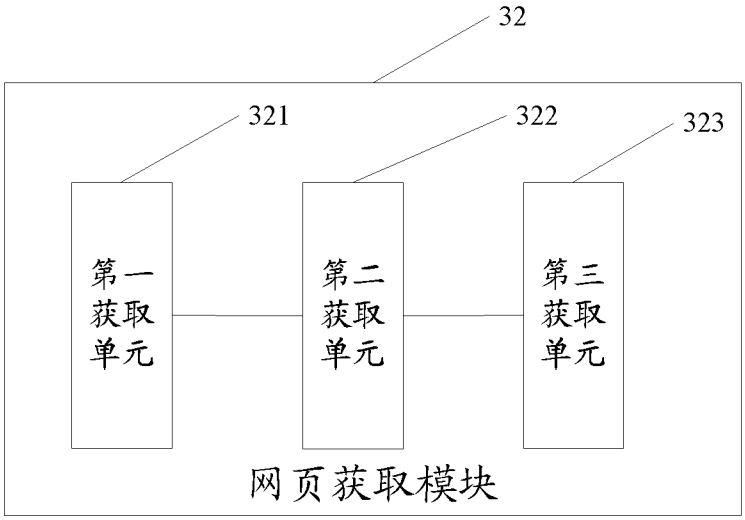
32
321
322
323
第一获取单元
第二获取单元
第三获取单元
网页获取模块

图 5

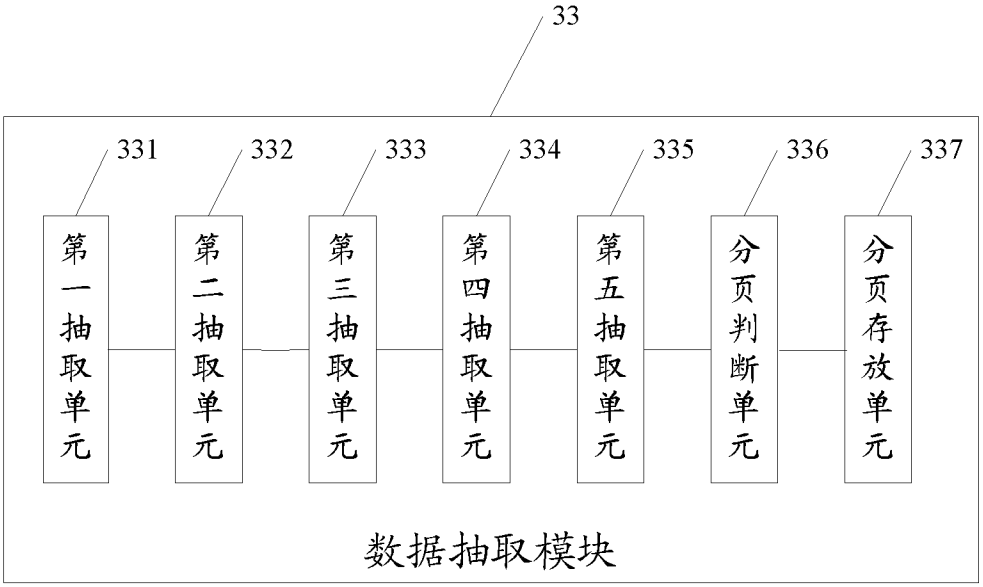
33
331
332
333
334
335
336
337
第一抽取单元
第二抽取单元
第三抽取单元
第四抽取单元
第五抽取单元
分页判断单元
分页存放单元
数据抽取模块

