(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 102469132 A
(43)申请公布日 2012.05.23

(21)申请号 201010546334.0
(22)申请日 2010.11.15
(71)申请人 北大方正集团有限公司
地址 100871 北京市海淀区成府路 298 号方正大厦
申请人 北京大学
北京北大方正电子有限公司
(72)发明人 李湘军 于晓明 杨建武 吴新丽
(74)专利代理机构 北京天悦专利代理事务所(普通合伙) 11311
代理人 田明 任晓航
(51)Int.Cl.
H04L 29/08(2006.01)
H04L 29/12(2006.01)
G06F 17/30(2006.01)

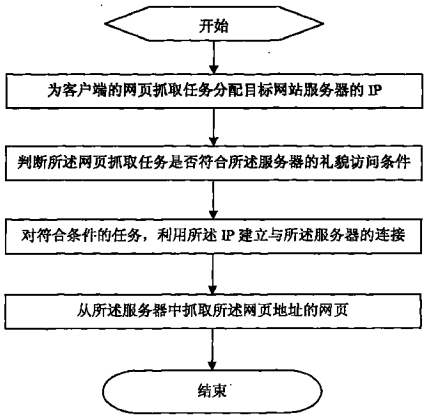
权利要求书 2 页 说明书 4 页 附图 2 页

(54)发明名称

从网站中多个不同 IP 的服务器抓取网页的方法及系统

(57)摘要

本发明公开了一种从网站中多个不同 IP 的服务器抓取网页的方法及系统。本发明首先为客户端的网页抓取任务分配目标网站服务器的 IP，所述网页抓取任务包括待抓取网页的网页地址；然后判断所述网页抓取任务是否符合所述服务器的礼貌访问条件；如果符合，则利用所述 IP 建立与所述服务器的连接，从所述服务器中抓取所述网页地址的网页。本发明的访问策略基于 IP 级，更便于控制采集工作线程对网站进行礼貌地访问；通过缓存 DNS，同时使用多个 IP 并优先分配速度最快 IP 的方式，极大地提高了网页抓取的效率；而且当目标网站有个别服务器不能访问时能够及时切换到其他 IP 的服务器，提高了容错能力。



CN 102469132 A

1. 一种从网站中多个不同 IP 的服务器抓取网页的方法，包括以下步骤：

(1) 为客户端的网页抓取任务分配目标网站服务器的 IP，所述网页抓取任务包括待抓取网页的网页地址；

(2) 判断所述网页抓取任务是否符合所述服务器的礼貌访问条件；如果符合，则利用所述 IP 建立与所述服务器的连接，从所述服务器中抓取所述网页地址的网页。

2. 如权利要求 1 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：步骤 (1) 中所述网页抓取任务可随时加载到抓取任务队列中；定时刷新抓取任务队列；如果抓取任务队列不空，则遍历抓取任务队列，获取符合所述礼貌访问条件的网页抓取任务。

3. 如权利要求 2 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：所述礼貌访问条件包括如下两个方面：①使用该 IP 的并发数不超过设定的并发数限制；②向该 IP 发送请求的时间间隔不小于设定的时间间隔限制。

4. 如权利要求 3 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：如果当前网页抓取任务符合所述礼貌访问条件，则从抓取任务队列中取下该网页抓取任务，并累加该网页抓取任务所使用 IP 的并发数；如果当前网页抓取任务不符合所述礼貌访问条件，则继续判断抓取任务队列中下一个网页抓取任务。

5. 如权利要求 1 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：步骤 (1) 中所述为网页抓取任务分配目标网站服务器的 IP 的过程如下：

判断客户端缓存中与所述网页地址对应的主机名是否有 IP 列表；如果没有 IP 列表，则预分配使用第一个 IP；如果有 IP 列表且已知 IP 列表中所有 IP 的抓取速度，则分配一个可用的且抓取速度最快的 IP，否则轮询分配一个 IP。

6. 如权利要求 5 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：如果网页抓取任务没有与该网页地址对应的主机名的 IP 列表，则对该主机名进行 DNS 解析，获取 IP 列表，将预分配的第一个 IP 转换为 IP 列表中实际的第一个 IP；然后将所述主机名和 IP 列表进行缓存。

7. 如权利要求 2 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：步骤 (2) 中，在利用所述 IP 建立与所述服务器的连接过程中，如果连接成功，则缓存该连接，下次使用该连接时，直接从缓存中获取该连接；如果连接失败，则对该 IP 进行标记，并将该网页抓取任务加载到失败任务队列。

