



(12)发明专利申请

(10)申请公布号 CN 107704535 A

(43)申请公布日 2018.02.16

(21)申请号 201710862871.8

(22)申请日 2017.09.21

(71)申请人 广州大学

地址 510000 广东省广州市番禺广州大学  
城外环西路230号

(72)发明人 胡勇军 李奕臻 谭钻华 刘洁怡

(74)专利代理机构 广州三环专利商标代理有限公司 44202

代理人 梁顺宜 郝传鑫

(51)Int.Cl.

G06F 17/30(2006.01)

H04L 29/08(2006.01)

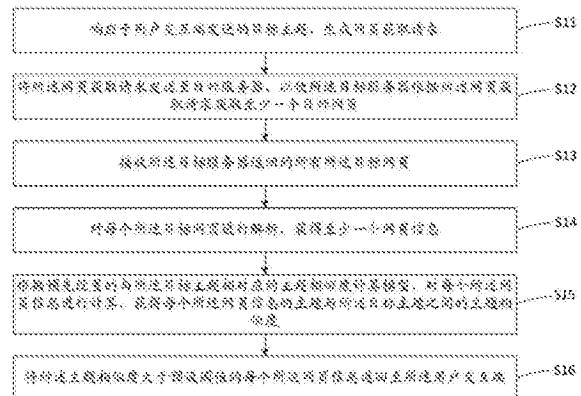
权利要求书2页 说明书8页 附图2页

(54)发明名称

基于主题相似度的网页信息获取方法、装置及系统

(57)摘要

本发明公开了一种基于主题相似度的网页信息获取方法、装置及系统。所述基于主题相似度的网页信息获取方法包括：响应于用户交互端发送的目标主题，生成网页获取请求；将所述网页获取请求发送至目标服务器；接收所述目标服务器返回的所有所述目标网页；对每个所述目标网页进行解析，获得至少一个网页信息；根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度；将所述主题相似度大于预设阈值的每个所述网页信息返回至所述用户交互端。采用本发明，能够提高所获取的网页信息的针对性和准确度。



1. 一种基于主题相似度的网页信息获取方法，其特征在于，包括：
  - 响应于用户交互端发送的目标主题，生成网页获取请求；
  - 将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页；
  - 接收所述目标服务器返回的所有所述目标网页；
  - 对每个所述目标网页进行解析，获得至少一个网页信息；
  - 根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度；
  - 将所述主题相似度大于预设阈值的每个所述网页信息返回至所述用户交互端。
2. 如权利要求1所述的基于主题相似度的网页信息获取方法，其特征在于，在所述将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页之前，还包括：
  - 对与本地相连的每个服务器的运行状态进行检测，并将其中运行状态为空闲的任意一个服务器设置为所述目标服务器。
3. 如权利要求1所述的基于主题相似度的网页信息获取方法，其特征在于，所述网页获取请求中包含预先设置的目标网页列表中的各个网页地址；
  - 则所述将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页，具体包括：
    - 将所述网页获取请求发送至所述目标服务器，以使所述目标服务器根据所述网页获取请求中的每个所述网页地址查找到对应的所述目标网页。
4. 如权利要求1所述的基于主题相似度的网页信息获取方法，其特征在于，所述目标网页为HTML格式的网页；所述网页信息为所述目标网页中的ASCII码文本内容。
5. 如权利要求1所述的基于主题相似度的网页信息获取方法，其特征在于，所述主题相似度计算模型包括主题生成模型和词向量获取模型；
  - 则所述根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度，还包括：
    - 获取与所述目标主题相对应的所述主题相似度计算模型；
    - 利用所述主题相似度计算模型中的主题生成模型对每个所述网页信息进行计算，获得每个所述网页信息的主题；
    - 根据所述主题相似度计算模型中的词向量获取模型，对每个所述网页信息的主题分别与所述目标主题进行余弦相似度计算，获得每个所述网页信息的主题与所述目标主题的所述主题相似度。
6. 如权利要求5所述的基于主题相似度的网页信息获取方法，其特征在于，在所述根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度之前，还包括：
  - 接收所述用户交互端发送的目标主题信息；
  - 根据所述目标主题和所述目标主题信息训练生成所述主题生成模型。
7. 如权利要求5或6所述的基于主题相似度的网页信息获取方法，其特征在于，所述主