图 6

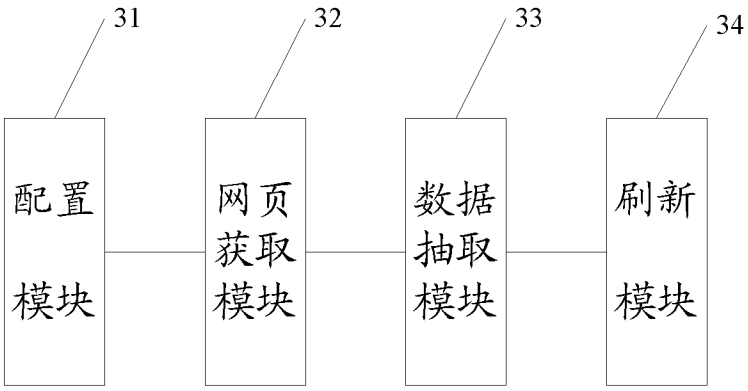
31
32
33
34
配置模块
网页获取模块
数据抽取模块
刷新模块

图 7

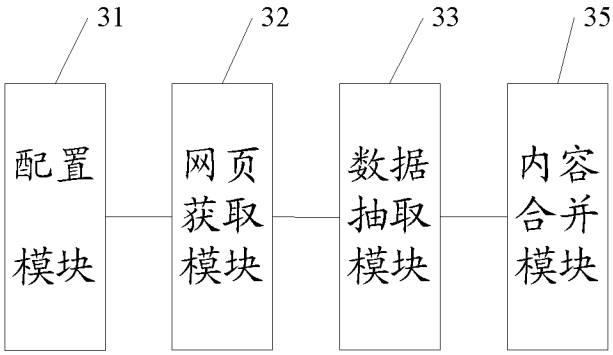
31
32
33
35
配置模块
网页获取模块
数据抽取模块
内容合并模块

图 8

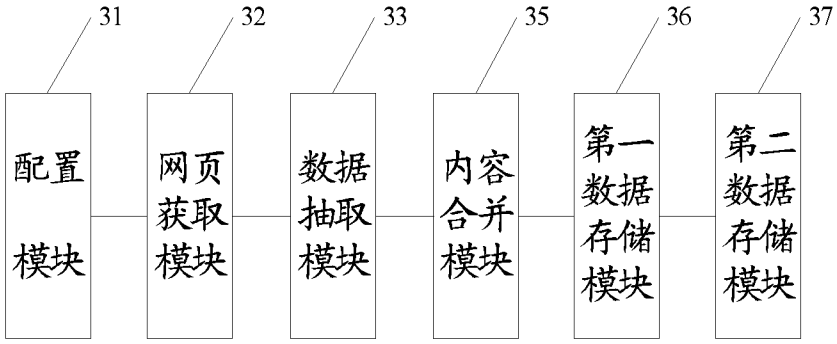
31
32
33
35
36
37
配置模块
网页获取模块
数据抽取模块
内容合并模块
第一数据存储模块
第二数据存储模块

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/086584

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (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)

IPC: G06F; G06Q

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)

CPRSABS VEN CNKI: text allocate URL web page link address document literary novel article classify category categorize source code uniform resource locator universal resource locator mining

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102118400 A (NAVINFO CO., LTD.), 06 July 2011 (06.07.2011), abstract, and description, paragraphs 3-92	1-11, 15-18
A		12-14
Y	CN 101593200 A (HUAIHAI INSTITUTE OF TECHNOLOGY), 02 December 2009 (02.12.2009), abstract, and description, page 1, line 16 to page 7, line 12	1-11, 15-18
A		12-14
Y	CN 101094135 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.), 26 December 2007 (26.12.2007), abstract, and description, page 1, line 13 to page 5, line 16	1-11, 15-18
Y	CN 101136026 A (BEIJING JUSHENG SCIENCE TECHNOLOGY CO., LTD.), 05 March 2008 (05.03.2008), abstract, and description, page 1, lines 7-27	1-11, 15-18

☐ 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
05 March 2013 (05.03.2013)	21 March 2013 (21.03.2013)
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 WANG, Xuelian Telephone No.: (86-10) 62411747

Form PCT/IS A/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/086584

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102118400 A	06.07.2011	None	
CN 101593200 A	02.12.2009	CN 101593200 B	03.10.2012
CN 101094135 A	26.12.2007	CN 100512181 C	08.07.2009
CN 101136026 A	05.03.2008	None	

国际检索报告

国际申请号
PCT/CN2012/086584

A．主题的分类

G06F 17/30 (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)

IPC: G06F; G06Q

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)，CPRSABS VEN CNKI 网页 链接 地址 文档 文本 文学 文章 类型 分类 类别 分配 源码 源代码 原始码 URL 统一资源定位符 挖掘 web page link address document literary novel article classify category categorize source code uniform resource locator universal resource locator mining

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
Y	CN 1021 18400 A (北京四维图新科技股份有限公司) 06.7 月 201 1 (06.07.201 1) 摘要，说明书第 3-92 段	1-1 1, 15-18
A		12-14
Y	CN 101593200 A (淮海工学院) 02. 12 月 2009 (02. 12.2009) 摘要，说明书第 1 页第 16 行-第 7 页第 12 行	1-1 1, 15-18
A		12-14
Y	CN 101094135 A (腾讯科技(深圳)有限公司) 26.12 月 2007 (26. 12.2007) 摘要，说明书第 1 页第 13 行-第 5 页第 16 行	1-1 1, 15-18
Y	CN 101 136026 A (北京聚生科技有限公司) 05.3 月 2008 (05.03.2008) 摘要，说明书第 1 页第 7-27 行	1-1 1, 15-18

口 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¾ ¾¾言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
05.3 月 2013 (05.03.2013)	21.3 月 2013 (21.03.2013)
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 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 王雪莲 电话号码：(86-10) 62411747

国际检索报告

关于同族专利的信息

国际申请号
PCT/CN2012/086584

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN 1021 18400 A	06.07.201 1	无	
CN 101593200 A	02. 12.2009	CN 101593200 B	03. 10.2012
CN 101094135 A	26. 12.2007	CN 100512181 C	08.07.2009
CN 101 136026 A	05.03.2008	无	