(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104077402 B
(45)授权公告日 2018.01.19

(21)申请号 201410318757.5

(22)申请日 2014.07.04

(65)同一申请的已公布的文献号
申请公布号 CN 104077402 A

(43)申请公布日 2014.10.01

(73)专利权人 用友网络科技股份有限公司
地址 100094 北京市海淀区北清路68号

(72)发明人 张欣

(74)专利代理机构 北京友联知识产权代理事务所(普通合伙) 11343
代理人 尚志峰 汪海屏

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 101334784 A,2008.12.31,
CN 103678665 A,2014.03.26,
CN 102799686 A,2012.11.28,
US 2004148278 A1,2004.07.29,
CN 101334784 A,2008.12.31,

审查员 袁冠群

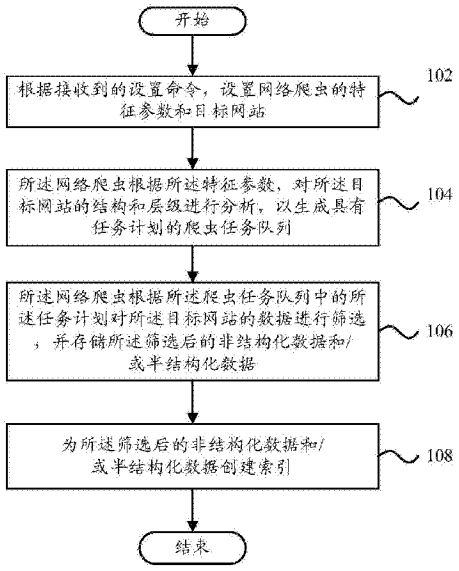
权利要求书2页 说明书7页 附图2页

(54)发明名称

数据处理方法和数据处理系统

(57)摘要

本发明提供了一种数据处理方法和一种数据处理系统，包括：设置网络爬虫的特征参数和目标网站；所述网络爬虫根据所述特征参数，对所述目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列；所述网络爬虫根据所述爬虫任务队列中的所述任务计划对所述目标网站的数据进行筛选，并存储所述筛选后的非结构化数据和/或半结构化数据；为所述筛选后的非结构化数据和/或半结构化数据创建索引。通过本发明的技术方案，能对网络爬虫的参数进行设置，并根据目标网站的结构和层级完善任务计划，同时还可以对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析，以保证系统的可用性和高效性。



CN 104077402 B

1.一种数据处理方法，其特征在于，包括：

根据接收到的设置命令，设置网络爬虫的特征参数和目标网站；

所述网络爬虫根据所述特征参数，对所述目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列；

所述网络爬虫根据所述爬虫任务队列中的所述任务计划对所述目标网站的数据进行筛选，并存储所述筛选后的非结构化数据和/或半结构化数据；

为所述筛选后的非结构化数据和/或半结构化数据创建索引；

存储所述筛选后的非结构化数据和/或半结构化数据，具体包括：

将所述非结构化数据以文件形式存储在分布式文件系统中；以及

通过分布式计算框架将所述半结构化数据进行分析后转换为结构化数据，并将所述结构化数据存储在分布式列式存储数据库中。

2.根据权利要求1所述的数据处理方法，其特征在于，还包括：

根据所述索引查询所述存储后的数据；以及

对所述存储后的数据进行统计和/或分析，并生成统计结果和/或分析图表。

3.根据权利要求1所述的数据处理方法，其特征在于，还包括：

根据用户设置的关键词，自动筛选出与所述关键词相关的数据，并为所述筛选后的与所述关键词相关的数据生成统计结果和/或分析图表。

4.根据权利要求2或3所述的数据处理方法，其特征在于，还包括：

将所述统计结果和/或所述分析图表在指定时间发送给指定用户。

5.一种数据处理系统，其特征在于，包括：

设置单元，根据接收到的设置命令，设置网络爬虫的特征参数和目标网站；

任务建立单元，所述网络爬虫根据所述特征参数，对所述目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列；