题生成模型为LDA模型；所述词向量获取模型为Word2vec模型。

8. 一种基于主题相似度的网页信息获取装置，其特征在于，包括：

网页获取请求生成模块，用于响应于用户交互端发送的目标主题，生成网页获取请求；

网页获取请求发送模块，用于将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页；

目标网页接收模块，用于接收所述目标服务器返回的所有所述目标网页；

网页信息提取模块，用于对每个所述目标网页进行解析，获得至少一个网页信息；

主题相似度计算模块，用于根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度；以及，

网页信息反馈模块，用于将所述主题相似度大于预设阈值的每个所述网页信息返回至所述用户交互端。

9. 如权利要求8所述的基于主题相似度的网页信息获取装置，其特征在于，所述基于主题相似度的网页信息获取装置，还包括：

目标服务器确定模块，用于对与本地相连的每个服务器的运行状态进行检测，并将其中运行状态为空闲的任意一个服务器设置为所述目标服务器。

10. 一种基于主题相似度的网页信息获取系统，其特征在于，包括客户端和服务器端；其中，所述客户端中包括用户交互端和数据处理端；

所述用户交互端，用于与用户进行交互；

所述数据处理端，为如权利要求8或9所述的基于主题相似度的网页信息获取装置；

所述服务器端，其中包含至少一个服务器，用于存储和管理网页。

## 基于主题相似度的网页信息获取方法、装置及系统

### 技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种基于主题相似度的网页信息获取方法、装置及系统。

### 背景技术

[0002] 网络爬虫，是一种智能程序，它根据给定策略，智能抓取互联网上各类信息，常在搜索引擎中作为搜索引擎的核心之一。网络爬虫通过预置的种子URL，利用网络访问引擎发送HTTP网络协议来进行网页访问与内容抓取，然后以抓取到的URL为新的起点，继续爬取。网络爬虫一般会往高效高可用方向发展，即在下载尽可能多的相关性高的实用信息的同时，消耗尽可能短的时间。

[0003] 然而，传统的通用网络爬虫负责面对所有的用户查询需求，不断地抓取全互联网的信息，返回的结果过于繁多，有时候不太适合特定信息的需求者使用。这类网络爬虫的抓取结果，往往追求大而全的结果，缺少清晰化的、领域相关的模型，也缺少精准化的搜索结果。

### 发明内容

[0004] 本发明实施例提出一种基于主题相似度的网页信息获取方法、装置及系统，能够提高所获取的网页信息的针对性和准确度。

[0005] 本发明实施例提供的一种基于主题相似度的网页信息获取方法，具体包括：

[0006] 响应于用户交互端发送的目标主题，生成网页获取请求；

[0007] 将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页；

[0008] 接收所述目标服务器返回的所有所述目标网页；

[0009] 对每个所述目标网页进行解析，获得至少一个网页信息；

[0010] 根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度；

[0011] 将所述主题相似度大于预设阈值的每个所述网页信息返回至所述用户交互端。

[0012] 进一步地，在所述将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页之前，还包括：

[0013] 对与本地相连的每个服务器的运行状态进行检测，并将其中运行状态为空闲的任意一个服务器设置为所述目标服务器。

[0014] 进一步地，所述网页获取请求中包含预先设置的目标网页列表中的各个网页地址；

[0015] 则所述将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页，具体包括：

[0016] 将所述网页获取请求发送至所述目标服务器，以使所述目标服务器根据所述网页

获取请求中的每个所述网页地址查找到对应的所述目标网页。

[0017] 进一步地，所述目标网页为HTML格式的网页；所述网页信息为所述目标网页中的ASCII码文本内容。

[0018] 进一步地，所述主题相似度计算模型包括主题生成模型和词向量获取模型；

[0019] 则所述根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度，具体包括：

[0020] 获取与所述目标主题相对应的所述主题相似度计算模型；

[0021] 利用所述主题相似度计算模型中的主题生成模型对每个所述网页信息进行计算，获得每个所述网页信息的主题；

[0022] 根据所述主题相似度计算模型中的词向量获取模型，对每个所述网页信息的主题分别与所述目标主题进行余弦相似度计算，获得每个所述网页信息的主题与所述目标主题的所述主题相似度。