8. 如权利要求 7 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：缓存该连接后，记录每次使用该连接的时间，当使用该连接的时间超过了设定的有效期限时，则删除该连接。

9. 如权利要求 7 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：如果连接失败的次数超过设定的阈值 M 时，标记该 IP 不可用，下次分配 IP 时不再分配该 IP。

10. 如权利要求 9 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：当不可用的 IP 个数超过了 IP 列表中 IP 总个数的设定比例 N 时，重新进行 DNS 解析，获取新的 IP 列表。

11. 如权利要求 2 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：步骤 (2) 中，在建立与所述服务器的连接后，访问所述网页地址的网页；如果访问成功，则

抓取网页内容，并标记该连接可以重用；如果访问失败，则关闭该连接，并将该网页抓取任务加载到失败任务队列。

12. 如权利要求 7 或 11 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：定时将所述失败任务队列中的任务加载到抓取任务队列中。

13. 如权利要求 11 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：步骤 (2) 中，在抓取网页内容后，记录该网页抓取任务所使用 IP 的本次抓取速度，并统计所述 IP 的综合抓取速度。

14. 如权利要求 13 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：采用如下公式统计所述 IP 的综合抓取速度：

$$\begin{cases} S_0 = 0 \\ S_n = \alpha S_{n-1} + \beta R_n, n \geq 1, \alpha + \beta = 1 \end{cases}$$

其中，S_n 表示第 n 次所述 IP 的抓取速度，S_{n-1} 表示第 n-1 次所述 IP 的抓取速度，α 表示历史抓取速度的权重，R_n 表示第 n 次所述 IP 的实际抓取速度，β 表示当前抓取速度的权重。

15. 如权利要求 2 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于：步骤 (2) 中，在从所述服务器中抓取所述网页地址的网页后，分析抓取的网页中的网页地址链接；先在抓取网页中的网页地址链接中进行排重，然后再在整个抓取任务队列中的网页地址链接中进行排重；排重后加载到抓取任务队列中。

16. 如权利要求 15 所述的从网站中多个不同 IP 的服务器抓取网页的方法，其特征在于，所述排重的方法为：先将网页地址字符串转换为 MD5 值，然后通过比较每个网页地址的 MD5 值进行排重。

17. 一种从网站中多个不同 IP 的服务器抓取网页的系统，包括为客户端的网页抓取任务分配目标网站服务器 IP 的分配装置 (11)，所述网页抓取任务包括待抓取网页的网页地址；

用于判断所述网页抓取任务是否符合所述服务器的礼貌访问条件的判断装置 (12)；

用于利用所述 IP 建立与所述服务器的连接，从所述服务器中抓取所述网页地址的网页的抓取装置 (13)。

从网站中多个不同 IP 的服务器抓取网页的方法及系统

技术领域

[0001] 本发明涉及一种从网站中抓取网页的方法及系统，尤其是涉及一种从网站中多个不同 IP 的服务器抓取网页的方法及系统。

背景技术

[0002] 随着互联网的飞速发展，互联网上的信息规模越来越大，网站访问量也越来越大。大多数信息规模较大或访问量较大的网站，为了满足目前的互联网访问需求，都提供了多台不同 IP(Internet Protocol，网络之间互连的协议）的服务器，通过智能 DNS(Domain Name System，域名系统）服务器，按照负载均衡的策略返回不同顺序的服务器 IP 列表，客户端会使用第一个服务器进行访问，从而将用户的访问请求分散到不同的服务器上。为了防止服务器压力过大或者被恶意攻击，这些网站尤其是论坛和博客，会对并发过大或者频率过快的访问采取临时性地拒绝服务或永久性地封杀对方 IP 的措施。对于一个客户端而言，如果该客户端发送的多个访问请求被分配到了网站的同一台服务器上，则可能会因为该服务器的限制而被拒绝服务，甚至被封杀 IP。

[0003] 现有的爬虫系统一般都按照网站来控制访问策略，由于受到网站访问并发数的限制，所以采集效率较低，如果增加抓取网页的工作线程数量又容易触发网站的限制访问条件，造成抓取失败或被封杀 IP。

发明内容

[0004] 针对现有技术中存在的缺陷，本发明要解决的技术问题是提供一种从网站中多个不同 IP 的服务器抓取网页的方法与系统，该方法及系统能够在礼貌访问网站的前提下，成倍地提高网页抓取的效率。

