(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 109977328 A
(43)申请公布日 2019.07.05

(21)申请号 201910168704.2

(22)申请日 2019.03.06

(71)申请人 杭州迪普科技股份有限公司
地址 310051 浙江省杭州市滨江区通和路68号中财大厦6楼

(72)发明人 黄晓炼

(74)专利代理机构 北京博思佳知识产权代理有限公司 11415
代理人 林祥

(51)Int.Cl.
*G06F 16/955*(2019.01)
*G06F 16/958*(2019.01)

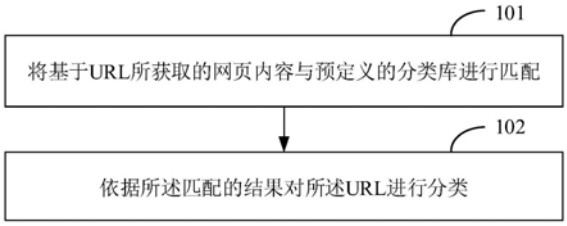
权利要求书1页 说明书6页 附图3页

(54)发明名称

一种URL分类方法及装置

(57)摘要

本申请提供一种URL分类方法及装置。本申请将基于URL所获取的网页内容与预定义的分类库进行匹配，依据所述匹配的结果对所述URL进行分类，基于URL对应的网页内容进行匹配的分类方式，能够全面匹配URL对应网页内容所涉及的各个关键词信息，依据网页内容的对分类库中的各个关键词的匹配情况，确定该URL的对应的分类，提高了分类准确率。




CN 109977328 A

1.一种URL分类方法,其特征在于,所述方法包括:

将基于URL所获取的网页内容与预定义的分类库进行匹配;

依据所述匹配的结果对所述URL进行分类。

2.根据权利要求1所述方法,其特征在于,所述依据所述匹配的结果对所述URL进行分类,包括:

统计所述分类库中匹配成功的类型对应的匹配次数;

当任一类型对应的匹配次数超过预设阈值时,将所述URL划分入所述类型中。

3.根据权利要求1或2所述方法,其特征在于,还包括:

统计匹配成功的所述分类库中的类型对应的匹配次数;

将所述分类库中匹配次数最高的类型确定为所述URL地址对应的分类类型。

4.根据权利要求1所述方法,其特征在于,所述方法还包括:

将所述URL地址按照所述分类类型进行分类存储。

5.根据权利要求1所述方法,其特征在于,所述方法还包括:

按照预设的时间间隔采集URL;

将所述URL加入到待提取队列中;

解析所述待提取队列中的URL,以获取所述URL对应的网页内容。

6.一种URL分类装置,其特征在于,所述装置包括:

匹配单元,将基于URL所获取的网页内容与预定义的分类库进行匹配;

第一分类单元,依据所述匹配的结果对所述URL进行分类。

7.根据权利要求6所述装置,其特征在于,所述第一分类单元,包括:

第一统计单元,统计所述分类库中匹配成功的类型对应的匹配次数;

第二分类单元,当任一类型对应的匹配次数超过预设阈值时,将所述URL划分入所述类型中。

8.根据权利要求1或2所述装置,其特征在于,还包括:

第二统计单元,统计匹配成功的所述分类库中的类型对应的匹配次数;

第三分类单元,确定所述分类库中匹配次数最高的类型为所述URL地址对应的分类类型。

9.根据权利要求1所述装置,其特征在于,还包括:

存储单元,将所述URL地址按照所述分类类型进行分类存储。

10.根据权利要求1所述装置,其特征在于,还包括:

采集单元,按照预设的时间间隔采集URL;

补充单元,将所述URL加入到待提取队列中;

解析单元,解析所述待提取队列中的URL,以获取所述URL对应的网页内容。

# 一种URL分类方法及装置

## 技术领域

[0001] 本申请涉及网络技术领域，具体涉及一种URL分类方法及装置。

## 背景技术

[0002] 随着互联网的迅速发展，网络数据量急剧膨胀，这使得网页信息的大量、广泛、分散等不易管理的特点愈发显著。面对数据庞大的网页信息资源，需要对网页信息进行分类整理，目前，将网页的统一资源定位符(URL)进行分类是较为有效的网页分类、整理的方法之一。

[0003] 现有技术中往往采用对URL的构成进行分析，从而依据分析结果进行分类，例如，将获取的URL地址与设定的正则表达式bbs.XXX.com(cn)或者XXX.com(cn)/bbs进行比较，凡是匹配成功的，则归类为bbs(论坛)。然而，随着URL构成的千变万化，使得上述方法误报漏报的概率逐渐升高，导致分类失真。