[0023] 进一步地，在所述根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度之前，还包括：

[0024] 接收所述用户交互端发送的目标主题信息；

[0025] 根据所述目标主题和所述目标主题信息训练生成所述主题生成模型。

[0026] 进一步地，所述主题生成模型为LDA模型；所述词向量获取模型为Word2vec模型。

[0027] 相应地，本发明实施例还提供了一种基于主题相似度的网页信息获取装置，具体包括：

[0028] 网页获取请求生成模块，用于响应于用户交互端发送的目标主题，生成网页获取请求；

[0029] 网页获取请求发送模块，用于将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页；

[0030] 目标网页接收模块，用于接收所述目标服务器返回的所有所述目标网页；

[0031] 网页信息提取模块，用于对每个所述目标网页进行解析，获得至少一个网页信息；

[0032] 主题相似度计算模块，用于根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度；以及，

[0033] 网页信息反馈模块，用于将所述主题相似度大于预设阈值的每个所述网页信息返回至所述用户交互端。

[0034] 进一步地，所述基于主题相似度的网页信息获取装置，还包括：

[0035] 目标服务器确定模块，用于对与本地相连的每个服务器的运行状态进行检测，并将其中运行状态为空闲的任意一个服务器设置为所述目标服务器。

[0036] 相应地，本发明实施例还提供了一种基于主题相似度的网页信息获取系统，具体包括客户端和服务器端；其中，所述客户端中包括用户交互端和数据处理端；

[0037] 所述用户交互端，用于与用户进行交互；

[0038] 所述数据处理端，为如上所述的基于主题相似度的网页信息获取装置；

[0039] 所述服务器端，其中包含至少一个服务器，用于存储和管理网页。

[0040] 实施本发明实施例，具有如下有益效果：

[0041] 本发明实施例提供的基于主题相似度的网页信息获取方法、装置及系统，通过对与本地相连的各个服务器的运行状态进行检测，并将其中运行状态为空闲的服务器设置为目标服务器，从而能够避免在网页获取过程中出现排队等待响应的现象，从而提高网页获取过程的效率，进而提高网页信息获取的效率，提高用户体验。

## 附图说明

[0042] 图1是本发明提供的基于主题相似度的网页信息获取方法的一个优选的实施例的流程示意图；

[0043] 图2是本发明提供的基于主题相似度的网页信息获取装置的一个优选的实施例的结构示意图；

[0044] 图3是本发明提供的基于主题相似度的网页信息获取系统的一个优选的实施例的结构示意图。

## 具体实施方式

[0045] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，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作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。

[0046] 本发明通过预生成若干主题相似度计算模型，并通过利用与用户输入的目标主题相对应的主题相似度计算模型计算获取的网页信息的主题和用户输入的目标主题之间的主题相似度，从而判断所获取的网页信息是否为用户想要获取的网页信息，并将其中符合用户输入的目标主题的网页信息返回给用户。本发明通过计算获取的网页信息的主题和用户输入的目标主题之间的主题相似度，从而能够从获取的所有网页信息中筛选出用户想要获取的网页信息，提高所获取的网页信息的针对性和准确度。

[0047] 如图1所示，为本发明提供的基于主题相似度的网页信息获取方法的一个优选的实施例的流程示意图，包括步骤S11至S16，具体如下：

[0048] S11：响应于用户交互端发送的目标主题，生成网页获取请求；

[0049] S12：将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页；

[0050] S13：接收所述目标服务器返回的所有所述目标网页；

[0051] S14：对每个所述目标网页进行解析，获得至少一个网页信息；

[0052] S15：根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度；

[0053] S16：将所述主题相似度大于预设阈值的每个所述网页信息返回至所述用户交互端。

[0054] 需要说明的是，本发明实施例由终端设备中的数据处理模块执行，该数据处理模块中包含网络访问内核、网页处理器以及主题匹配器。该终端设备中还包括用户交互模块，