[0005] 为解决上述技术问题，本发明采用的技术方案如下：

[0006] 一种从网站中多个不同 IP 的服务器抓取网页的方法，包括以下步骤：

[0007] (1) 为客户端的网页抓取任务分配目标网站服务器的 IP，所述网页抓取任务包括待抓取网页的网页地址；

[0008] (2) 判断所述网页抓取任务是否符合所述服务器的礼貌访问条件；如果符合，则利用所述 IP 建立与所述服务器的连接，从所述服务器中抓取所述网页地址的网页。

[0009] 一种从网站中多个不同 IP 的服务器抓取网页的系统，包括为客户端的网页抓取任务分配目标网站服务器 IP 的分配装置，所述网页抓取任务包括待抓取网页的网页地址；

[0010] 用于判断所述网页抓取任务是否符合所述服务器的礼貌访问条件的判断装置；

[0011] 用于利用所述 IP 建立与所述服务器的连接，从所述服务器中抓取所述网页地址的网页的抓取装置。

[0012] 本发明所述的方法及系统，访问策略基于 IP 级，更便于控制采集工作线程对网站进行礼貌地访问；通过缓存 DNS，同时使用多个 IP 并优先分配速度最快 IP 的方式，极大地提高了网页抓取的效率；当目标网站有个别服务器不能访问时能够及时切换到其他 IP 的

服务器，提高了容错能力。

附图说明

[0013] 图 1 是本发明所述从网站中多个不同 IP 的服务器抓取网页的系统结构框图；

[0014] 图 2 是本发明所述从网站中多个不同 IP 的服务器抓取网页的方法流程图；

[0015] 图 3 是本发明所述方法一具体实施方式的流程图。

具体实施方式

[0016] 下面结合具体实施方式和附图对本发明进行详细描述。

[0017] 图 1 示出了本发明所述从网站中多个不同 IP 的服务器抓取网页的系统结构。如图 1 所示，该系统包括分配装置 11，与分配装置 11 连接的判断装置 12，与判断装置 12 连接的抓取装置 13。

[0018] 分配装置 11 用于为客户端的网页抓取任务分配目标网站服务器 IP。所述网页抓取任务包括待抓取网页的 URL（网页地址）；所述目标网站是指待抓取网页所在的网站。

[0019] 判断装置 12 用于判断网页抓取任务是否符合服务器的礼貌访问条件。所述礼貌访问条件包括如下两个方面：①使用该 IP 的并发数不超过设定的并发数限制；②向该 IP 发送请求的时间间隔不小于设定的时间间隔限制。

[0020] 抓取装置 13 用于利用分配的 IP 建立与该 IP 的服务器的连接，从该 IP 的服务器中抓取所述 URL 的网页。

[0021] 图 2 示出了采用图 1 所示系统从网站中多个不同 IP 的服务器抓取网页的方法流程。该方法首先为客户端的网页抓取任务分配目标网站服务器的 IP；然后判断该网页抓取任务是否符合被分配 IP 的服务器的礼貌访问条件；如果符合，则利用所述 IP 建立与所述服务器的连接，从所述服务器中抓取所述网页地址的网页。

[0022] 图 3 示出了本发明所述方法一具体实施方式的流程。客户端可随时将网页抓取任务加载到抓取任务队列中，所述网页抓取任务包括待抓取网页的 URL。如图 3 所示，对抓取任务队列中的网页抓取任务进行如下操作：

[0023] (1) 定时刷新抓取任务队列，如果抓取任务队列为空，则重复该步骤。

[0024] (2) 遍历抓取任务队列。为当前网页抓取任务分配网站服务器的 IP 地址，并判断当前网页抓取任务是否符合所述礼貌访问条件。只有满足了礼貌访问条件，抓取网页才不会被目标网站拒绝。如果符合礼貌访问条件，则从抓取任务队列中取下该网页抓取任务，并累加该网页抓取任务所使用 IP 的并发数，作为下次判断礼貌访问条件的依据；如果不符合，则继续判断抓取任务队列中下一个网页抓取任务，直到本次遍历结束。

