(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102347930 A
(43)申请公布日 2012.02.08

(21)申请号 201010240330.X

(22)申请日 2010.07.26

(71)申请人 中国电信股份有限公司
地址 100032 北京市西城区金融大街 31 号

(72)发明人 王爱宝 张涛 李屹 杨德利

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 孙宝海

(51)Int.Cl.
H04L 29/06(2006.01)
H04L 29/08(2006.01)
G06F 17/30(2006.01)

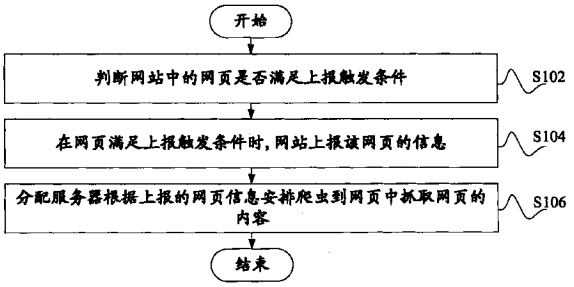
权利要求书 1 页 说明书 6 页 附图 3 页

(54)发明名称

网页内容获取方法和系统

(57)摘要

本发明公开了一种网页内容获取的方法与系统。其中，该方法包括判断网站中的网页是否满足上报触发条件；在网页满足上报触发条件时，网站上报网页信息；分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容。本发明在网页满足上报触发条件时上报网页信息，爬虫根据网页信息到指定网页中抓取网页的内容。该方法节约了爬虫的工作量，缓解了目标网站的压力，并且增加了获取实时信息的能力，为实时搜索提供了有利的条件。



CN 102347930 A

1. 一种网页内容获取方法，其特征在于，所述方法包括：

判断网站中的网页是否满足上报触发条件；

在所述网页满足所述上报触发条件时，所述网站上报网页信息；

分配服务器根据上报的所述网页信息安排爬虫到所述网页中抓取所述网页的内容。

2. 根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

根据网站类型的不同，为每类网站设置不同的上报触发条件。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述上报触发条件包括发表新文章触发上报、文章内容更新触发上报、依据回帖数量触发上报、依据浏览量触发上报以及定时触发上报中的至少一种。

4. 根据权利要求 1 所述的方法，其特征在于，所述网页信息包括所述网页的 URL 地址、所述网页的关键词、所述网页的摘要以及所述网页的 SP 信息。

5. 根据权利要求 1 所述的方法，其特征在于，所述分配服务器根据上报的所述网页信息安排爬虫到所述网页中抓取所述网页的内容的步骤包括：

所述分配服务器从所述网页信息中提取所述网页的 URL 地址；

调用分配优先级策略；

将接收到的符合所述分配优先级策略的 URL 地址发送给所述爬虫；

所述爬虫从接收到的 URL 地址中抓取所述网页的内容。

6. 根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

将上报的所述网页信息和获取的所述网页的内容发送给索引服务器。

7. 一种网页内容获取系统，其特征在于，所述系统包括：

内容触发服务器，用于判断网站中的网页是否满足上报触发条件，如果满足所述上报触发条件，则上报网页信息；

分配服务器，与所述内容触发服务器相连，用于根据上报的所述网页信息安排爬虫到所述网页中抓取所述网页的内容。

8. 根据权利要求 7 所述的系统，其特征在于，所述内容触发服务器还用于：

根据网站类型的不同，为每类网站设置不同的上报触发条件。

9. 根据权利要求 7 或 8 所述的系统，其特征在于，所述上报触发条件包括发表新文章触发上报、文章内容更新触发上报、依据回帖数量触发上报、依据浏览量触发上报以及定时触发上报中的至少一种。

10. 根据权利要求 7 所述的系统，其特征在于，所述网页信息包括所述网页的 URL 地址、所述网页的关键词、所述网页的摘要以及所述网页的 SP 信息。

11. 根据权利要求 7 所述的系统，其特征在于，所述分配服务器包括：

地址提取模块，用于从所述网页信息中提取所述网页的 URL 地址；

策略调用模块，用于调用分配优先级策略；

抓取分配模块，分别与所述地址提取模块和所述策略调用模块相连，用于将提取的符合所述分配优先级策略的 URL 地址发送给所述爬虫以抓取所述网页的内容。

12. 根据权利要求 7 所述的系统，其特征在于，所述系统还包括：

索引服务器，与所述分配服务器相连，用于对上报的所述网页信息和获取的所述网页的内容进行分类并建立关系索引。

网页内容获取方法和系统

技术领域

[0001] 本发明涉及信息检索领域，更具体地，涉及一种网页内容获取方法和系统。

背景技术

[0002] 随着网络上大量涌现的博客、微博等网站，使得用户对获取网络内容的实时性有了很高的要求，而面对突发的海量信息的管理更是将实时信息的获取推上了最重要的位置。