执行单元，所述网络爬虫根据所述爬虫任务队列中的所述任务计划对所述目标网站的数据进行筛选，并存储所述筛选后的非结构化数据和/或半结构化数据；

索引创建单元，用于为所述筛选后的非结构化数据和/或半结构化数据创建索引；

所述执行单元包括：

将所述非结构化数据以文件形式存储在分布式文件系统中；以及

通过分布式计算框架将所述半结构化数据进行分析后转换为结构化数据，并将所述结构化数据存储在分布式列式存储数据库中。

6.根据权利要求5所述的数据处理系统，其特征在于，还包括：

查询单元，用于根据所述索引查询所述存储后的数据；以及

分析单元，用于对所述存储后的数据进行统计和/或分析，并生成统计结果和/或分析图表。

7.根据权利要求5所述的数据处理系统，其特征在于，所述执行单元还包括：

根据用户设置的关键词，自动筛选出与所述关键词相关的数据；以及

所述分析单元还包括：

为所述筛选后的与所述关键词相关的数据生成统计结果和/或分析图表。

8.根据权利要求6或7所述的数据处理系统，其特征在于，还包括：

发送单元，用于将所述统计结果和/或所述分析图表在指定时间发送给指定用户。

数据处理方法和数据处理系统

技术领域

[0001] 本发明涉及数据处理技术领域，具体而言，涉及一种数据处理方法和一种数据处理系统。

背景技术

[0002] 目前，随着互联网在社会上的普及，每天在互联网中传播的信息量都呈爆炸式增长，统计发现，每天有约200万篇博客文章在网上被发布，每天在社交网站被发布的状态更新有约5亿3200万条，面对互联网中浩如烟海的信息，随时了解互联网的所有动态变得愈加艰难。

[0003] 在现有技术中，一般只能将使用爬虫抓取到的海量数据存储在关系型数据库中，关系型数据库的存储容量会受到单台服务器硬件配置的极大制约，很难或根本无法对系统进行横向扩展，并且，只能存储经过处理后的结构化数据，无法对互联网中大量的非结构化数据进行存储和分析，这导致了部分有价值的数据只能被舍弃，要及时汇总关注的某个领域的信息只能通过耗费大量的人力和时间来实现。另外，现有技术还遭遇了数据量积累到一定程度后出现的查询、分析、知识转移等方面的困难。

[0004] 因此，提出一种高效且灵活的数据处理方法显得十分必要。

发明内容

[0005] 本发明正是基于上述技术问题，提出了一种新的技术方案，能对网络爬虫的参数进行设置，根据目标网站的结构和层级完善任务计划，还可以同时对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析，保证系统的可用性和高效性，比现有的爬虫技术更具灵活性。

[0006] 有鉴于此，本发明提出了一种数据处理方法，包括：根据接收到的设置命令，设置网络爬虫的特征参数和目标网站；所述网络爬虫根据所述特征参数，对所述目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列；所述网络爬虫根据所述爬虫任务队列中的所述任务计划对所述目标网站的数据进行筛选，并存储所述筛选后的非结构化数据和/或半结构化数据；为所述筛选后的非结构化数据和/或半结构化数据创建索引。

[0007] 在该技术方案中，通过Web界面可以对网络爬虫进行多种参数的设置，可以分类大量收集对应每种特征参数的数据信息，通过智能检索目标网站的URL(Uniform Resource Locator，统一资源定位器)及网站结构和层级来创建完善的执行计划，并依此可以做到对收集到的海量信息进行过滤，还可同时存储采集到的结构化或非结构化数据，还可以同时对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析。这样，利用大数据技术搭建的分布式存储系统来存储和处理采集到的数据，可以横向扩展存储容量和并行数据处理能力，保证系统的可用性和高效性，比现有的爬虫技术更具灵活性，这些都是传统的数据库所无法比拟的，可以应用到舆情监控、商业信息收集、市场行情分析、电子商务推广等领域中去。

[0008] 在上述技术方案中，优选地，存储所述筛选后的非结构化数据和/或半结构化数据，具体包括：将所述非结构化数据以文件形式存储在分布式文件系统中；以及通过分布式计算框架将所述半结构化数据进行分析后转换为结构化数据，并将所述结构化数据存储在分布式列式存储数据库中。

[0009] 在该技术方案中，非结构化数据包括图片、视频等，半结构化数据包括html、xml等类型的文件，非结构化数据将会以文件形式存储在分布式文件系统(HDFS)中，利用分布式计算框架(MapReduce)对半结构化数据进行分析，并转换为结构化数据存储在分布式列式存储数据库(HBase)中。这样解决了无法对非结构化数据进行存储和分析的问题，实现了对海量半结构化和非结构化内容的采集，成功保留了有价值的数据。

[0010] 在上述技术方案中，优选地，还包括：根据所述索引查询所述存储后的数据；以及对所述存储后的数据进行统计和/或分析，并生成统计结果和/或分析图表。

[0011] 在该技术方案中，使用分布式文件系统和分布式列式存储数据库作为搜索引擎技术的底层支撑，利用搜索引擎技术为收集的所有数据建立索引，便于对已有的数据进行快速检索和查询，实现了搜索引擎的分布式索引，用户可以通过Web用户界面对所有采集到的数据进行查询，搜索引擎会快速返回查询结果，还可以对图片、视频等内容进行在线展示，在保证搜索引擎系统高效响应的同时也兼顾了系统整体的易扩容和高可用。另外，数据汇总后会形成有价值的信息，如统计结果和/或分析图表，以供用户读取。