[0025] 所述分配 IP 地址的过程如下：

[0026] 判断客户端缓存中与该网页抓取任务中的 URL 对应的主机名是否有 IP 列表。如果没有 IP 列表，则预分配使用第一个 IP，即标记该网页抓取任务使用 IP 列表中的第一个 IP，等到进行了 DNS 解析后再转换为实际的第一个 IP 地址。此处之所以不立即进行 DNS 解析是为了避免影响获取网页抓取任务的效率。如果有 IP 列表且已知 IP 列表中所有 IP 的抓取速度，则分配一个可用的且抓取速度最快的 IP。如果 IP 的抓取速度未知，则轮询分配一个 IP，即按照顺序分配。例如，将 IP1 分配给任务 a，将 IP2 分配给任务 b，将 IP3 分配给

任务 c ;当所有 IP 均被分配一遍后,再从 IP1 开始分配。

[0027] 如果客户端缓存中与取下的网页抓取任务中的 URL 对应的主机名没有 IP 列表,则对目标网站主机名进行 DNS 解析,获取 IP 列表,将预分配使用的第一个 IP 转换为 IP 列表实际的第一个 IP 地址。对主机名和与该主机名对应的 IP 列表进行缓存,这样在以后分配 IP 时,只需根据主机名从缓存中直接查找 IP 即可,不必每个网页抓取任务都重新进行 DNS 解析,从而减小了解析的代价,减轻了对 DNS 服务器的压力。

[0028] (3) 对取下的网页抓取任务,用所分配的 IP 与目标网站该 IP 的服务器建立 Socket 连接。记录每次发送连接请求的时间,作为下次判断礼貌访问条件的依据。

[0029] 如果连接成功,则缓存该连接以便重复利用,下次使用该连接时,直接从缓存中获取该连接。如果连接失败,则标记该 IP 建立连接失败,关闭该连接,并将该网页抓取任务加载到失败任务队列。定时将失败任务队列中的任务加载到抓取任务队列中。

[0030] 如果与某 IP 建立连接的失败次数超过设定的阈值 M 时,标记该 IP 不可用,分配 IP 时不再分配该 IP。本实施方式中,M 取值为 20。当不可用的 IP 个数超过了 IP 列表中 IP 总数的设定比例 N 时,重新进行 DNS 解析,获取新的 IP 列表。本实施方式中,N 取值为 50%。

[0031] 缓存客户端与目标网站建立的 Socket 连接,并在短时间内重复利用这个连接,能够减少建立连接的代价,提高运行的效率。为了防止重用连接超过设定的有效期限造成访问网页失败,还需要记录每次使用该连接的时间,如果某次使用该连接的时间超过了有效期限,则删除缓存中的该连接,下次使用时再重新建立新的 Socket 连接。

[0032] 网站服务器可能由于某些原因如网络不稳定,服务器宕机,更换 IP 等造成当前 IP 不能使用,此时建立 Socket 连接就会失败。IP 不能使用,可能是暂时的,也可能是长期的,因此本实施方式中记录连接失败的次数,只有当连接失败次数超过设定的阈值 M 时,才标记该 IP 为不可用,分配 IP 时不再分配该 IP。

[0033] (4) 根据建立的连接通过 HTTP 协议访问网页。

[0034] 如果访问成功,则抓取网页内容,并标记该连接可以重用。记录并统计该 IP 的抓取速度,为下一次分配 IP 时提供依据。由于 IP 的访问速度是不断变化的,记录的速度值应该能够反映历史和当前的抓取速度,这样既能防止当前偶尔的网络不好导致不能反映历史速度,又能防止过去网络一直不好导致不能反映当前的速度。在抓取网页内容后,记录该网页抓取任务所使用 IP 的本次抓取速度,并统计该 IP 的综合抓取速度。可以采用如下公式统计 IP 的综合抓取速度 :

[0035] $$\begin{cases} S_0 = 0 \\ S_n = \alpha S_{n-1} + \beta R_n, n \geq 1, \alpha + \beta = 1 \end{cases}$$

[0036] 其中,S_n 表示第 n 次该 IP 的记录速度,S_{n-1} 表示第 n-1 次该 IP 的记录速度,α 表示历史速度的权重,R_n 表示第 n 次该 IP 的实际抓取速度,β 表示当前速度的权重。α 和 β 值之和应为 1,这两个值直接影响到统计的 IP 访问速度的可靠性。本实施方式中,α 和 β 的取值均为 0.5。

[0037] 如果访问网页失败,则关闭该连接,将该任务加入到失败任务队列,定时将失败任务队列中的任务加入到抓取任务队列。访问完毕后,还需要递减使用该 IP 的并发数,以使得可以继续分配该 IP。