[0003] 目前，搜索引擎获取网页信息的方式多采用爬虫抓取，随着网页中包含的链接不停地传递获取下去，导致对信息获取效率的降低，浪费了大量资源。并且，这种获取网页的方式因为访问量极大，对于同一位置网页内容的更新不能在第一时间获取，基本不能实现实时信息的呈现。

[0004] 具体地，爬虫抓取方法中存在的下述问题严重地影响了对网页信息的实时获取：(1) 爬虫获取需要大量冗余地抓取不相关或者重复的网页，效率很低；(2) 无法实时获取网页的内容更新；(3) 为了获取同一网页的信息需要反复访问该网页，对服务器和带宽产生了巨大的压力。

[0005] 可见，传统的网页信息获取方式无法实现实时信息的获取，从而不能满足实际使用的需要。

发明内容

[0006] 本发明要解决的一个技术问题是提供一种网页内容获取方法，能够实现实时信息的获取。

[0007] 本发明提供了一种网页内容获取方法，包括判断网站中的网页是否满足上报触发条件；在网页满足上报触发条件时，网站上报网页信息；分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容。

[0008] 根据本发明方法的一个实施例，该方法还包括：根据网站类型的不同，为每类网站设置不同的上报触发条件。

[0009] 根据本发明方法的另一实施例，上报触发条件包括发表新文章触发上报、文章内容更新触发上报、依据回帖数量触发上报、依据浏览量触发上报以及定时触发上报中的至少一种。

[0010] 根据本发明方法的又一实施例，网页信息包括网页的同一资源定位符 (Uniform Resource Locator，URL) 地址、网页的关键词、网页的摘要以及网页的服务提供商 (Service Provider，SP) 信息。

[0011] 根据本发明方法的再一实施例，分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容的步骤包括：分配服务器从网页信息中提取网页的 URL 地址；调用分配优先级策略；将接收到的符合分配优先级策略的 URL 地址发送给爬虫；爬虫从接收到的 URL 地址中抓取网页的内容。

[0012] 根据本发明方法的再一实施例，该方法还包括：将上报的网页信息和获取的网页的内容发送给索引服务器。

[0013] 本发明的网页内容获取方法，在网页满足上报触发条件时上报网页信息，爬虫根据网页信息到指定网页中抓取网页的内容。该方法节约了爬虫的工作量，缓解了目标网站的压力，并且增加了获取实时信息的能力，为实时搜索提供了有利的条件。

[0014] 本发明要解决的另一技术问题是提供一种网页内容获取系统，能够实现实时信息的获取。

[0015] 本发明提供了一种网页内容获取系统，包括：内容触发服务器，用于判断网站中的网页是否满足上报触发条件，如果满足上报触发条件，则上报网页信息；分配服务器，与内容触发服务器相连，用于根据上报的网页信息安排爬虫到网页中抓取网页的内容。

[0016] 根据本发明系统的一个实施例，内容触发服务器还用于：根据网站类型的不同，为每类网站设置不同的上报触发条件。

[0017] 根据本发明系统的另一实施例，上报触发条件包括发表新文章触发上报、文章内容更新触发上报、依据回帖数量触发上报、依据浏览量触发上报以及定时触发上报中的至少一种。

[0018] 根据本发明系统的又一实施例，网页信息包括网页的 URL 地址、网页的关键词、网页的摘要以及网页的 SP 信息。

[0019] 根据本发明系统的再一实施例，分配服务器包括：地址提取模块，用于从网页信息中提取网页的 URL 地址；策略调用模块，用于调用分配优先级策略；抓取分配模块，分别与地址提取模块和策略调用模块相连，用于将提取的符合分配优先级策略的 URL 地址发送给爬虫以抓取网页的内容。

[0020] 根据本发明系统的再一实施例，该系统还包括：索引服务器，与分配服务器相连，用于对上报的网页信息和获取的网页的内容进行分类并建立关系索引。

[0021] 本发明的网页内容获取系统，在网页满足上报触发条件时上报网页信息，爬虫根据网页信息到指定网页中抓取网页的内容。该方法节约了爬虫的工作量，缓解了目标网站的压力，并且增加了获取实时信息的能力，为实时搜索提供了有利的条件。

