[19] 中华人民共和国国家知识产权局

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

[12] 发明专利说明书

专利号 ZL 200610099277.X

[45] 授权公告日 2009 年 7 月 29 日

[11] 授权公告号 CN 100520778C

[22] 申请日 2006.7.25

[21] 申请号 200610099277.X

[73] 专利权人 腾讯科技(深圳)有限公司

地址 518044 广东省深圳市福田区振兴路赛格科技园 2 栋东 410 室

[72] 发明人 余祥鑫 杨 卫

[56] 参考文献

US 6707470B1 2004.3.16

CN 1851705A 2006.10.25

US 6665658B1 2003.12.16

JP2003 - 196144A 2003.7.11

审查员 沈乐平

[74] 专利代理机构 北京同达信恒知识产权代理有限公司

代理人 郭润湘

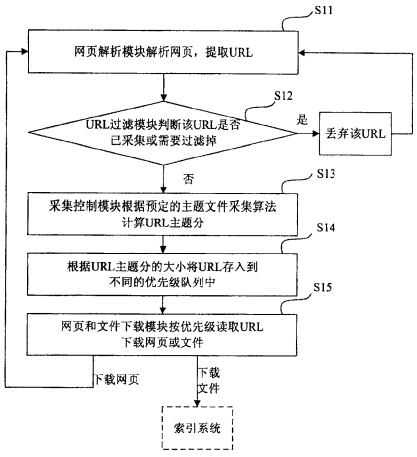
权利要求书 3 页 说明书 9 页 附图 3 页

[54] 发明名称

一种互联网主题文件搜索方法、爬虫系统和搜索引擎

[57] 摘要

本发明公开了一种互联网主题文件搜索方法，包括：解析下载的网页，提取网页中包含的统一资源定位符 URL；确定出各 URL 的对应优先级；按优先级从高到低的顺序采集各 URL，建立索引，搜索所需互联网主题文件。 本发明还公开了一种互联网主题文件的搜索引擎的爬虫系统和搜索引擎。 本发明提供的爬虫系统至少包括：URL 队列存储模块、网页和文件下载模块、网页解析模块和采集控制模块。 采用本发明可以提高互联网主题文件搜索效率。



1、一种互联网主题文件搜索方法，其特征在于，包括:

A、解析下载的网页，提取网页中包含的统一资源定位符 URL;

B、计算包含所述 URL 的已采集网页的网页主题分，累加所述网页主题分作为所述 URL 的 URL 主题分;

根据所述 URL 主题分的分值大小确定出所述 URL 的对应优先级;

C、按优先级从高到低的顺序采集各 URL，建立索引，搜索所需互联网主题文件。

2、如权利要求 1 所述的互联网主题文件搜索方法，其特征在于，还包括:

保存已采集的 URL 历史记录;

所述步骤 B 中，根据所述历史记录判断下载网页中包含的 URL 是否已采集，仅对未采集过的 URL 确定优先级。

3、如权利要求 2 所述的互联网主题文件搜索方法，其特征在于，还包括:

设置 URL 过滤条件，仅对未采集过的不符合所述过滤条件的 URL 确定优先级。

4、如权利要求 3 所述的互联网主题文件搜索方法，其特征在于，所述网页主题分具体计算公式为:

$F(p) = a \times numFileLink \times FactorLink + b*numKeyWord \times FactorWord;$

式中，F（p）为计算出的网页主题分;

numFileLink 为该网页含有的主题文件 URL 的个数;

FactorLink 为 URL 链接的积分因子;

numKeyWord 为该网页含有的主题关键词个数;

FactorWord 为主题关键词的积分因子;

a，b 为权重因子，且 a+b =1。

5、一种搜索引擎的爬虫系统，其特征在于包括: URL 队列存储模块、网

页和文件下载模块、网页解析模块和采集控制模块;

所述 URL 队列存储模块，按优先级顺序存储待采集的 URL;

所述网页和文件下载模块，按 URL 优先级从高到低的顺序下载网页或文件；将下载的网页发送到所述网页解析模块，将下载的文件发送到搜索引擎的索引系统处理;