该用户交互模块中包含自然语言处理工具和爬虫操作接口。

[0055] 该终端设备在对网页信息的主题相似度进行计算之前,预先生成一个或者多个主题相似度计算模型,并将这些主题相似度计算模型存储于本地。

[0056] 当用户需要获取某一主题的网页信息时,将欲获取的目标主题通过用户交互端发送给上述终端设备。该用户交互端在接收到该目标主题之后,利用自然语言处理工具对该目标主题进行解析和基本的自然语言处理,并将经过解析和处理的目标主题输入至爬虫操作接口。该爬虫操作接口在接收到前述目标主题后,调用网络访问内核,并将该目标主题发送至该网络访问内核。该网络访问内核在接收到爬虫操作接口发送的目标主题之后,根据该目标主题生成网页获取请求,并将该网页获取请求发送给目标服务器。其中,该网页获取请求的格式一般为HTTP请求格式。该目标服务器在接收到该网页获取请求之后,根据该网页获取请求在本地中获取一个或者多个目标网页,并将获取的各个目标网页返回至上述网络访问内核。上述网络访问内核在接收到目标服务器返回的一个或者多个目标网页之后,将这些目标网页传送至网页处理器中进行分析和处理,从而从中获得一个或者多个网页信息,并将这些网页信息传送至主题匹配器中。该主题匹配器在接收到网页处理器传送过来的网页信息之后,从预先存储的主题相似度计算模型中获取与上述用户输入的目标主题相对应的主题相似度计算模型,并利用该主题相似度计算模型对每个接收到的网页信息依次进行计算,从而获得每个网页信息的主题和上述目标主题的主题相似度。随后,该主题匹配器将接收到的网页信息中主题与上述目标主题的主题相似度大于预设的阈值网页信息返回至上述爬虫操作接口,从而使得该爬虫操作接口将这些网页信息发送至用户交互端的显示屏中显示,从而使得用户获得与上述目标主题相关的网页信息。

[0057] 本发明实施例通过计算获取的网页信息的主题和用户输入的目标主题之间的主题相似度,从而能够从获取的所有网页信息中筛选出用户想要获取的网页信息,提高所获取的网页信息的针对性和准确度。

[0058] 在另一个优选的实施例中,在上述实施例的基础上,在所述将所述网页获取请求发送至目标服务器,以使所述目标服务器根据所述网页获取请求获取至少一个目标网页之前,还包括:

[0059] 对与本地相连的每个服务器的运行状态进行检测,并将其中运行状态为空闲的任意一个服务器设置为所述目标服务器。

[0060] 需要说明的是,上述数据处理模块中还包含资源调度器,用于对各个服务器的资源进行调度。具体地,在上述网络访问内核生成网页获取请求之后,该资源调度器对与本地相连的各个服务器的资源进行检测,从而判断各个服务器的运行状态是否为空闲或者忙碌,并从中选出运行状态为空闲的一个或者多个服务器,并将这些运行状态为空闲的服务器中的任意一个服务器设置为目标服务器,从而使得上述网络访问内核将生成的网页获取请求发送至该目标服务器以获得一个或者多个目标网页。

[0061] 需要进一步说明的是,资源调度器在对与本地相连的各个服务器的运行状态进行检测的过程中,在接收各个服务器的响应的同时,还会接收各个服务器返回的“Set-Cookie”信息,并将这些“Set-Cookie”信息存入本地的Cookie池中,以便于后续的服务器访问和通信。在与服务器通信的过程中,可以从该Cookie池中随机选取一个“Set-Cookie”信息,并对对应的服务器进行访问,若访问失败的次数超过预设阈值,则该“Set-Cookie”信息

失效，因此在该Cookie池中重新选取一个“Set-Cookie”信息，重新进行服务器的访问，从而提高服务器访问和通信的成功率。其中，需要说明的是，上述“Set-Cookie”信息为一种用于鉴别不同用户身份的键值信息对集合，可用于设置生成Cookie；Cookie池为若干个Cookie的集合。

[0062] 本发明实施例通过对与本地相连的各个服务器的运行状态进行检测，并将其中运行状态为空闲的服务器设置为目标服务器，从而能够避免在网页获取过程中出现排队等待响应的现象，从而提高网页获取过程的效率，进而提高网页信息获取的效率，提高用户体验。

[0063] 在又一个优选的实施例中，在上述实施例的基础上，所述网页获取请求中包含预先设置的目标网页列表中的各个网页地址；

[0064] 则所述将所述网页获取请求发送至目标服务器，以使所述目标服务器根据所述网页获取请求获取至少一个目标网页，具体包括：

[0065] 将所述网页获取请求发送至所述目标服务器，以使所述目标服务器根据所述网页获取请求中的每个所述网页地址查找到对应的所述目标网页。

[0066] 需要说明的是，上述终端设备中还存储有目标网页列表，用于记录欲获取的网页信息所在的网页的网页地址。具体地，上述网页获取请求中还包含上述目标网页列表中记录的各个网页地址，因此，在上述网络访问内核将该网页获取请求发送至目标服务器之后，该目标服务器在接收到该网页获取请求之后，根据该网页获取请求触发获取目标网页的操作，查找获得与该网页获取请求中的各个网页地址分别相对应的目标网页，并将各个目标网页返回至上述网络访问内核。