附图说明

[0022] 此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分。在附图中：

[0023] 图 1 是本发明方法的第一实施例的流程示意图。

[0024] 图 2 是本发明方法的第二实施例的流程示意图。

[0025] 图 3 是本发明方法的第四实施例的流程示意图。

[0026] 图 4 是本发明系统的第一实施例的结构示意图。

[0027] 图 5 是本发明系统的第三实施例的结构示意图。

[0028] 图 6 是本发明系统的第四实施例的结构示意图。

[0029] 图 7 是本发明系统的第五实施例的结构示意图。

具体实施方式

[0030] 下面参照附图对本发明进行更全面的描述，其中说明本发明的示例性实施例。本发明的示例性实施例及其说明用于解释本发明，但并不构成对本发明的不当限定。

[0031] 实现实时搜索的一个非常困难的问题是从大量的网络信息中查找用户更新的数据并获取。为了能够第一时间掌握网页中内容的更新，本发明基于上报触发条件的网页内容获取方法在网页满足上报触发条件时，主动地上报该网页的信息，以便搜索平台对网页内容的获取。例如，当博客的博主对其一篇文章更新了当日的内容时，该网页会主动上报其 URL 地址、关键词、摘要和 SP 等信息给搜索平台的服务器，服务器再安排爬虫去该网页获取更新的内容信息。

[0032] 图 1 是本发明方法的第一实施例的流程示意图。

[0033] 如图 1 所示，该实施例包括以下步骤：

[0034] S102，判断网站中的网页是否满足上报触发条件，例如，可以在网页中添加计数器等功能，当某种计数满足条件时触发上报，例如，可添加浏览计数器、回复计数器和时钟功能等；

[0035] S104，在网页满足上报触发条件时，网站上报该网页的信息，例如，包括网页提取的自身 URL 地址以及预先设定的关键词、摘要和 SP 信息等；

[0036] S106，分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容。

[0037] 该实施例在网页满足上报触发条件时上报网页信息，爬虫根据网页信息到指定网页中抓取网页的内容。该方法节约了爬虫的工作量，缓解了目标网站的压力，并且增加了获取实时信息的能力，为实时搜索提供了有利的条件。

[0038] 图 2 是本发明方法的第二实施例的流程示意图。

[0039] 如图 2 所示，该实施例包括以下步骤：

[0040] S202，根据网站类型的不同，为每类网站设置不同的上报触发条件；

[0041] 例如，对于论坛博客类网站，其内容更新频率快、内容多、浏览量大、有大量的回复内容、并且是重要信息监管的重要站点，因此需要加强对论坛博客类网站信息的上报频度和内容，因而其上报触发条件可以是：发表新文章和更新文章内容触发上报，或根据回帖数量和浏览数量触发上报，或每天定时上报；

[0042] 对于门户类网站，其内容更新较快、浏览量大、但是回复量与论坛博客类相比较少，因此，其上报触发条件可以是：发表新文章触发上报，或根据回帖数量和浏览数量触发上报，或每天定时上报；

[0043] 对于资源信息类网页，其一般都有自己的更新频率，并且内容更新较少，格式也统一，因此可以遵循其本身的更新频率，其上报触发条件可以是：发表新文章触发上报，或根据回帖数量和浏览数量触发上报，或每天定时上报；

[0044] S204，判断网站中的网页是否满足上报触发条件；

[0045] S206，在网页满足上报触发条件时，网站上报网页信息；

[0046] S208，分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容。

[0047] 该实施例能够针对不同类型的网站分别设置不同的上报触发条件，在满足需求的同时，不仅大大缓解了网络带宽的压力，而且显著提高了工作效率，进而提高了对网页内容的实时获取能力。

[0048] 在本发明方法的第三实施例中，分配服务器根据上报的网页信息安排爬虫到网页

中抓取网页的内容的步骤包括：

[0049] 分配服务器从网页信息中提取网页的 URL 地址；

[0050] 调用分配优先级策略；

[0051] 将提取的符合分配优先级策略的 URL 地址发送给爬虫；

[0052] 爬虫从接收到的 URL 地址中抓取网页的内容。

[0053] 具体地，当大量信息超过爬虫获取能力时，为了提高爬虫获取信息的实时性和效率，可以设置下述分配优先级策略：

[0054] (1) 基于网站权重

[0055] 根据网站流量和重要程度可以设置不同的权重，例如，可以将门户类网站、博客类网站等流量大而且重要的网站的权重设置为高，其他网站随着流量和重要程度其权重逐渐降低。