所述网页解析模块，对网页进行解析，提取网页中包含的 URL 发送到所述采集控制模块;

所述采集控制模块，计算包含所述 URL 的已采集网页的网页主题分，累加所述网页主题分作为所述 URL 的 URL 主题分；根据所述 URL 主题分的分值大小确定出所述 URL 的优先级，并将该 URL 按其优先级存入所述 URL 队列存储模块中的对应优先级队列中。

6、如权利要求 5 所述的爬虫系统，其特征在于，还包括 URL 过滤模块连接在所述网页解析模块和采集控制模块之间;

所述 URL 过滤模块判断所述网页解析模块解析出的 URL 是否已采集，仅保留未采集过的 URL；并进一步判断未采集过的 URL 是否符合设置的 URL 过滤条件，仅将不符合所述过滤条件的未采集过的 URL 发送给所述采集控制模块。

7、一种搜索引擎，包括爬虫系统、索引系统和检索系统，其特征在于，所述爬虫系统包括: URL 队列存储模块、网页和文件下载模块、网页解析模块和采集控制模块;

所述 URL 队列存储模块，按优先级顺序存储待采集的 URL;

所述网页和文件下载模块，按 URL 优先级从高到低的顺序下载网页或文件；将下载的网页发送到所述网页解析模块，将下载的文件发送到搜索引擎的索引系统处理;

所述网页解析模块，对网页进行解析，提取网页中包含的 URL 发送到所述采集控制模块;

所述采集控制模块，计算包含所述 URL 的已采集网页的网页主题分，累加所述网页主题分作为所述 URL 的 URL 主题分；根据所述 URL 主题分的分值大小确定出所述 URL 的优先级，并将该 URL 按其优先级存入所述 URL 队列存储模块中的对应优先级队列中。

一种互联网主题文件搜索方法、爬虫系统和搜索引擎

技术领域

本发明涉及互联网文件搜索，尤其涉及一种互联网主题文件搜索方法，以及相应的爬虫系统和搜索引擎。

背景技术

Intemet 已经成为计算机领域最热门的一项技术，Internet 的普及使人们可以突破空间、地域的限制，方便地共享信息资源。www 是 Internet 上提供的最主要、应用最广泛的一种信息服务，自诞生以来得到了迅猛发展，已经成为一个巨大的信息库，存储着大量有价值的信息，人们可以在其上查找到自己感兴趣的各种内容。但在实际使用中，web 网上庞大的数据量会给用户的信息查询工作带来极大的困难。在这种情况下，各种信息检索服务应运而生，而全文检索技术是广泛采用的一项重要信息检索技术。目前，基于 www 网的全文检索技术正得到日益广泛的应用，已经有不少颇具影响的大型全文检索工具，其中比较著名的中文搜索引擎系统有www.soso.com, www.baidu.com 等，这些全文检索系统的应用对 www 网上文档信息的查询起到了巨大作用。

目前互联网搜索引擎一般由爬虫系统、索引系统、检索系统组成，爬虫系统需要从网络上不同的网站采集网页和各种文件，比如 web 网页、mp3 文件等，然后交给索引系统建立索引数据库，检索系统接收用户的检索请求，检索索引数据库，返回符合用户需求的结果。

一般互联网搜索引擎系统架构如图 1 所示，包括:

网页服务器: 提供中文搜索引擎系统网页访问服务，是用户使用中文搜索引擎系统的用户接口;

检索系统: 根据用户提交的检索关键词检索索引数据库，根据一定算法对

符合检索需求的文档进行排序、过滤，返回给网页服务器；

索引系统：对爬虫系统采集的文档进行处理，建立索引数据库；

爬虫系统：采集互联网的网页和各种文档数据。

现有技术一：采集所有的 web 网站和网页。

执行特定互联网主题文件搜索的搜索引擎中，其爬虫系统一般只采集特定主题的文件，然后建立索引，提供检索。但要采集特定主题的文件，需要采集网页，找到特定主题文件的统一资源定位符（Uniform Resource Locator，URL）链接。