[0067] 更优选地，所述目标网页为HTML格式的网页；所述网页信息为所述目标网页中的ASCII码文本内容。

[0068] 需要说明的是，上述网络访问内核从目标服务器中获取的目标网页为HTML格式的网页。上述网页处理器从目标网页中解析获得的页面信息为该目标网页中的ASCII码文本内容。

[0069] 在又一个优选的实施例中，在上述实施例的基础上，所述主题相似度计算模型包括主题生成模型和词向量获取模型；

[0070] 则所述根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度，具体包括：

[0071] 获取与所述目标主题相对应的所述主题相似度计算模型；

[0072] 利用所述主题相似度计算模型中的主题生成模型对每个所述网页信息进行计算，获得每个所述网页信息的主题；

[0073] 根据所述主题相似度计算模型中的词向量获取模型，对每个所述网页信息的主题分别与所述目标主题进行余弦相似度计算，获得每个所述网页信息的主题与所述目标主题的所述主题相似度。

[0074] 进一步地，在所述根据预先设置的与所述目标主题相对应的主题相似度计算模型，对每个所述网页信息进行计算，获得每个所述网页信息的主题与所述目标主题之间的主题相似度之前，还包括：

- [0075] 接收所述用户交互端发送的目标主题信息；  
[0076] 根据所述目标主题和所述目标主题信息训练生成所述主题生成模型。  
[0077] 更优选地，所述主题生成模型为LDA模型；所述词向量获取模型为Word2vec模型。  
[0078] 需要说明的是，上述主题相似度计算模型中包含主题生成模型和词向量获取模型。上述主题匹配器在接收到网页处理器传送过来的网页信息之后，获取与用户输入的目标主题相对应的主题相似度计算模型，并利用该主题相似度计算模型中的主题生成模型对每个网页信息进行计算，从而获得每个网页信息的主题。随后，上述主题匹配器利用上述主题相似度计算模型中的词向量获取模型对计算获得的各个网页信息的主题进行解析，并按照余弦相似度计算方法计算获得各个网页信息的主题和上述目标主题之间的主题相似度。最后，该主题匹配器将这些网页信息中主题与上述目标主题之间的主题相似度大于预设的阈值网页信息返回至上述爬虫操作接口，从而使得该爬虫操作接口将这些网页信息发送至用户交互端的显示屏中显示，从而使得用户获得与上述目标主题相关的网页信息。  
[0079] 在一些更优选的实施例中，终端设备还可以根据各个网页信息的主题和预先设置的各个目标主题之间的主题相似度来对各个网页信息进行分类。具体地，在计算网页信息的主题和各个目标主题之间的主题相似度之前，还可以首先对该网页信息的主题类型进行预测，从而提高对网页信息进行分类的准确度，具体步骤如下：  
[0080] S1：对网页信息进行预处理，包括噪声消除、分词、去除停用词，删除一些与主题无关或者重复的信息，得到相对规范和整洁的数据；  
[0081] S2：特征提取，获取能够代表上述网页信息的最小特征项集合，降低特征空间维度；  
[0082] S3：选择一个“合适”的外部辅助语料库，为辅助语料库进行主题分析，建立主题模型和主题特征描述；  
[0083] S4：利用S3中已经建立好的主题模型为上述网页信息进行主题推断，得到上述网页信息的文档-主题概率分布矩阵，以此来表达网页信息的结构特征；  
[0084] S5：利用S4所得到的文档-主题概率以及S3中外部辅助语料库所得到的主题特征描述对网页信息进行特征扩展；  
[0085] S6：应用支持向量机(SVM)训练分类模型，预测网页信息所属的主题类别。  
[0086] 需要进一步说明的是，上述主题生成模型可以根据用户输入的目标主题和与该目标主题相关的目标主题信息训练生成。  
[0087] 在一些更优选的实施例中，上述主题生成模型可以为LDA模型(Latent Dirichlet Allocation，文档主题生成模型，又称三层贝叶斯概率模型)，上述词向量获取模型为Word2vec模型。其中，每个LDA模型可以通过对每个主题类别的高频词进行训练获得，具体地，通过抽取每个类别的高频词作为向量空间模型的特征空间，用TF-IDF方法将短文本表示成向量，再利用初始的LDA模型得到每个文本的隐主题特征，将概率大于某一阈值的隐主题对应的高频词扩展到文本中，以降低短文本的噪声和稀疏性影响。  
[0088] 在一些更优选的实施例中，还可以通过将LDA模型和pagerank(网页排名，又称网页级别、Google左侧排名或佩奇排名)技术相结合的方式进行网页信息的爬取，从而进一步提高网页信息获取的针对性和准确度。具体地，通过采用pagerank技术，通过分析网页的链接，赋予每个网站不同的链接权重，具体地，重要的网站链接权重更大。同时，通过采用上述