[0056] (2) 基于时间权重

[0057] 上报信息随着等待时间的增加权重不断降低。

[0058] (3) 排序

[0059] 按照网站权重与时间权重相乘的结果降序排列，依次将地址分配给爬虫。

[0060] 该实施例根据分配优先级策略对爬虫抓取网页内容进行了优化，提高了爬虫的工作效率，对降低系统带宽的负担起到了至关重要的作用，在很大程度上也提高了网页内容获取的实时性。

[0061] 图 3 是本发明方法的第四实施例的流程示意图。

[0062] 如图 3 所示，该实施例包括以下步骤：

[0063] S302，判断网站中的网页是否满足上报触发条件；

[0064] S304，在网页满足上报触发条件时，网站上报网页信息；

[0065] S306，分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容；

[0066] S308，将上报的网页信息和获取的网页的内容发送给索引服务器，由索引服务器进行分类整理，然后建立关系索引再存入数据库中，其中，Flag 是一个标记位，0 代表未处理，1 代表已处理，2 代表正在处理，3 代表已删除。

[0067] 在上述实施例中，上报触发条件包括发表新文章触发上报、文章内容更新触发上报、依据回帖数量触发上报、依据浏览量触发上报以及定时触发上报中的至少一种。

[0068] 网页信息包括网页的 URL 地址、网页的关键词、网页的摘要以及网页的 SP 信息。

[0069] 在本发明方法的第五实施例中，以网页内容更新为例说明如何抓取网页内容：

[0070] 如果网页有内容更新，则判断是否满足设定的上报触发条件，如果满足，则获取自身的 URL 地址，并读取预设的关键词、摘要、SP 信息等，将这些网页信息上传至分配服务器；

[0071] 分配服务器获得上报的网页信息，分析并提取上报信息中的 URL 地址，再从数据库中调用分配优先级策略，匹配分配优先级策略，将符合策略的 URL 地址发送给状态空闲的爬虫；

[0072] 状态空闲的爬虫获取分配服务器发送的 URL 地址，从分配到的 URL 地址中抓取该链接的网页内容，不进行其他链接的抓取，再将抓取到的网页内容发送给索引服务器。

[0073] 分配服务器也将 URL 地址、关键词、摘要、SP 信息等从网页获取的信息发送给索引服务器。

[0074] 图 4 是本发明系统的第一实施例的结构示意图。

[0075] 如图 4 所示，该实施例的系统包括 ：

[0076] 内容触发服务器 11，用于判断网站中的网页是否满足上报触发条件，如果满足上报触发条件，则上报网页信息 ；

[0077] 分配服务器 12，与内容触发服务器 11 相连，用于根据上报的网页信息安排爬虫到网页中抓取网页的内容。

[0078] 该实施例在网页满足上报触发条件时上报网页信息，爬虫根据网页信息到指定网页中抓取网页的内容。该方法节约了爬虫的工作量，缓解了目标网站的压力，并且增加了获取实时信息的能力，为实时搜索提供了有利的条件。

[0079] 在本发明系统的第二实施例中，与图 4 中的实施例相比，该实施例的系统中的内容触发服务器还用于 ：根据网站类型的不同，为每类网站设置不同的上报触发条件。

[0080] 例如，对于论坛博客类网站，其内容更新频率快、内容多、浏览量大、有大量的回复内容、并且是重要信息监管的重要站点，因此需要加强对论坛博客类网站信息的上报频度和内容，因而其上报触发条件可以是 ：发表新文章和更新文章内容触发上报，或根据回帖数量和浏览数量触发上报，或每天定时上报。

[0081] 该实施例能够针对不同类型的网站分别设置不同的上报触发条件，在满足需求的同时，不仅大大缓解了网络带宽的压力，而且显著提高了工作效率，进而提高了网页内容获取的实时性。

[0082] 图 5 是本发明系统的第三实施例的结构示意图。

[0083] 如图 5 所示，与图 4 中的实施例相比，该实施例的系统中的分配服务器包括 21 ：

[0084] 地址提取模块 211，用于从网页信息中提取网页的 URL 地址 ；

[0085] 策略调用模块 212，用于调用分配优先级策略 ；

[0086] 抓取分配模块 213，分别与地址提取模块 211 和策略调用模块 212 相连，用于将提取的符合分配优先级策略的 URL 地址发送给爬虫以抓取网页的内容。