目前爬虫系统一般采用遍历所有网页，即采集所有的网页和文件，然后保存需要的特定主题的文件。由于含有特定主题文件的网页很少，导致下载特定主题文件的效率很低，下载几万个网页才包含有一个特定主题文件，而且还很可能是死链。因此需要一种技术提高下载包含特定主题文件的网页的概率。

现有技术二：采集特定的主题网站和网页。

根据对采集的网页进行分析，发现一般网页间的链接具有以下特征：主题聚集性和本地性。网页普遍具有这两个特性，本地性决定同一主机的网页互相链接概率比较大，主题聚集性决定同一主题的网页互相链接概率大。

网页之间的链接特性可以用图 2 来进行模拟表示，图 2 中，每一个圆圈代表一个网页，实心圆圈代表包含 mp3 文件的网页；假设需要采集 MP3 文件，图 2 中显示出新闻主题和音乐主题的网页之间的链接和包含的 MP3 文件，结果表明：新闻主题的网页之间互相链接比较多，音乐主题的网页之间互相链接比较多，音乐主题和新闻主题之间的网页链接比较少。音乐主题的网页包含 MP3 文件的 URL 概率要比新闻主题的网页包含的 MP3 文件的 URL 概率大。

因此，现有技术二中采用对特定的主题网页进行搜索的方法。以上述采集 MP3 文件为例，MP3 搜索引擎的爬虫系统采集音乐主题网站和网页，发现和采集 MP3 文件的效率会比较高。

尽管现有技术二的采集效率较高，但由于只采集特定的少数网站，导致整

个采集的特定主题文件比较少，无法采集互联网上尽可能多的文件。

发明内容

本发明提供一种互联网主题文件搜索方法，用以解决现有技术中存在的搜索互联网主题文件效率低或采集不全面的问题。

为解决所述技术问题，本发明采用的技术方案是，提供一种互联网主题文件搜索方法，该方法包括:

A、解析下载的网页，提取网页中包含的统一资源定位符 URL;

B、计算包含所述 URL 的已采集网页的网页主题分，累加所述网页主题分作为所述 URL 的 URL 主题分;

根据所述 URL 主题分的分值大小确定出所述 URL 的对应优先级;

C、按优先级从高到低的顺序采集各 URL，建立索引，搜索所需互联网主题文件。

根据本发明的上述方法，还包括:

保存已采集的 URL 历史记录;

所述步骤 B 中，根据所述历史记录判断下载网页中包含的 URL 是否已采集，仅对未采集过的 URL 确定优先级。

根据本发明的上述方法，还包括:

设置 URL 过滤条件，仅对未采集过的不符合所述过滤条件的 URL 确定优先级。

所述网页主题分具体计算公式为:

$F(p) = a \times numFileLink \times FactorLink + b*numKeyWord \times FactorWord;$

式中，F（p）为计算出的网页主题分;

numFileLink 为该网页含有的主题文件 URL 的个数;

FactorLink 为 URL 链接的积分因子;

numKeyWord 为该网页含有的主题关键词个数;

FactorWord 为主题关键词的积分因子;

a，b 为权重因子，且 a+b =1。

同时，本发明还提供一种搜索引擎的爬虫系统，包括: URL 队列存储模块、网页和文件下载模块、网页解析模块和采集控制模块;

所述 URL 队列存储模块，按优先级顺序存储待采集的 URL;

所述网页和文件下载模块，按 URL 优先级从高到低的顺序下载网页或文件；将下载的网页发送到所述网页解析模块，将下载的文件发送到搜索引擎的索引系统处理;

所述网页解析模块，对网页进行解析，提取网页中包含的 URL 发送到所述采集控制模块;

所述采集控制模块，计算包含所述 URL 的已采集网页的网页主题分，累加所述网页主题分作为所述 URL 的 URL 主题分;

根据所述 URL 主题分的分值大小确定出所述 URL 的优先级，并将该 URL 按其优先级存入所述 URL 队列存储模块中的对应优先级队列中。

根据本发明提供的上述爬虫系统，还包括 URL 过滤模块连接在所述网页解析模块和采集控制模块之间;