## 发明内容

[0004] 有鉴于此，本申请提供一种URL分类方法及装置，能够降低相关技术中对URL匹配结果的误报、漏报的情况，保证了对URL所述类别评估的准确性，提高了分类效率。

[0005] 为实现上述目的，本申请提供技术方案如下：

[0006] 将基于URL所获取的网页内容与预定义的分类库进行匹配；

[0007] 依据所述匹配的结果对所述URL进行分类。

[0008] 根据本申请的第二方面，提出了一种URL分类装置，包括：

[0009] 匹配单元，将基于URL所获取的网页内容与预定义的分类库进行匹配；

[0010] 第一分类单元，依据所述匹配的结果对所述URL进行分类。

[0011] 由以上技术方案可见，将所获取的URL对应的网页内容与预定义的分类库进行匹配，从而依据匹配的结果对URL进行分类，基于URL对应的网页内容进行匹配的分类方式，能够全面匹配URL对应网页内容所涉及的各个关键词信息，依据网页内容的对分类库中的各个关键词的匹配情况，确定该URL的对应的分类，提高了分类准确率，实现对所采集的URL的有效分类。

## 附图说明

[0012] 图1是根据本申请一示例性实施例中的一种URL分类方法的流程图；

[0013] 图2是根据本申请一示例性实施例中的另一种URL分类方法的流程图；

[0014] 图3是根据本申请一示例性实施例中的一种电子设备的示意结构图；

[0015] 图4是根据本申请一示例性实施例中的一种URL分类装置的框图。

## 具体实施方式

[0016] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及

附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

[0017] 在本申请使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本申请。在本申请和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0018] 应当理解，尽管在本申请可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本申请范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0019] 统一资源定位符(Uniform Resource Locator，简称URL)是对可以从互联网上得到的资源的位置和访问方法的一种简洁的标识，是互联网上标准资源的地址，也被称为网页地址，以用来表示网页的位置。从URL的具体构成上来看，一个URL往往含有协议信息、服务器名、域名等信息，例如当一个URL为http://www.***.com/index.html时，通过该URL便可获知：该URL中含有的协议为HTTP超文本传输协议，网站名为www.***.com，其由服务器名www和域名***.com所组成，且URL的根目录下的网页为默认网页index.html。

[0020] 在相关技术中通过对URL的构成进行分析或者人工查看所获取的网页的URL地址，从而实现对URL进行分类。例如：将获取的URL地址与设定的正则表达式bbs.XXX.com或者XXX.com(cn)/bbs进行匹配，凡是匹配成功的，则归类为bbs(论坛)，但是对于一些构成上不含有bbs的论坛类URL，则无法进行正确的归类，因而在URL的构成千变万化的应用场景中，通过正则表达式进行匹配的分类方法的误报概率较高。而通过人工查看所获取到的URL对应的网页内容并将其进行归类，该方法虽然能够保证一定的准确率，但针对大量的待分类的URL的情况下，需要投入大量的人工，用工成本高且耗费时间。

[0021] 有鉴于此，本申请提供一种URL分类方法，以解决相关技术在实施URL分类时存在的技术问题。

[0022] 下面结合附图，对本申请的具体实施方案进行详细阐述。

[0023] 为对本申请进行进一步说明，提供下列实施例：

[0024] 图1是本申请一示例性实施例提供的一种URL分类方法的流程图，如图1所示，该方法可以包括以下步骤：

[0025] 步骤101，将基于URL所获取的网页内容与预定义的分类库进行匹配。

[0026] 步骤102，依据所述匹配的结果对所述URL进行分类。

[0027] 在一实施例中，可以统计所述分类库中匹配成功的类型对应的匹配次数，从而当任一类型对应的匹配次数超过预设阈值时，将所述URL划分入所述类型中。

[0028] 在本实施例中，由于网页内容往往具有信息量较大的文字信息，因而基于网页内容的信息匹配将涉及到分类库中较多的类型，但相较于网页内容的主旨信息的关键词，非必要的网页内容信息的关键词所匹配到的次数也往往在于少数，所以通过设定一定的阈值，使得仅对于超过阈值的匹配次数对应的分类类型作为该URL对应的分类，避免了在基于