[0087] 该实施例根据分配优先级策略对爬虫抓取网页内容进行了优化，提高了爬虫的工作效率，对降低系统带宽的负担起到了至关重要的作用。

[0088] 图 6 是本发明系统的第四实施例的结构示意图。

[0089] 如图 6 所示，与图 4 中的实施例相比，该实施例的系统还包括 ：

[0090] 索引服务器 31，与分配服务器 12 相连，用于对上报的网页信息和获取的网页的内容进行分类并建立关系索引。

[0091] 在上述实施例中，上报触发条件包括发表新文章触发上报、文章内容更新触发上报、依据回帖数量触发上报、依据浏览量触发上报以及定时触发上报中的至少一种。

[0092] 网页信息包括网页的 URL 地址、网页的关键词、网页的摘要以及网页的 SP 信息。

[0093] 图 7 是本发明系统的第五实施例的结构示意图。

[0094] 如图 7 所示，该实施例的系统包括 ：内容触发服务器 11、分配服务器 12、内容触发式爬虫 13。这三者的主要目的是将满足上报触发条件的网页的信息交由分配服务器，再由分配服务器将该网页的 URL 地址提供给内容触发式爬虫，爬虫将网页中的内容抓取至搜索平台。

[0095] 其中，内容触发服务器 11 用于 ：在网页满足上报触发条件时，获取网页的 URL 地

址，读取预设的关键词、摘要、SP 信息等网页信息，再将这些网页信息上传至分配服务器。

[0096] 分配服务器 12 用于：接收上报的网页信息，分析上报的信息，提取其中的 URL 地址，再从数据库 15 中调用分配优先级策略，匹配分配优先级策略，将符合策略的 URL 地址发送给状态空闲的爬虫，再将 URL 地址、关键词、摘要、SP 信息等从网页获取的信息发送给索引服务器 14。

[0097] 内容触发式爬虫 13 用于：发送空闲状态至分配服务器，获取分配服务器发送的 URL 地址，返回状态忙值，再从分配到的 URL 地址中抓取该链接的网页信息，不进行其他链接的抓取，最后将抓取到的网页发送给索引服务器，并返回空闲状态至分配服务器。

[0098] 索引服务器 14 对网页信息和网页内容进行分类整理，并建立关系索引，再将关系索引存储到数据库 15 中。

[0099] 另外，内容触发服务器还在网页中添加计数器等功能，当某种计数满足条件时触发上报。例如，可以添加浏览计数器、回复计数器等，还可添加时钟功能。

[0100] 上报的网页信息包括：该网页所在 URL 地址、关键词（不超过 10 个）、摘要、SP 信息等。

[0101] 此外，上报触发方式可以包括：发表新文章触发上报、文章内容有更新触发上报、依据回帖数量触发上报、依据浏览量触发上报、设定时间触发上报中的至少一种。

[0102] 其中，可以针对不同网站类型的特点设置不同的上报触发条件以提高效率。

[0103] 本发明的描述是为了示例和描述起见而给出的，而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显而易见的。选择和描述实施例是为了更好说明本发明的原理和实际应用，并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