所述 URL 过滤模块判断所述网页解析模块解析出的 URL 是否已采集，仅保留未采集过的 URL；并进一步判断未采集过的 URL 是否符合设置的 URL 过滤条件，仅将不符合所述过滤条件的未采集过的 URL 发送给所述采集控制模块。

对应于所述爬虫系统，本发明还提供一种搜索引擎，包括爬虫系统、索引系统和检索系统，所述爬虫系统包括: URL 队列存储模块、网页和文件下载模块、网页解析模块和采集控制模块;

所述 URL 队列存储模块，按优先级顺序存储待采集的 URL;

所述网页和文件下载模块，按 URL 优先级从高到低的顺序下载网页或文件；将下载的网页发送到所述网页解析模块，将下载的文件发送到搜索引擎的

索引系统处理;

所述网页解析模块，对网页进行解析，提取网页中包含的 URL 发送到所述采集控制模块;

所述采集控制模块，计算包含所述 URL 的已采集网页的网页主题分，累加所述网页主题分作为所述 URL 的 URL 主题分；根据所述 URL 主题分的分值大小确定出所述 URL 的优先级，并将该 URL 按其优先级存入所述 URL 队列存储模块中的对应优先级队列中。

本发明有益效果如下:

（1）本发明通过解析下载网页，提取网页中包含的统一资源定位符 URL;对各 URL 根据预定规则确定优先级，优先采集优先级较高的 URL，搜索所需主题文件；由于优先级较高的 URL 与主题文件的关系较密切，搜索出相关主题文件的可能性较大，因此，采用本发明能提高搜索效率。

（2）本发明不局限于对某些特定网站进行搜索，可以根据 URL 优先级搜索各相关网页，因此，可以做到在整个 Internet 上进行搜索。

附图说明

图 1 为现有技术中文信息检索系统架构图;

图 2 为不同主题之间的网页链接示意图;

图 3 为本发明提供的爬虫系统结构示意图;

图 4 为本发明方法流程图。

具体实施方式

参见图 3，为本发明提供的爬虫系统 1 结构示意图。包括：网页和文件下载模块 11、网页解析模块 12、URL 过滤模块 13、采集控制模块 14 和 URL 队列存储模块 15。

下面对各模块的功能进行详细描述。

网页和文件下载模块 11：使用 HTTP、FTP 协议下载网页或文件，并把下载的网页提交给网页解析模块 12，把下载的文件提交给搜索引擎的索引系统建立索引数据库；

爬虫系统 1 刚开始启动运行时，设置一些种子 URL 放入 URL 队列存储模块 15 的最高优先级 URL 队列（其对应 URL 主题分为一个默认初始值），例如一些常见的目录导航网页，如www.hao123.com，网页和文件下载模块 11 从 URL 队列获取种子 URL，然后下载网页并发送到网页解析模块 12 进行解析。

网页解析模块 12：解析 HTML 网页，提取网页包含的 URL 链接，并提交给 URL 过滤模块 13。

URL 过滤模块 13：判断各 URL 是否已经采集，如果未采集，判断是否符合设定的过滤条件，如果当前 URL 未采集且不符合设定的过滤条件，则作为待采集 URL 发送给采集控制模块 14；

在该 URL 过滤模块 13 中，保存已采集的 URL 历史记录；根据保存的历史记录判断下载网页中包含的 URL 是否已采集，并将已采集的 URL 实时存入历史记录中进行记录更新；

在该 URL 过滤模块 13 中，还可以存储过滤条件，例如：过滤条件为设定的 URL 黑名单，URL 过滤模块 13 根据该过滤条件判断当前的 URL 是否位于黑名单中，如果当前 URL 位于设置的黑名单中，则判定该 URL 符合设定的过滤条件，该 URL 将被过滤掉，不被发送到采集控制模块 14；否则，URL 过滤模块 13 将网页解析模块 12 发送过来的判断为未采集且不符合过滤条件的 URL 全部发送到采集控制模块 14 进行处理。

采集控制模块 14，采用预定算法计算出待采集 URL 的 URL 的主题分，根据各 URL 主题分的分值大小确定对应 URL 的优先级；并将各 URL 根据其对应的优先级存入到 URL 队列存储模块 15 的不同优先级队列中；