LDA模型对网页信息进行主题相似度计算,对不同的网页赋予不同的分类权重,具体地,与上述目标主题的主题相似度大的分类权重更大。最后,将网页信息的链接权重和分类权重进行综合,从中选择与上述目标主题最相关的网页信息反馈给用户。

[0089] 本发明实施例提供的基于主题相似度的网页信息获取方法,通过对与本地相连的各个服务器的运行状态进行检测,并将其中运行状态为空闲的服务器设置为目标服务器,从而能够避免在网页获取过程中出现排队等待响应的现象,从而提高网页获取过程的效率,进而提高网页信息获取的效率,提高用户体验。另外,通过对与本地相连的各个服务器的运行状态进行检测,并将其中运行状态为空闲的服务器设置为目标服务器,从而能够避免在网页获取过程中出现排队等待响应的现象,从而提高网页获取过程的效率,进而提高网页信息获取的效率,提高用户体验。

[0090] 相应地,本发明还提供一种基于主题相似度的网页信息获取装置,能够实现上述实施例中的基于主题相似度的网页信息获取方法的所有流程。

[0091] 如图2所示,为本发明提供的基于主题相似度的网页信息获取装置的一个优选的实施例的结构示意图,具体包括:

[0092] 网页获取请求生成模块21,用于响应于用户交互端发送的目标主题,生成网页获取请求;

[0093] 网页获取请求发送模块22,用于将所述网页获取请求发送至目标服务器,以使所述目标服务器根据所述网页获取请求获取至少一个目标网页;

[0094] 目标网页接收模块23,用于接收所述目标服务器返回的所有所述目标网页;

[0095] 网页信息提取模块24,用于对每个所述目标网页进行解析,获得至少一个网页信息;

[0096] 主题相似度计算模块25,用于根据预先设置的与所述目标主题相对应的主题相似度计算模型,对每个所述网页信息进行计算,获得每个所述网页信息的主题与所述目标主题之间的主题相似度;以及,

[0097] 网页信息反馈模块26,用于将所述主题相似度大于预设阈值的每个所述网页信息返回至所述用户交互端。

[0098] 在另一个优选的实施例中,在上述实施例的基础上,所述基于主题相似度的网页信息获取装置,还包括:

[0099] 目标服务器确定模块,用于对与本地相连的每个服务器的运行状态进行检测,并将其中运行状态为空闲的任意一个服务器设置为所述目标服务器。

[0100] 在又一个优选的实施例中,在上述实施例的基础上,所述网页获取请求中包含预先设置的目标网页列表中的各个网页地址;