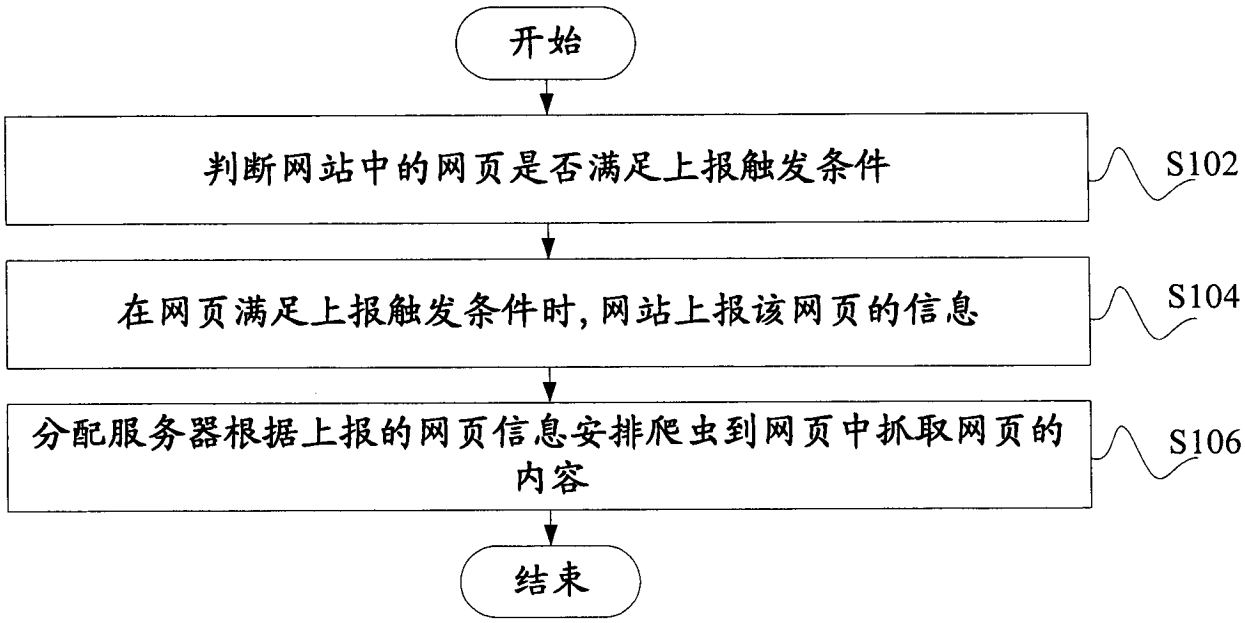

开始
判断网站中的网页是否满足上报触发条件
S102
在网页满足上报触发条件时，网站上报该网页的信息
S104
分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容
S106
结束

图 1

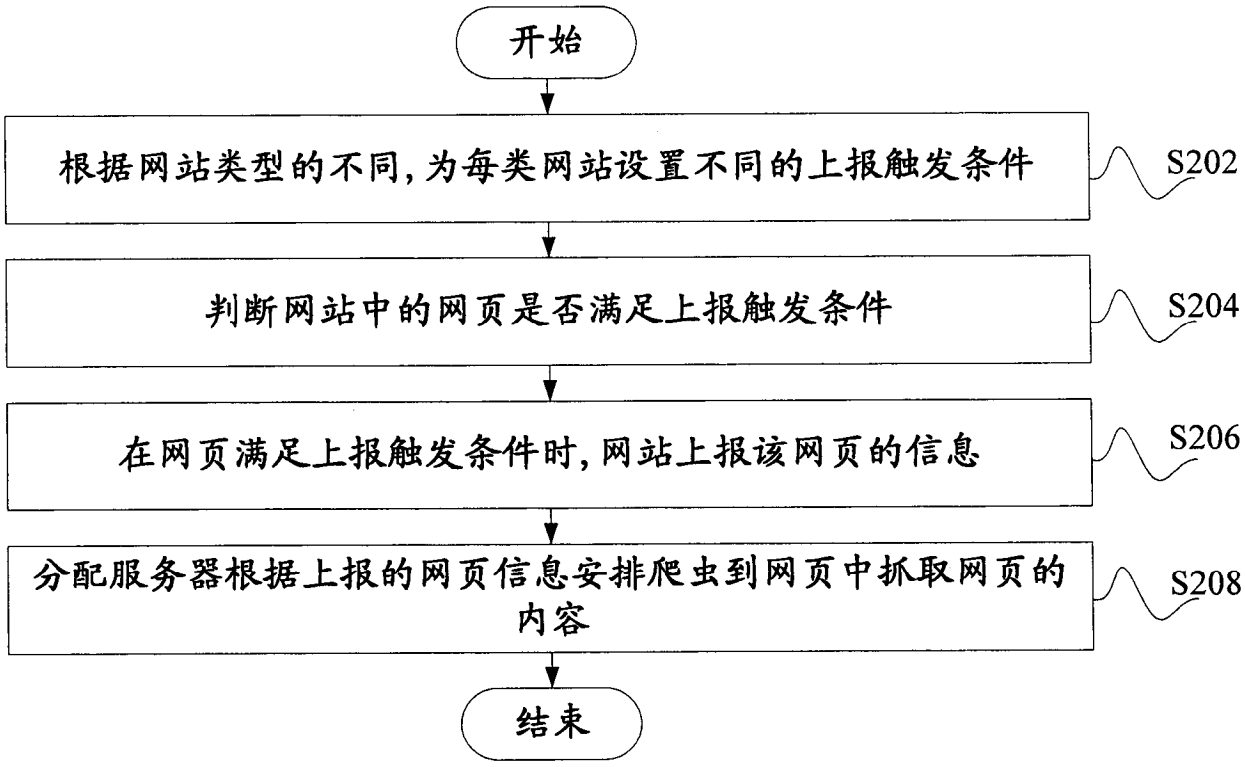

开始
根据网站类型的不同，为每类网站设置不同的上报触发条件
S202
判断网站中的网页是否满足上报触发条件
S204
在网页满足上报触发条件时，网站上报该网页的信息
S206
分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容
S208
结束