[0038] (5) 分析被抓取的网页中的 URL 链接,排重后加载到抓取任务队列中。

[0039] 分析被抓取的网页中的URL链接的方法可采用现有方法，如中国专利申请公开的“一种基于网页抽取的搜索系统及搜索方法”（公开日：2008.06.04；公开号：CN101192234），此处不再详细说明。

[0040] 分析出被抓取的网页中的URL链接后，先在抓取的网页中的网页地址链接中进行排重，然后再在整个抓取任务队列中的网页地址链接中进行排重，避免重复抓取。

[0041] 本实施方式中，排重的方法如下：先将URL字符串转换为MD5值，然后通过比较每个URL的MD5值进行排重。在网页中的网页地址链接中排重时，如果有相同的，则只取一个。在整个网页抓取任务中的网页地址链接中进行排重时，如有相同的，则说明任务中存在该URL链接，否则说明不存在，可以加载到抓取任务队列中。

[0042] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其同等技术的范围之内，则本发明也意图包含这些改动和变型在内。

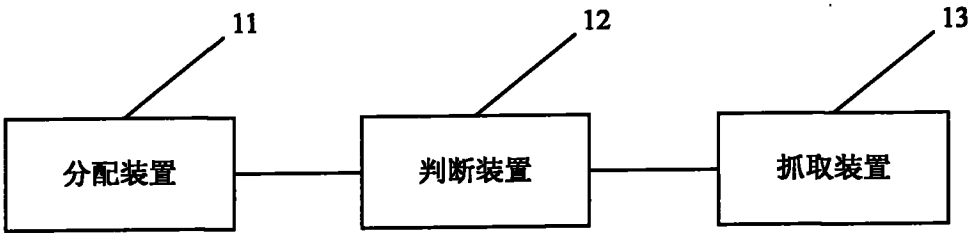

11
12
13
分配装置
判断装置
抓取装置

图 1

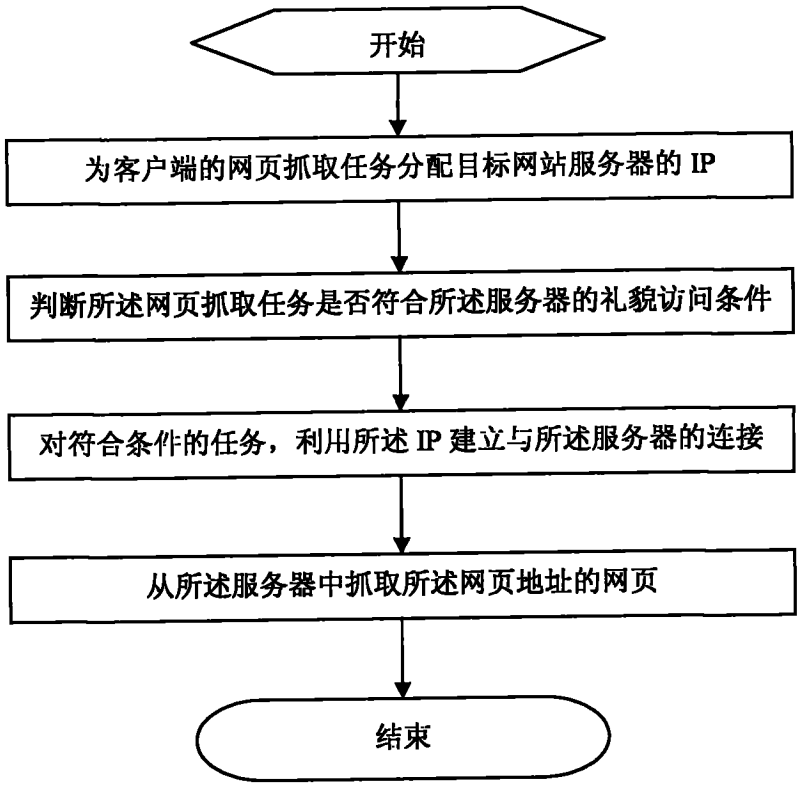

开始
为客户端的网页抓取任务分配目标网站服务器的 IP
判断所述网页抓取任务是否符合所述服务器的礼貌访问条件
对符合条件的任务，利用所述 IP 建立与所述服务器的连接
从所述服务器中抓取所述网页地址的网页
结束