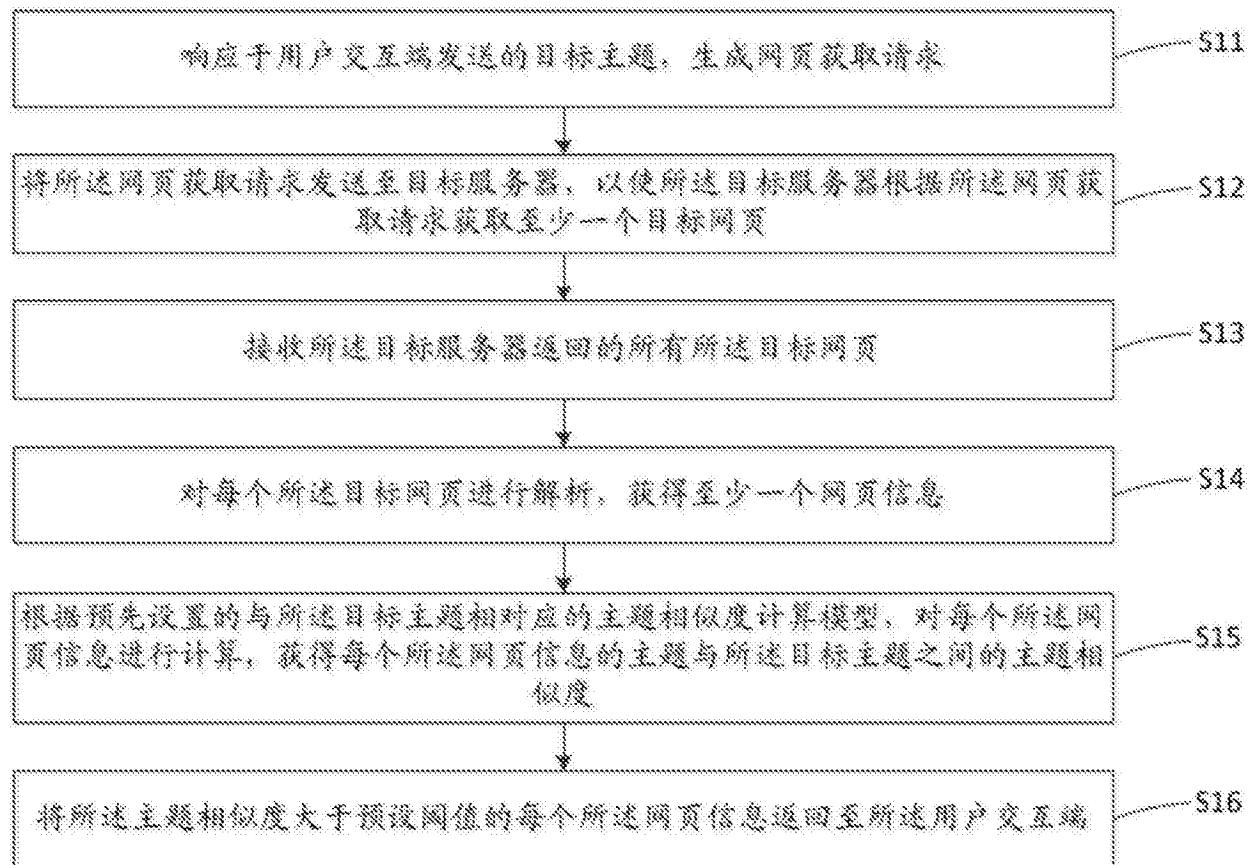
[0101] 则所述网页获取请求发送模块,具体包括:

[0102] 请求发送单元,用于将所述网页获取请求发送至所述目标服务器,以使所述目标服务器根据所述网页获取请求中的每个所述网页地址查找到对应的所述目标网页。

[0103] 更优选地,所述目标网页为HTML格式的网页;所述网页信息为所述目标网页中的ASCII码文本内容。

[0104] 在又一个优选的实施例中,在上述实施例的基础上,所述主题相似度计算模型包括主题生成模型和词向量获取模型;

- [0105] 则所述主题相似度计算模块，具体包括：
- [0106] 主题相似度计算模型获取单元，用于获取与所述目标主题相对应的所述主题相似度计算模型；
- [0107] 网页信息主题计算获得单元，用于利用所述主题相似度计算模型中的主题生成模型对每个所述网页信息进行计算，获得每个所述网页信息的主题；以及，
- [0108] 主题相似度计算获得单元，用于根据所述主题相似度计算模型中的词向量获取模型，对每个所述网页信息的主题分别与所述目标主题进行余弦相似度计算，获得每个所述网页信息的主题与所述目标主题的所述主题相似度。
- [0109] 进一步地，所述基于主题相似度的网页信息获取装置，还包括：
- [0110] 目标主题信息接收模块，用于接收所述用户交互端发送的目标主题信息；以及，
- [0111] 主题生成模型训练模块，用于根据所述目标主题和所述目标主题信息训练生成所述主题生成模型。
- [0112] 更优选地，所述主题生成模型为LDA模型；所述词向量获取模型为Word2vec模型。
- [0113] 本发明实施例提供的基于主题相似度的网页信息获取装置，通过对与本地相连的各个服务器的运行状态进行检测，并将其中运行状态为空闲的服务器设置为目标服务器，从而能够避免在网页获取过程中出现排队等待响应的现象，从而提高网页获取过程的效率，进而提高网页信息获取的效率，提高用户体验。另外，通过对与本地相连的各个服务器的运行状态进行检测，并将其中运行状态为空闲的服务器设置为目标服务器，从而能够避免在网页获取过程中出现排队等待响应的现象，从而提高网页获取过程的效率，进而提高网页信息获取的效率，提高用户体验。
- [0114] 相应地，本发明还提供一种基于主题相似度的网页信息获取系统。
- [0115] 如图3所示，为本发明提供的基于主题相似度的网页信息获取系统的一个优选的实施例的结构示意图，具体包括客户端31和服务器端32；其中，所述客户端31中包括用户交互端311和数据处理端312；
- [0116] 所述用户交互端311，用于与用户进行交互；
- [0117] 所述数据处理端312，为如上任一实施例所述的基于主题相似度的网页信息获取装置；
- [0118] 所述服务器端32，其中包含至少一个服务器，用于存储和管理网页。
- [0119] 本发明实施例提供的基于主题相似度的网页信息获取系统，通过对与本地相连的各个服务器的运行状态进行检测，并将其中运行状态为空闲的服务器设置为目标服务器，从而能够避免在网页获取过程中出现排队等待响应的现象，从而提高网页获取过程的效率，进而提高网页信息获取的效率，提高用户体验。另外，通过对与本地相连的各个服务器的运行状态进行检测，并将其中运行状态为空闲的服务器设置为目标服务器，从而能够避免在网页获取过程中出现排队等待响应的现象，从而提高网页获取过程的效率，进而提高网页信息获取的效率，提高用户体验。
- [0120] 以上所述是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以作出若干改进和润饰，这些改进和润饰也视为本发明的保护范围。