网页内容匹配的过程中因非重要信息而匹配到的类型对URL进行准确匹配的干扰。

[0029] 在一实施例中，可以统计在所述分类库中匹配成功的类型对应的匹配次数，并将所述分类库中匹配次数最高的类型确定为待匹配的URL对应的分类类型。

[0030] 在需要对URL进行精准化匹配的应用场景中，通过所设定的阈值对URL匹配次数进行过滤的方法则无法保证网页URL仅匹配到的一个类型，例如：在基于URL1的内容进行分析和匹配后确定该URL1在分类库中匹配到的分类类型及匹配情况为：类型A匹配到10次、类型B匹配到8次、类型C匹配到1次、类型D匹配到4次和类型E匹配到18次，则超过预设的阈值5次的分类类型涉及到类型A、类型B和类型E，显然该URL1所匹配到的类型有3种，而并非精准化匹配要求的1种，因而可采用对所匹配到的类型的匹配次数的大小进行比较，将匹配次数最高的类型确定为待匹配的URL对应的分类类型。

[0031] 在一实施例中，可以将所述URL按照所述分类类型进行分类存储，从而便于后续实现对于各个分类的URL进行批量处理，例如在阻断或者放通与某一敏感词有关的网页内容的应用场景中，可通过获取符合该敏感词的分类类型，从而实现对基于该分类类型的URL进行统一处理，提高处理效率。

[0032] 在一实施例中，可以按照预设的时间间隔触发采集URL数据，并将所述URL数据中的URL地址加入到待提取队列中，逐次解析所述待提取队列中的URL地址，并记录其对应的网页内容。

[0033] 在本实施例中，经过预设时间间隔自动触发对URL的获取，并将所获取的URL添加到待提取的队列中，从而无需人工提取工作便可触发对待解析的URL进行分析并记录该URL对应的网页内容，形成了关于URL的获取、解析、匹配和存储集成式的分类系统，提高了对于URL分类效率。

[0034] 通过上述实施例，将所获取的URL对应的网页内容与预定义的分类库进行匹配，从而依据匹配的结果对URL进行分类，基于URL对应的网页内容进行匹配的分类方式，能够全面匹配URL对应网页内容所涉及的各个关键词信息，依据网页内容的对分类库中的各个关键词的匹配情况，确定该URL的对应的分类，提高了分类准确率，弥补了相关技术中的缺陷。

[0035] 下面结合附图，对本发明的具体实施方案进行详细阐述。

[0036] 请参见图2，图2是根据本申请一示例性实施例中的另一种URL分类方法的流程图，所述方法包括：

[0037] 步骤201，按照预设的时间间隔采集URL。

[0038] 在一实施例中，可以从任一预配置的URL存储空间中采集URL数据，亦可以对已获取的URL对应的网页内容中的URL进行采集，或者在已进行分类的网站上的URL进行采集，本申请对URL的具体采集来源不做限定。

[0039] 在一实施例中，具体的采集方式可以为通过爬虫技术爬取，或是其他能够实现对于网页的URL进行捕捉的技术，诸如借用嗅探器等方式，本申请对获取URL的具体方式不做限定。

[0040] 步骤202，将所采集的URL加入到待提取队列中。

[0041] 步骤203，解析所述待提取队列中的URL，以获取所述URL对应的网页内容。

[0042] 在一实施例中，通过对待提取队列中的URL进行解析，从而能够获取该URL对应的网页内容，诸如网页标题或是正文内容等，以便对于网页内容进行匹配分析，提高网页内容

的类型匹配的全面性。

[0043] 步骤204,将基于URL所获取的网页内容与预定义的分类库进行匹配。

[0044] 在一实施例中,若所获取的URL已含有类型的标识信息,则可获取该类型的标识信息,并在第一分类库中进行初步筛选,以定义出与该类型的标识相关的第二分类库,从而通过第二分类库对URL对应的网页内容进行匹配。例如所采集的URL为:http://www.***.com/index.html,并获得该URL中的***对应的分类标识信息为A,则可在进行匹配之前,可预先在第一分类库中筛选出与分类标识A相关的类型信息,例如筛选后得到类型M、类型N、类型O、类型P和类型Q,从而使用类型M、类型N、类型O、类型P和类型Q所构成的第二分类库作为对该URL对应的网页内容进行匹配的分类库。