[0012] 在上述技术方案中，优选地，还包括：根据用户设置的关键词，自动筛选出与所述关键词相关的数据，并为所述筛选后的与所述关键词相关的数据生成统计结果和/或分析图表。

[0013] 在该技术方案中，用户还可以使用关键词设置功能对自己感兴趣的内容进行定义，系统会利用分词技术自动匹配与用户设置的关键词相关的内容，用户可以根据系统给出的统计结果和/或分析图表作为参考，对商业和市场行为进行更准确的决策。

[0014] 在上述技术方案中，优选地，还包括：将所述统计结果和/或所述分析图表在指定时间发送给指定用户。

[0015] 在该技术方案中，系统可以根据用户的设置命令，自动将统计结果和/或所述分析图表通过邮件等方式定时发送给指定用户，提高了用户使用的便利性。

[0016] 根据本发明的另一方面，还提供了一种数据处理系统，包括：设置单元，根据接收到的设置命令，设置网络爬虫的特征参数和目标网站；任务建立单元，所述网络爬虫根据所述特征参数，对所述目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列；执行单元，所述网络爬虫根据所述爬虫任务队列中的所述任务计划对所述目标网站的数据进行筛选，并存储所述筛选后的非结构化数据和/或半结构化数据；索引创建单元，用于为所述筛选后的非结构化数据和/或半结构化数据创建索引。

[0017] 在该技术方案中，通过Web界面可以对网络爬虫进行多种参数的设置，可以分类大量收集对应每种特征参数的数据信息，通过智能检索目标网站的URL(Uniform Resource Locator，统一资源定位器)及网站结构和层级来创建完善的执行计划，并依此可以做到对收集到的海量信息进行过滤，还可同时存储采集到的结构化或非结构化数据，还可以同时对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析。这样，利用大数据技术搭建的分布式存储系统来存储和处理采集到的数据，可以横向扩展存储容量和并行

数据处理能力，保证系统的可用性和高效性，比现有的爬虫技术更具灵活性，这些都是传统的数据库所无法比拟的，可以应用到舆情监控、商业信息收集、市场行情分析、电子商务推广等领域中去。

[0018] 在上述技术方案中，优选地，所述执行单元包括：将所述非结构化数据以文件形式存储在分布式文件系统中；以及通过分布式计算框架将所述半结构化数据进行分析后转换为结构化数据，并将所述结构化数据存储在分布式列式存储数据库中。

[0019] 在该技术方案中，非结构化数据包括图片、视频等，半结构化数据包括html、xml等类型的文件，非结构化数据将会以文件形式存储在分布式文件系统(HDFS)中，利用分布式计算框架(MapReduce)对半结构化数据进行分析，并转换为结构化数据存储在分布式列式存储数据库(HBase)中。这样解决了无法对非结构化数据进行存储和分析的问题，实现了对海量半结构化和非结构化内容的采集，成功保留了有价值的数据。

[0020] 在上述技术方案中，优选地，还包括：查询单元，用于根据所述索引查询所述存储后的数据；以及分析单元，用于对所述存储后的数据进行统计和/或分析，并生成统计结果和/或分析图表。

[0021] 在该技术方案中，使用分布式文件系统和分布式列式存储数据库作为搜索引擎技术的底层支撑，利用搜索引擎技术为收集的所有数据建立索引，便于对已有的数据进行快速检索和查询，实现了搜索引擎的分布式索引，用户可以通过Web用户界面对所有采集到的数据进行查询，搜索引擎会快速返回查询结果，还可以对图片、视频等内容进行在线展示，在保证搜索引擎系统高效响应的同时也兼顾了系统整体的易扩容和高可用。另外，数据汇总后会形成有价值的信息，如统计结果和/或分析图表，以供用户读取。

[0022] 在上述技术方案中，优选地，所述执行单元还包括：根据用户设置的关键词，自动筛选出与所述关键词相关的数据；以及所述分析单元还包括：为所述筛选后的与所述关键词相关的数据生成统计结果和/或分析图表。

[0023] 在该技术方案中，用户还可以使用关键词设置功能对自己感兴趣的内容进行定义，系统会利用分词技术自动匹配与用户设置的关键词相关的内容，用户可以根据系统给出的统计结果和/或分析图表作为参考，对商业和市场行为进行更准确的决策。