图 2

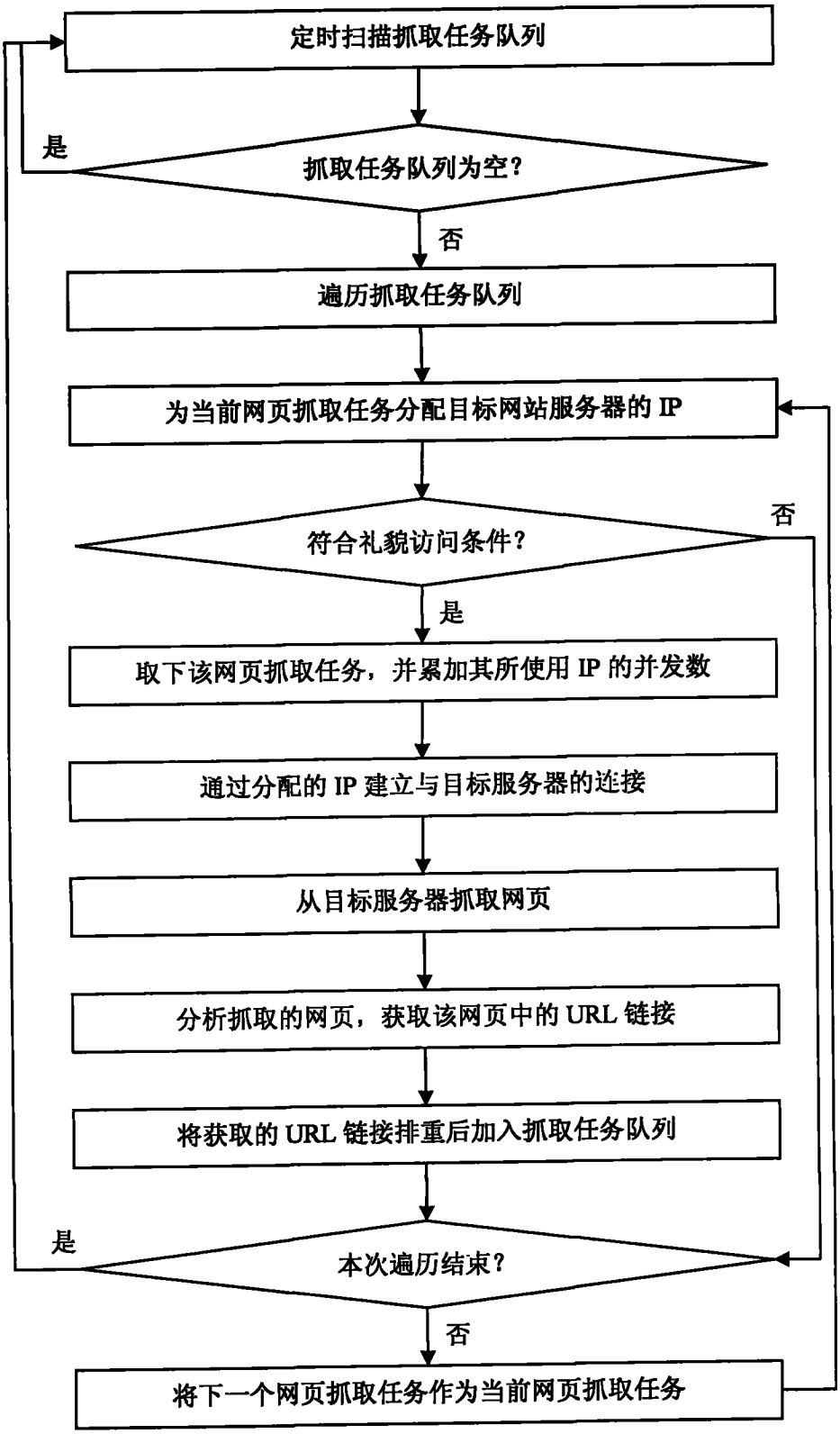

定时扫描抓取任务队列
是
抓取任务队列为空？
否
遍历抓取任务队列
为当前网页抓取任务分配目标网站服务器的 IP
否
符合礼貌访问条件？
是
取下该网页抓取任务，并累加其所使用 IP 的并发数
通过分配的 IP 建立与目标服务器的连接
从目标服务器抓取网页
分析抓取的网页，获取该网页中的 URL 链接
将获取的 URL 链接排重后加入抓取任务队列
是
本次遍历结束？
否
将下一个网页抓取任务作为当前网页抓取任务

图 3