URL 主题分的具体计算方法如下：

$$S(url) = \sum_{\text{已经采集的包含该url的网页}} F(p) \qquad \text{式（1）}$$

式（1）中，S（url）为该URL的URL主题分，F（p）为网页的主题分。即一个URL的主题分为所有已经采集的包含该URL的网页的主题分之和。

其中：

F(p) = a*numFileLink*FactorLink + b*numKeyWord*FactorWord 式(2)

式（2）中，F（p）为计算出的包含该URL的网页对应的网页主题分；

numFileLink为该网页含有的主题文件URL的个数；

FactorLink为URL链接的积分因子；

numKeyWord为该网页含有的主题关键词个数；

FactorWord为主题关键词的积分因子；

a，b为权重因子，且a+b=1；

也就是说一个网页的主题分与包含的主题文件个数和主题关键词个数的相关，其包括主题文件越多，主题关键词越多，则该网页的主题分越大。

URL队列存储模块15：保存有多个不同优先级的URL队列，并根据URL的主题分大小把待采集URL放入到不同的优先级队列；例如：保存有三个队列，分别为第一优先级队列，第二优先级队列和第三优先级队列，URL根据主题分大小划分成三个不同的区间，其中，第一优先级队列级别最高，存储主题分最大区间的待采集URL，第二优先级队列次之，第三优先级队列级别最低；网页和文件下载模块11首先采集级别最高的第一优先级队列中的URL，只有当第一优先级队列为空后（由于已采集过的URL将从队列中删除，如果第一优先级队列中的URL都被采集，则该队列将为空），才顺序采集第二优先级队列和第三优先级队列中的URL；

该URL队列存储模块15中存储的URL队列个数可随意设置，本发明对此不作限定。

根据本发明提供的上述爬虫系统1，本发明提供一种主题文件搜索方法，其具体流程如图4所示，包括：

步骤S11、网页解析模块解析网页和文件下载模块下载的网页，并对网页

进行解析，提取网页包含的 URL，并发送到 URL 过滤模块；

步骤 S12、URL 过滤模块判断当前的 URL 是否已采集，或者是否符合设定的过滤条件需要被过滤掉；如果判断结果表明当前 URL 已被采集或符合设定的过滤条件，则丢弃该 URL，流程转至步骤 S11，由网页解析模块继续提取网页中包含的其它 URL；如果判断结果表明当前 URL 未被采集或不符合设定的过滤条件，则发送该 URL 到采集控制模块，继续下列步骤；

步骤 S13、采集控制模块根据主题文件采集算法（如采用上述式（1）、式（2）所定义的具体算法）计算出该 URL 对应的 URL 主题分；

步骤 S14、采集控制模块根据设定的 URL 主题分与优先级的对应关系，确定出该 URL 的优先级，将该 URL 存入到 URL 队列存储模块的对应优先级队列中；

步骤 S15、网页和文件下载模块从高优先级队列开始依次读取 URL 进行下载；将下载的网元发送到网页解析模块处理，将下载的文件发送到搜索引擎的索引系统。

综上所述，本发明通过解析下载网页，提取网页中包含的 URL；对各 URL 根据 URL 主题分计算方法计算主题分，根据预定规则确定优先级，放入不同的优先级队列，优先采集优先级较高的 URL，搜索所需主题文件；由于优先级较高的 URL 与主题文件的关系较密切，搜索出相关主题文件的可能性较大，因此，采用本发明能提高搜索效率。