[0024] 在上述技术方案中，优选地，还包括：发送单元，用于将所述统计结果和/或所述分析图表在指定时间发送给指定用户。

[0025] 在该技术方案中，系统可以根据用户的设置命令，自动将统计结果和/或所述分析图表通过邮件等方式定时发送给指定用户，提高了用户使用的便利性。

[0026] 通过本发明的技术方案，能对网络爬虫的参数进行设置，根据目标网站的结构和层级完善任务计划，还可以同时对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析，可以横向扩展存储容量和并行数据处理能力，保证系统的可用性和高效性，比现有的爬虫技术更具灵活性，用户还可以根据系统可定时给出的分析结果对商业和市场行为进行更准确的决策，提高了用户使用的便利性。

附图说明

[0027] 图1示出了根据本发明的实施例的数据处理方法的流程图；

[0028] 图2示出了根据本发明的实施例的数据处理系统的框图；

[0029] 图3示出了根据本发明的一个实施例的数据处理系统的结构示意图；

[0030] 图4示出了根据本发明的另一个实施例的数据处理方法的示意流程图。

具体实施方式

[0031] 为了能够更清楚地理解本发明的上述目的、特征和优点，下面结合附图和具体实施方式对本发明进行进一步的详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

[0032] 在下面的描述中阐述了很多具体细节以便于充分理解本发明，但是，本发明还可以采用其他不同于在此描述的其他方式来实施，因此，本发明的保护范围并不受下面公开的具体实施例的限制。

[0033] 图1示出了根据本发明的实施例的数据处理方法的流程图。

[0034] 如图1所示，根据本发明的实施例的数据处理方法，包括：

[0035] 步骤102，根据接收到的设置命令，设置网络爬虫的特征参数和目标网站。

[0036] 步骤104，网络爬虫根据特征参数，对目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列。

[0037] 步骤106，网络爬虫根据爬虫任务队列中的任务计划对目标网站的数据进行筛选，并存储筛选后的非结构化数据和/或半结构化数据。

[0038] 步骤108，为筛选后的非结构化数据和/或半结构化数据创建索引。

[0039] 在该技术方案中，通过Web界面可以对网络爬虫进行多种参数的设置，可以分类大量收集对应每种特征参数的数据信息，通过智能检索目标网站的URL(Uniform Resource Locator，统一资源定位器)及网站结构和层级来创建完善的执行计划，并依此可以做到对收集到的海量信息进行过滤，还可同时存储采集到的结构化或非结构化数据，还可以同时对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析。这样，利用大数据技术搭建的分布式存储系统来存储和处理采集到的数据，可以横向扩展存储容量和并行数据处理能力，保证系统的可用性和高效性，比现有的爬虫技术更具灵活性，这些都是传统的数据库所无法比拟的，可以应用到舆情监控、商业信息收集、市场行情分析、电子商务推广等领域中去。

[0040] 在上述技术方案中，优选地，步骤106中，存储筛选后的非结构化数据和/或半结构化数据，具体包括：将非结构化数据以文件形式存储在分布式文件系统中；以及通过分布式计算框架将半结构化数据进行分析后转换为结构化数据，并将结构化数据存储在分布式列式存储数据库中。

[0041] 在该技术方案中，非结构化数据包括图片、视频等，半结构化数据包括html、xml等类型的文件，非结构化数据将会以文件形式存储在分布式文件系统(HDFS)中，利用分布式计算框架(MapReduce)对半结构化数据进行分析，并转换为结构化数据存储在分布式列式存储数据库(HBase)中。这样解决了无法对非结构化数据进行存储和分析的问题，实现了对海量半结构化和非结构化内容的采集，成功保留了有价值的数据。

[0042] 在上述技术方案中，优选地，在步骤108之后还包括：根据索引查询存储后的数据；以及对存储后的数据进行统计和/或分析，并生成统计结果和/或分析图表。

[0043] 在该技术方案中，使用分布式文件系统和分布式列式存储数据库作为搜索引擎技

术的底层支撑，利用搜索引擎技术为收集的所有数据建立索引，便于对已有的数据进行快速检索和查询，实现了搜索引擎的分布式索引，用户可以通过Web用户界面对所有采集到的数据进行查询，搜索引擎会快速返回查询结果，还可以对图片、视频等内容进行在线展示，在保证搜索引擎系统高效响应的同时也兼顾了系统整体的易扩容和高可用。另外，数据汇总后会形成有价值的信息，如统计结果和/或分析图表，以供用户读取。