图 2

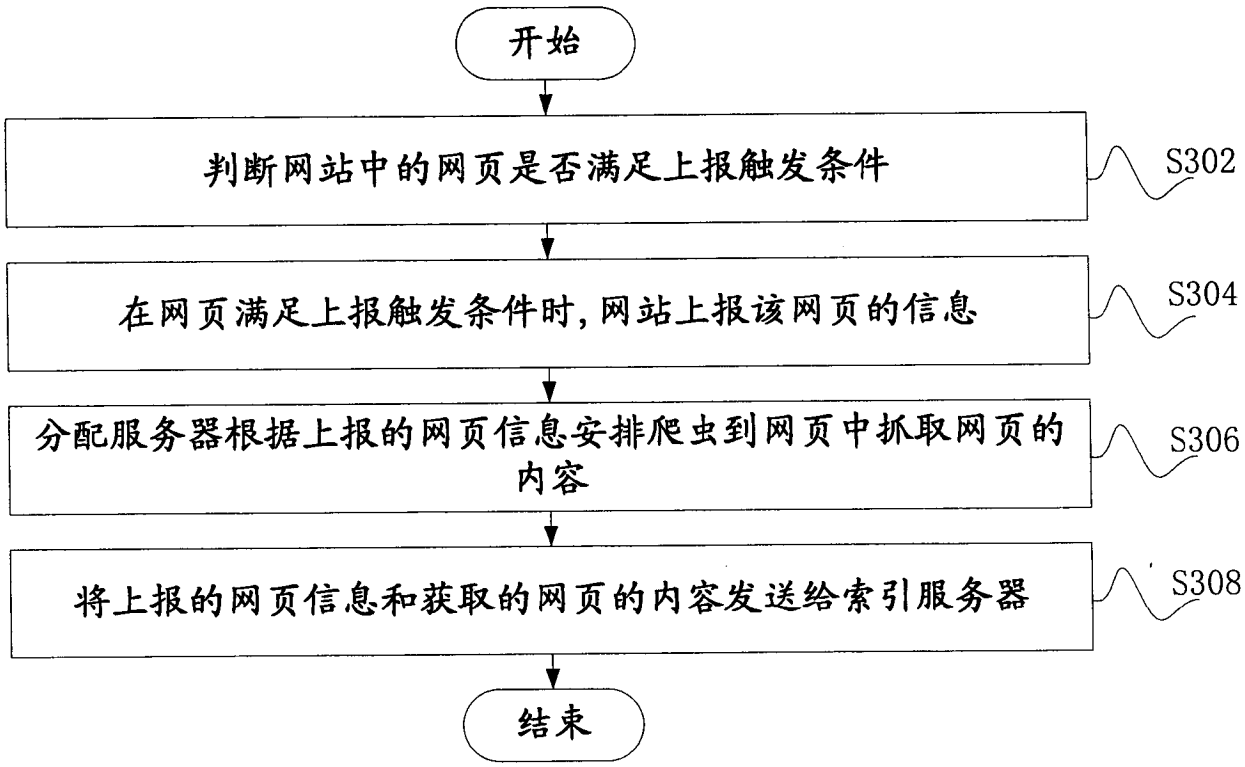

开始
判断网站中的网页是否满足上报触发条件
S302
在网页满足上报触发条件时，网站上报该网页的信息
S304
分配服务器根据上报的网页信息安排爬虫到网页中抓取网页的内容
S306
将上报的网页信息和获取的网页的内容发送给索引服务器
S308
结束