另外，本发明可以做到在整个 Internet 上进行搜索，不局限于对某些特定网站，搜索全面，充分满足用户需要。

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。

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。

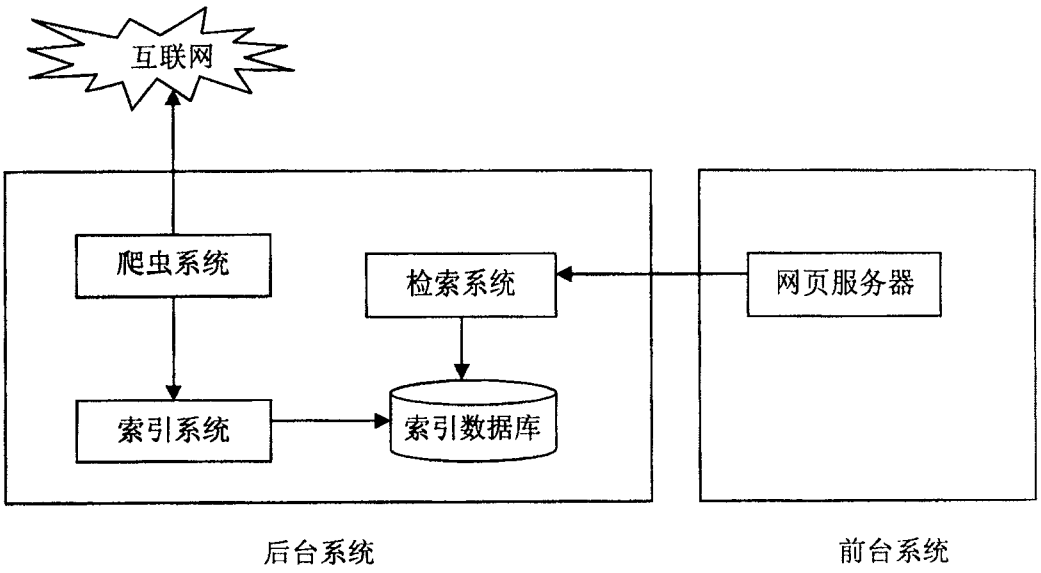


图 1

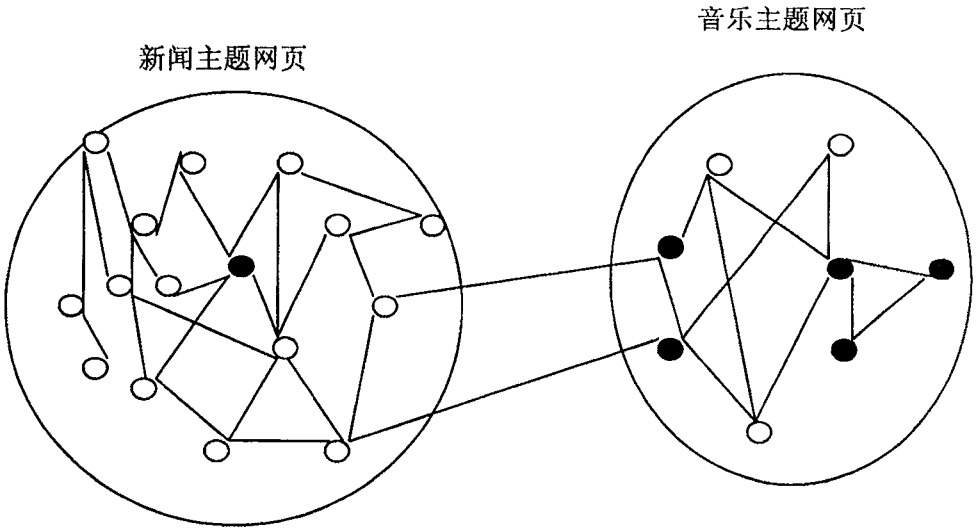

新闻主题网页
音乐主题网页

图 2

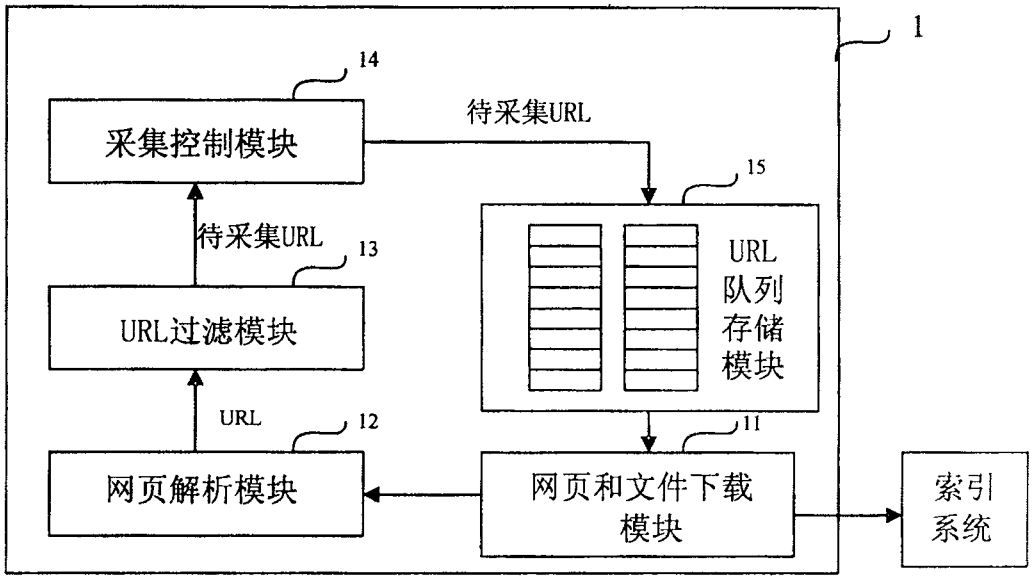