[0044] 在上述技术方案中，优选地，还包括：根据用户设置的关键词，自动筛选出与关键词相关的数据，并为筛选后的与关键词相关的数据生成统计结果和/或分析图表。

[0045] 在该技术方案中，用户还可以使用关键词设置功能对自己感兴趣的内容进行定义，系统会利用分词技术自动匹配与用户设置的关键词相关的内容，用户可以根据系统给出的统计结果和/或分析图表作为参考，对商业和市场行为进行更准确的决策。

[0046] 在上述技术方案中，优选地，还包括：将统计结果和/或分析图表在指定时间发送给指定用户。

[0047] 在该技术方案中，系统可以根据用户的设置命令，自动将统计结果和/或分析图表通过邮件等方式定时发送给指定用户，提高了用户使用的便利性。

[0048] 图2示出了根据本发明的实施例的数据处理系统的框图。

[0049] 如图2所示，根据本发明的实施例的数据处理系统200，包括：设置单元202，根据接收到的设置命令，设置网络爬虫的特征参数和目标网站；任务建立单元204，网络爬虫根据特征参数，对目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列；执行单元206，网络爬虫根据爬虫任务队列中的任务计划对目标网站的数据进行筛选，并存储筛选后的非结构化数据和/或半结构化数据；索引创建单元208，用于为筛选后的非结构化数据和/或半结构化数据创建索引。

[0050] 在该技术方案中，通过Web界面可以对网络爬虫进行多种参数的设置，可以分类大量收集对应每种特征参数的数据信息，通过智能检索目标网站的URL(Uniform Resource Locator，统一资源定位器)及网站结构和层级来创建完善的执行计划，并依此可以做到对收集到的海量信息进行过滤，还可同时存储采集到的结构化或非结构化数据，还可以同时对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析。这样，利用大数据技术搭建的分布式存储系统来存储和处理采集到的数据，可以横向扩展存储容量和并行数据处理能力，保证系统的可用性和高效性，比现有的爬虫技术更具灵活性，这些都是传统的数据库所无法比拟的，可以应用到舆情监控、商业信息收集、市场行情分析、电子商务推广等领域中去。

[0051] 在上述技术方案中，优选地，执行单元206包括：将非结构化数据以文件形式存储在分布式文件系统中；以及通过分布式计算框架将半结构化数据进行分析后转换为结构化数据，并将结构化数据存储在分布式列式存储数据库中。

[0052] 在该技术方案中，非结构化数据包括图片、视频等，半结构化数据包括html、xml等类型的文件，非结构化数据将会以文件形式存储在分布式文件系统(HDFS)中，利用分布式计算框架(MapReduce)对半结构化数据进行分析，并转换为结构化数据存储在分布式列式存储数据库(HBase)中。这样解决了无法对非结构化数据进行存储和分析的问题，实现了对海量半结构化和非结构化内容的采集，成功保留了有价值的数据。

[0053] 在上述技术方案中，优选地，还包括：查询单元210，用于根据索引查询存储后的数

据;以及分析单元212,用于对存储后的数据进行统计和/或分析,并生成统计结果和/或分析图表。

[0054] 在该技术方案中,使用分布式文件系统和分布式列式存储数据库作为搜索引擎技术的底层支撑,利用搜索引擎技术为收集的所有数据建立索引,便于对已有的数据进行快速检索和查询,实现了搜索引擎的分布式索引,用户可以通过Web用户界面对所有采集到的数据进行查询,搜索引擎会快速返回查询结果,还可以对图片、视频等内容进行在线展示,在保证搜索引擎系统高效响应的同时也兼顾了系统整体的易扩容和高可用。另外,数据汇总后会形成有价值的信息,如统计结果和/或分析图表,以供用户读取。

[0055] 在上述技术方案中,优选地,执行单元206还包括:根据用户设置的关键词,自动筛选出与关键词相关的数据;以及分析单元212还包括:为筛选后的与关键词相关的数据生成统计结果和/或分析图表。

[0056] 在该技术方案中,用户还可以使用关键词设置功能对自己感兴趣的内容进行定义,系统会利用分词技术自动匹配与用户设置的关键词相关的内容,用户可以根据系统给出的统计结果和/或分析图表作为参考,对商业和市场行为进行更准确的决策。