[0045] 通过本实施例,通过基于URL已含有的分类信息对第一分类库中的类型进行筛选,以构成与URL已含有的分类信息更为接近匹配的第二类型库,相较于类型范围广、数量众多的第一分类库,第二分类库因数量少、更为接近的类型信息实现大大降低了匹配进程的耗时的技术效果,从而提高了匹配效率。

[0046] 步骤205,统计在分类库中匹配成功的类型对应的匹配次数。

[0047] 在一实施例中,可以由管理员对URL分类库进行预先配置,或直接调用已有的URL分类库;进一步地,分类库中含有关键字的标识信息与一个或多个类型的标识信息之间的对应关系,使得当对网页内容进行匹配时,可以通过统计各个关键字在网页内容中出现次数,以确定关键字对应类型的匹配次数。

[0048] 步骤206,判断是否需要进行精细化的匹配,若不需要,则进行步骤207,否则进行步骤208。

[0049] 步骤207,当任一类型对应的匹配次数超过预设阈值时,将所述URL划分入所述类型中。

[0050] 在一实施例中,可以对各个类型信息均设置同一阈值,或者针对各个类型分别设置相应的阈值,本申请对此不做限制。

[0051] 步骤208,将所述分类库中匹配次数最高的类型确定为所述URL地址对应的分类类型。

[0052] 步骤209,将所述URL地址按照已划分的分类类型进行分类存储。

[0053] 步骤210,判断是否存在未被分类的URL,若是,则进行步骤211,否则结束分类配置。

[0054] 步骤211,显示未进行分类的URL对应的网页内容。

[0055] 步骤212,接收对该网页内容对应的URL的分类设置,返回步骤209。

[0056] 通过上述实施例,通过设定的阈值实现对URL的自动分类,减少低效的人工投入,节省了时间,提高了工作效率。

[0057] 图3是根据本申请一示例性实施例中的一种电子设备的示意结构图。请参考图3,在硬件层面,该电子设备包括处理器、内部总线、网络接口、内存以及非易失性存储器,当然还可能包括其他业务所需要的硬件。处理器从非易失性存储器中读取对应的计算机程序到内存中然后运行,在逻辑层面上形成URL分类装置。当然,除了软件实现方式之外,本申请并不排除其他实现方式,比如逻辑器件抑或软硬件结合的方式等等,也就是说以下处理流程的执行主体并不限定于各个逻辑单元,也可以是硬件或逻辑器件。

[0058] 请参考图4,在软件实施方式中,该URL分类装置可以包括:

[0059] 匹配单元401,将基于URL所获取的网页内容与预定义的分类库进行匹配;

[0060] 第一分类单元402,依据所述匹配的结果对所述URL进行分类。

[0061] 可选的,所述第一分类单元402具体用于:

[0062] 第一统计单元403,统计所述分类库中匹配成功的类型对应的匹配次数;

[0063] 第二分类单元404,当任一类型对应的匹配次数超过预设阈值时,将所述URL划分入所述类型中。

[0064] 可选的,还包括:

[0065] 第二统计单元405,统计匹配成功的所述分类库中的类型对应的匹配次数;

[0066] 第三分类单元406,确定所述分类库中匹配次数最高的类型为所述URL地址对应的分类类型。

[0067] 可选的,还包括:

[0068] 存储单元407,将所述URL地址按照所述分类类型进行分类存储。

[0069] 可选的,还包括:

[0070] 采集单元408,按照预设的时间间隔采集URL;

[0071] 补充单元409,将所述URL加入到待提取队列中;

[0072] 解析单元410,解析所述待提取队列中的URL,以获取所述URL对应的网页内容。

[0073] 所述装置与上述方法相对应,更多相同的细节不再一一赘述。

[0074] 在一个典型的配置中,计算设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0075] 内存可能包括计算机可读介质中的非永久性存储器,随机存取存储器(RAM)和/或非易失性内存等形式,如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。

[0076] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括,但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,可用于存储可以被计算设备访问的信息。按照本文中的界定,计算机可读介质不包括暂存电脑可读媒体(transitory media),如调制的数据信号和载波。

[0077] 对于装置实施例而言,由于其基本对应于方法实施例,所以相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本申请方案的目的。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0078] 虽然本说明书包含许多具体实施细节,但是这些不应被解释为限制任何发明的范围或所要求保护的范围,而是主要用于描述特定发明的具体实施例的特征。本说明书内在