1
14
采集控制模块
待采集URL
15
待采集URL
13
URL过滤模块
URL
队列
存储
模块
URL
12
11
网页解析模块
网页和文件下载
模块
索引
系统

图 3

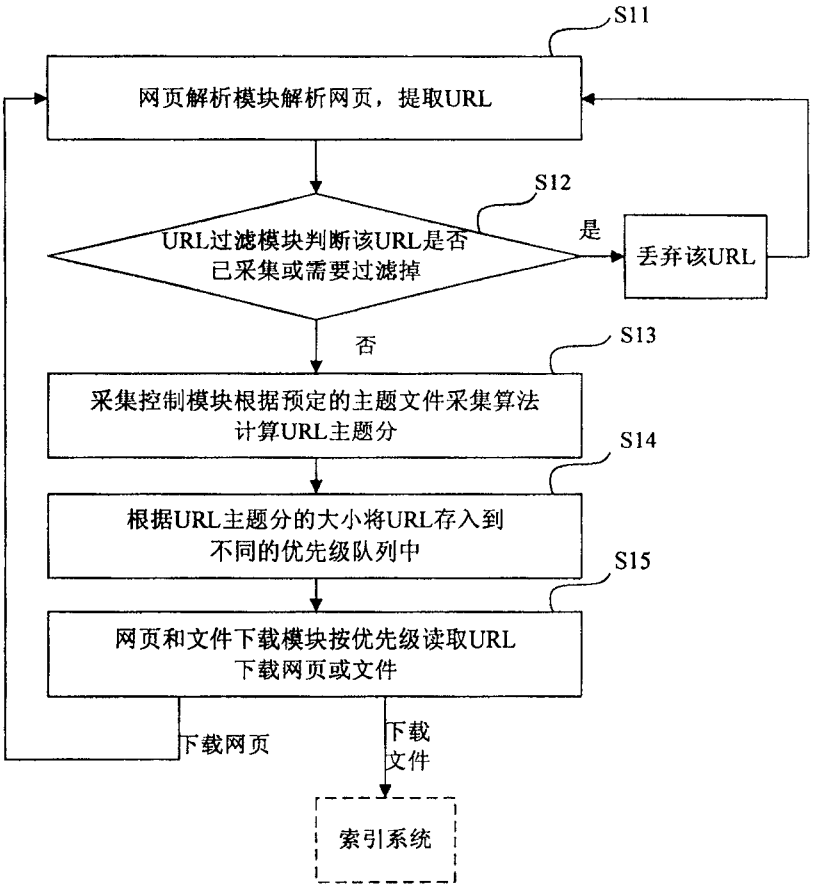

S11
网页解析模块解析网页，提取URL
S12
URL过滤模块判断该URL是否
已采集或需要过滤掉
是
丢弃该URL
否
S13
采集控制模块根据预定的主题文件采集算法
计算URL主题分
S14
根据URL主题分的大小将URL存入到
不同的优先级队列中
S15
网页和文件下载模块按优先级读取URL
下载网页或文件
下载网页
下载
文件
索引系统

图 4