[0057] 在上述技术方案中,优选地,还包括:发送单元214,用于将统计结果和/或分析图表在指定时间发送给指定用户。

[0058] 在该技术方案中,系统可以根据用户的设置命令,自动将统计结果和/或分析图表通过邮件等方式定时发送给指定用户,提高了用户使用的便利性。

[0059] 图3示出了根据本发明的一个实施例的数据处理系统的结构示意图。

[0060] 如图3所示,根据本发明的实施例的数据处理系统300,包括:自动化内容采集平台302,可以设置多个目标网站供爬虫爬取大量来自互联网的数据信息,比如,新浪微博、腾讯微博、Twitter和各种资讯网站;大数据处理平台304,可以将非结构化数据将会以文件形式存储在分布式文件系统HDFS中,并利用MapReduce对半结构化数据进行分析,并转换为结构化数据存储在HBase中,这样解决了无法对非结构化数据进行存储和分析的问题,实现了对海量半结构化和非结构化内容的采集,成功保留了有价值的数据;数据统计分析平台306,可以设置网络爬虫的参数,以抓取所需的信息,也可以对采集的数据进行统计分析,使用户可以根据系统给出的统计结果和/或分析图表作为参考,对商业和市场行为进行更准确的决策。

[0061] 大数据处理平台304上具有YARN(Yet Another Resource Negotiator),YARN是一种Hadoop(分布式系统基础架构)的编程模型框架;大数据处理平台304上还具有Solr(搜索应用服务器),用户通过Solr可以对已存储的海量数据进行检索。

[0062] 数据统计分析平台306还具备关键词设置功能,用户还可以使用关键词设置功能对自己感兴趣的内容进行定义,系统会利用分词技术自动匹配与用户设置的关键词相关的内容,用户可以根据系统给出的统计结果和/或分析图表作为参考,对商业和市场行为进行更准确的决策。除此之外,数据统计分析平台306还可以向用户定时发送邮件,该邮件可以包括系统给出的统计结果和/或分析图表。

[0063] 图4示出了根据本发明的另一个实施例的数据处理方法的示意流程图。

[0064] 如图4所示,首先,通过数据统计分析平台406的自定义爬虫功能对智能爬虫的目标网站及参数信息进行设置,智能爬虫会对目标网站的结构和层级进行分析,生成智能爬

虫的任务队列;自动化内容采集平台402设置多个目标网站比如,新浪微博、腾讯微博和各种资讯网站,并依照任务队列中的任务计划启动并发任务,驱动智能爬虫对目标网站的内容进行抓取,并对无效数据进行过滤。

[0065] 过滤后的得到的有效数据被传送至大数据处理平台404,过滤后的有效数据可分为两种格式:非结构化数据,如图片、视频等,和半结构化数据,如html、xml等格式的文件。非结构化数据将会以文件形式存储在HDFS中,半结构化数据会由MapReduce进行分析,并被转换为结构化数据存储在HBase中。同时,所有数据都会通过搜索引擎技术创建索引,大数据处理平台404上具有Solr(搜索应用服务器),用户通过Solr可以对已存储的海量数据进行索引创建和数据检索与查询。

[0066] 用户可以在数据统计分析平台406中通过Web UI(网页用户界面)对所有采集到的数据进行数据统计和采集内容查询,搜索引擎会快速返回查询的结果,并可以对图片、视频等内容进行在线展示。用户还可以在数据统计分析平台406上对统计后的数据的状态进行监控,并查看统计后的数据结果和各种分析图表。

[0067] 用户还可以在数据统计分析平台406实现关键词设置功能,用户对自己感兴趣的内容进行定义,系统就会利用分词技术自动匹配与用户设置的关键词相关的内容,并自动将统计和分析结果通过邮件定时发送的方式或着其他方式定时发送给指定用户,用户可根据邮件中的统计结果或分析图表作为参考,对商业和市场行为进行更准确的决策。

[0068] 以上结合附图详细说明了本发明的技术方案,通过本发明的技术方案,能对网络爬虫的参数进行设置,根据目标网站的结构和层级完善任务计划,还可以同时对采集到的结构化数据和非结构化数据进行收集、过滤、存储、检索和分析,可以横向扩展存储容量和并行数据处理能力,保证系统的可用性和高效性,比现有的爬虫技术更具灵活性,用户还可以根据系统可定时给出的分析结果对商业和市场行为进行更准确的决策,提高了用户使用的便利性。

[0069] 以上所述仅为本发明的优选实施例而已,并不用于限制本发明,对于本领域的技术人员来说,本发明可以有各种更改和变化。凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