图 3

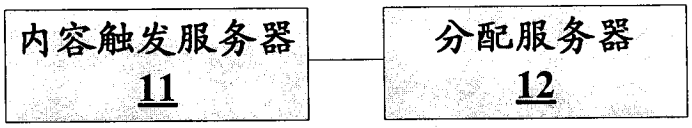

内容触发服务器
11
分配服务器
12

图 4

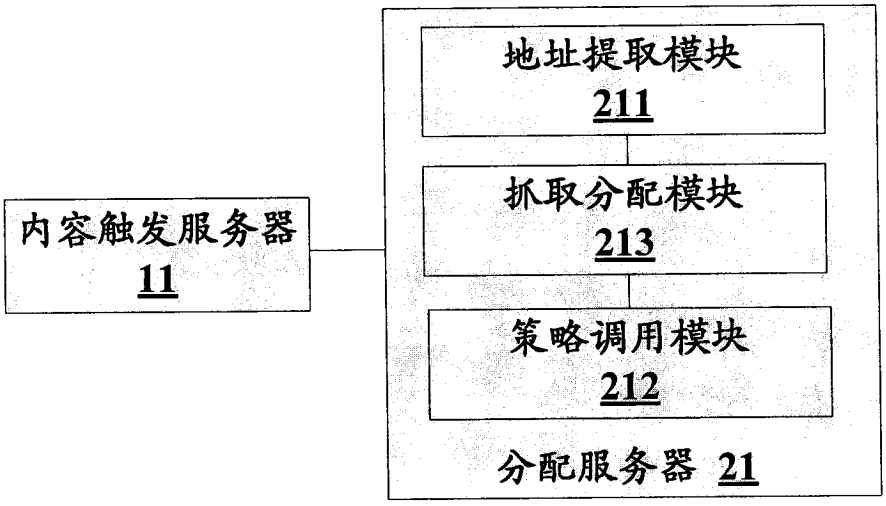

地址提取模块
211
抓取分配模块
213
策略调用模块
212
分配服务器 21
内容触发服务器
11

图 5

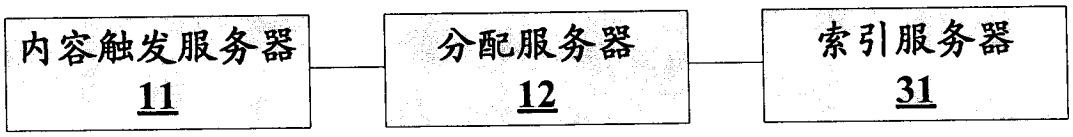

内容触发服务器
11
分配服务器
12
索引服务器
31

图 6

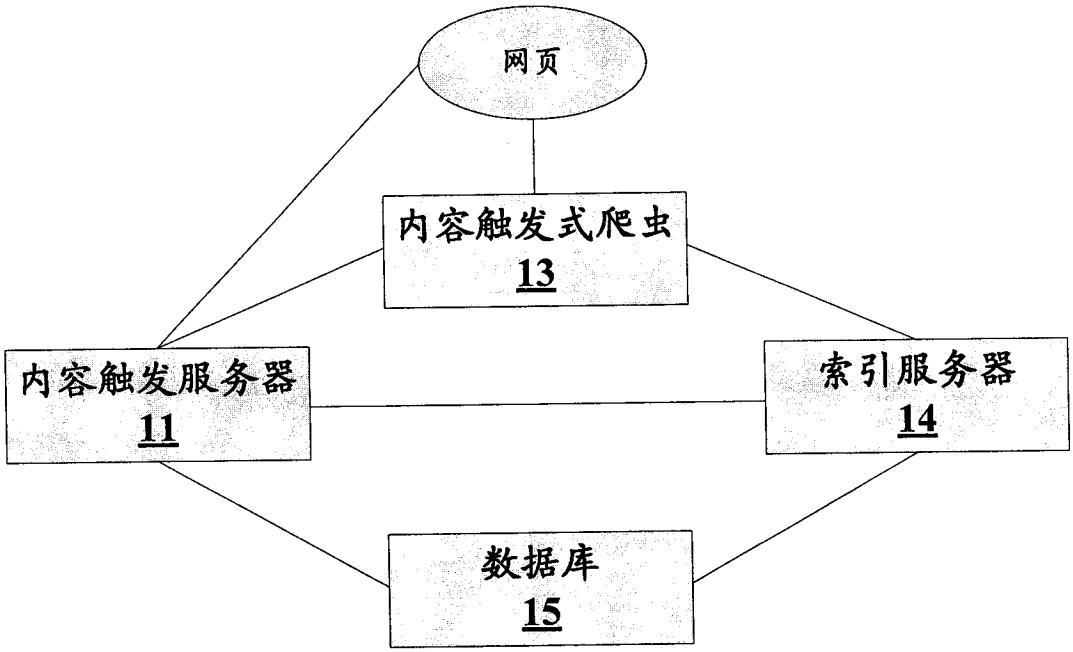

网页
内容触发式爬虫
13
内容触发服务器
11
索引服务器
14
数据库
15

图 7