多个实施例中描述的某些特征也可以在单个实施例中被组合实施。另一方面，在单个实施例中描述的各种特征也可以在多个实施例中分开实施或以任何合适的子组合来实施。此外，虽然特征可以如上所述在某些组合中起作用并且甚至最初如此要求保护，但是来自所要求保护的组合中的一个或多个特征在一些情况下可以从该组合中去除，并且所要求保护的组合可以指向子组合或子组合的变型。

[0079] 以上所述仅为本申请的较佳实施例而已，并不用以限制本申请，凡在本申请的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本申请保护的范围之内。

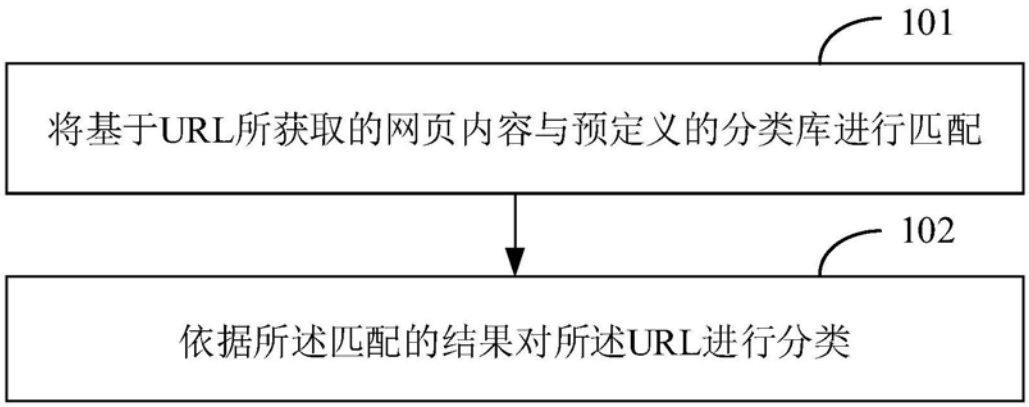

101
将基于URL所获取的网页内容与预定义的分类库进行匹配
102
依据所述匹配的结果对所述URL进行分类

图1

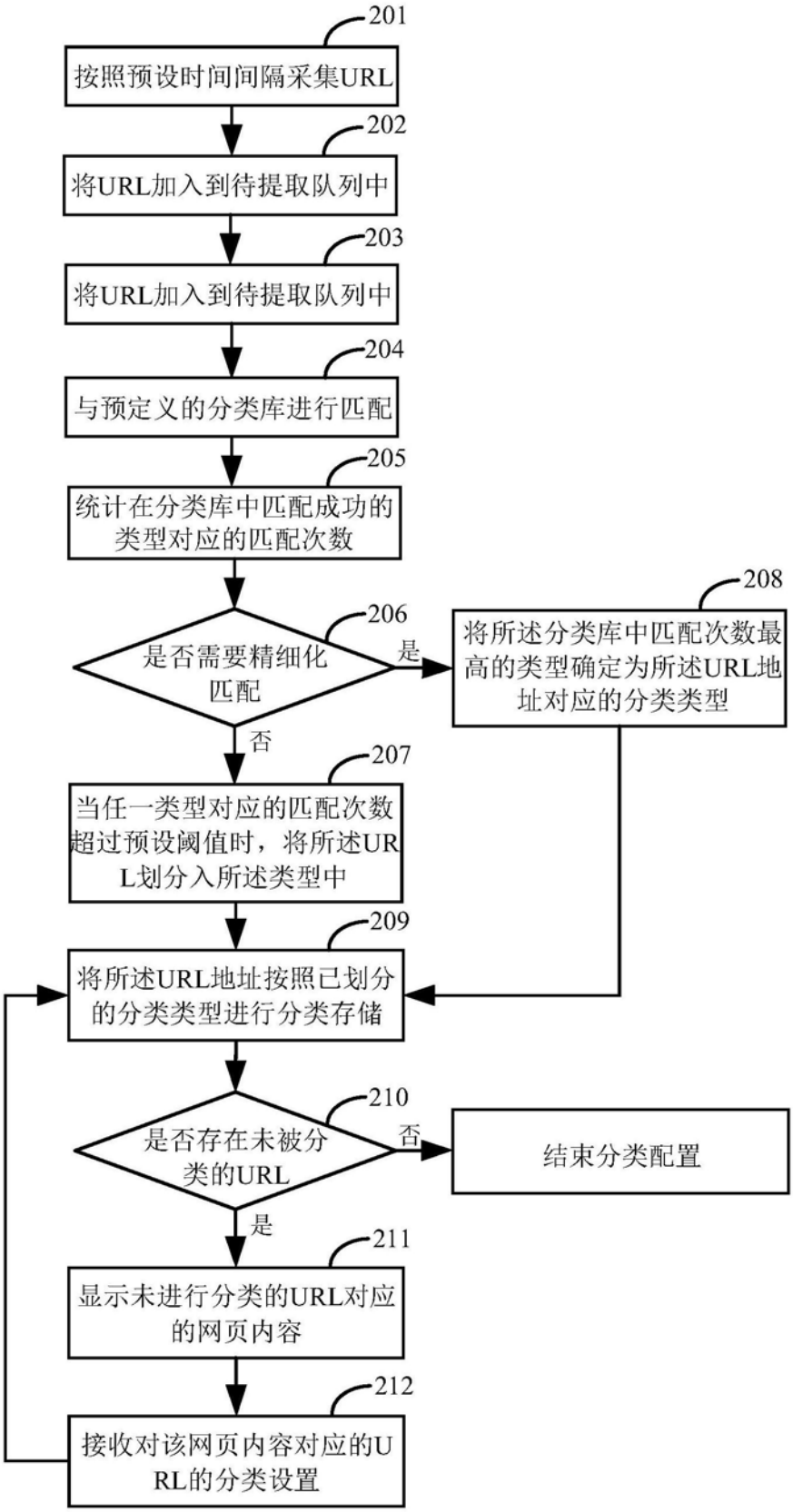