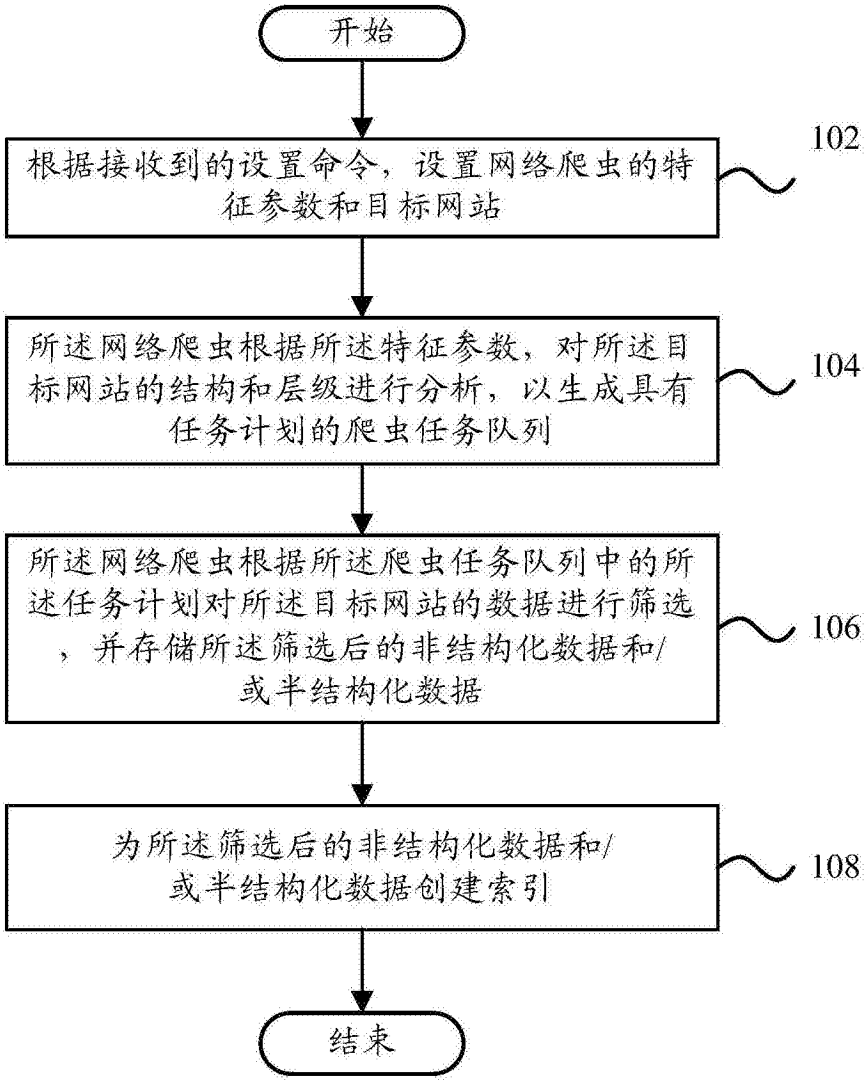

开始
根据接收到的设置命令，设置网络爬虫的特征参数和目标网站
102
所述网络爬虫根据所述特征参数，对所述目标网站的结构和层级进行分析，以生成具有任务计划的爬虫任务队列
104
所述网络爬虫根据所述爬虫任务队列中的所述任务计划对所述目标网站的数据进行筛选，并存储所述筛选后的非结构化数据和/或半结构化数据
106
为所述筛选后的非结构化数据和/或半结构化数据创建索引
108
结束