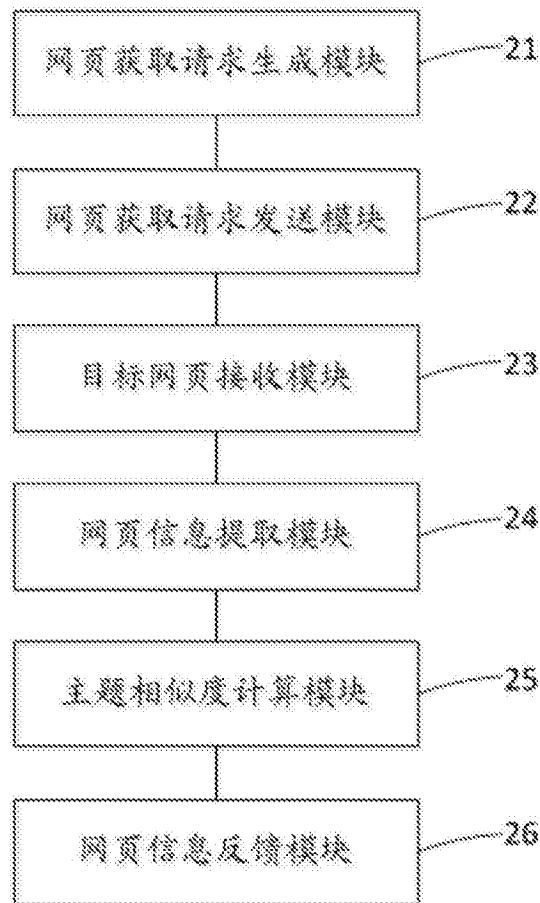


图2

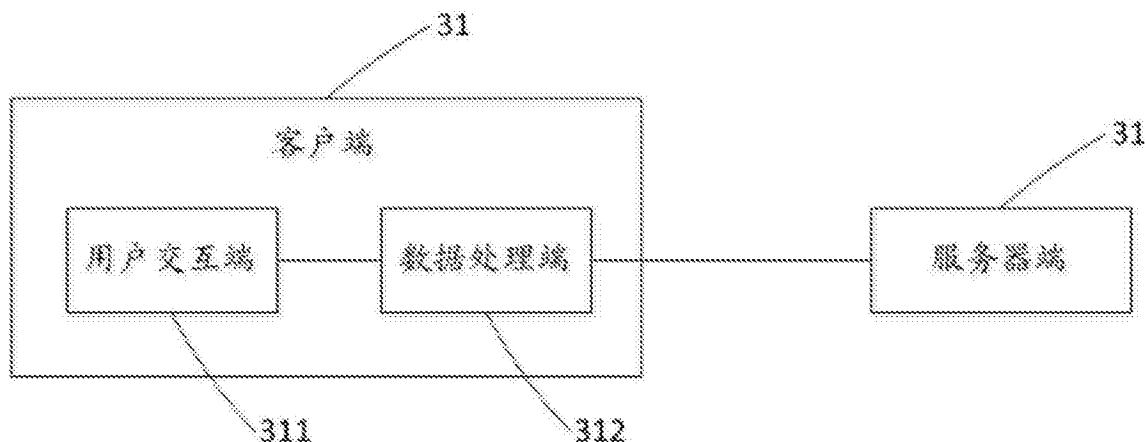


图3