图1


设置单元202
查询单元210
任务建立单元204
分析单元212
执行单元206
发送单元214
索引创建单元208
数据处理系统200

图2

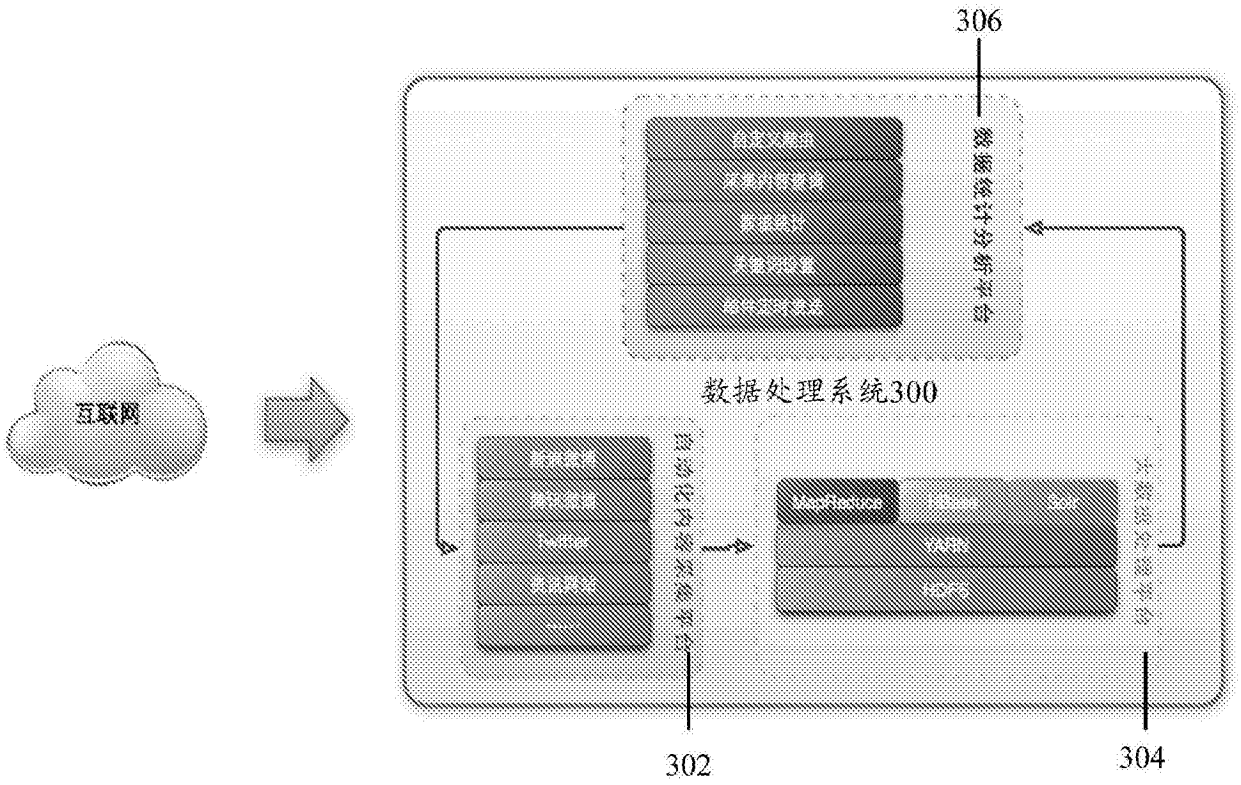

306
互联网
数据处理系统300
302
304

图3

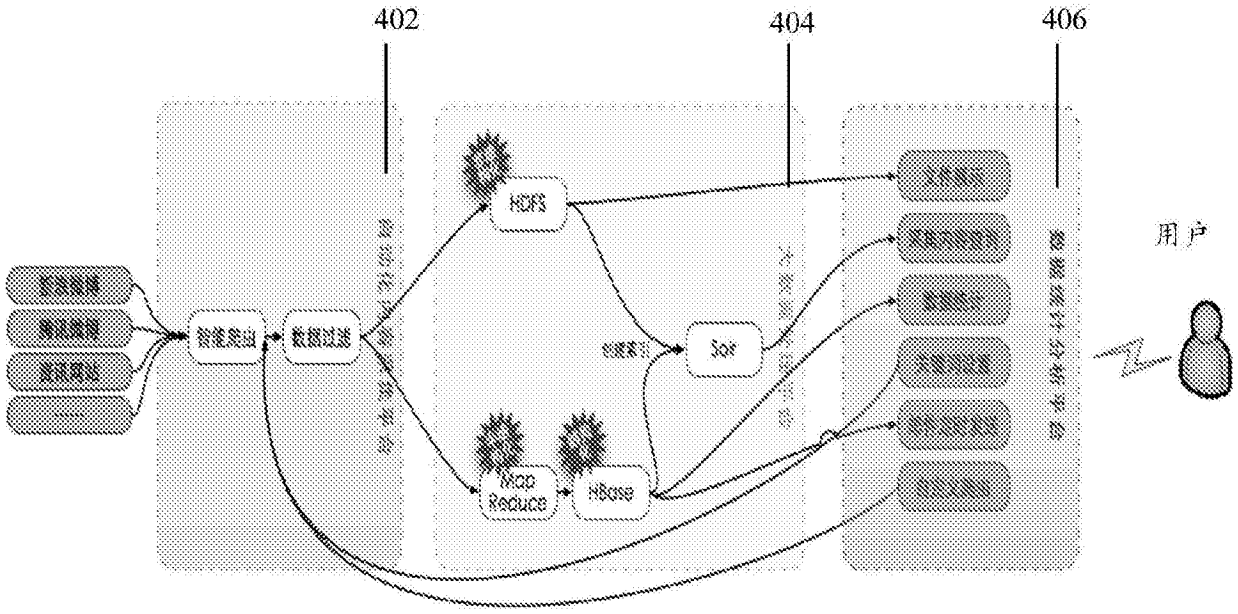

402
404
406
HDFS
Sqoop
Map Reduce
HBase
用户

图4