201
按照预设时间间隔采集URL
202
将URL加入到待提取队列中
203
将URL加入到待提取队列中
204
与预定义的分类库进行匹配
205
统计在分类库中匹配成功的类型对应的匹配次数
206
是否需要精细化匹配
是
208
将所述分类库中匹配次数最高的类型确定为所述URL地址对应的分类类型
否
207
当任一类型对应的匹配次数超过预设阈值时，将所述URL划分入所述类型中
209
将所述URL地址按照已划分的分类类型进行分类存储
210
是否存在未被分类的URL
否
结束分类配置
是
211
显示未进行分类的URL对应的网页内容
212
接收对该网页内容对应的URL的分类设置

图2

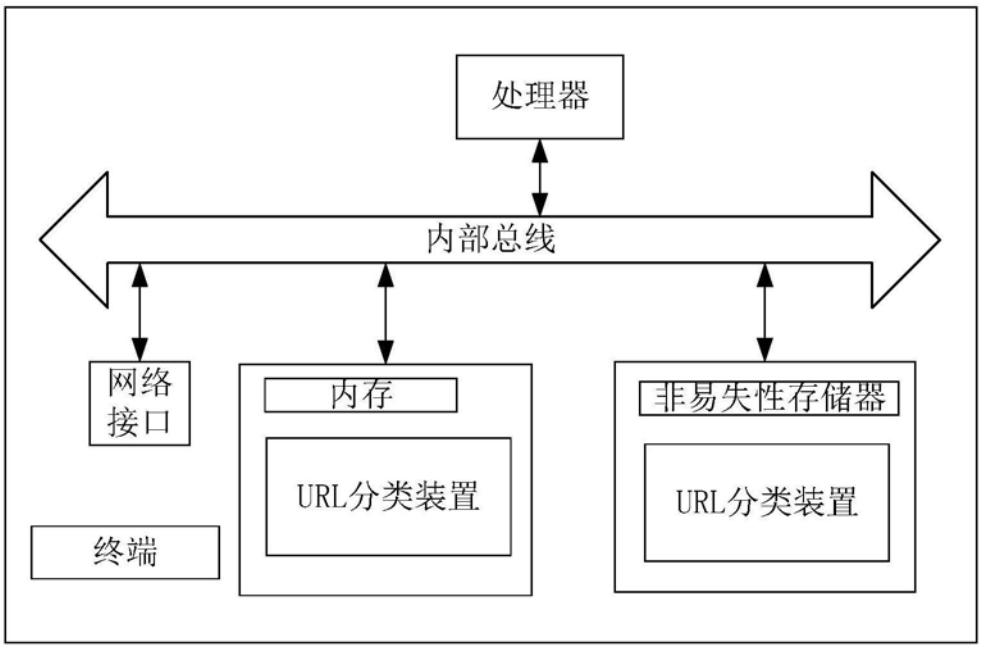

处理器
内部总线
网络接口
内存
URL分类装置
非易失性存储器
URL分类装置
终端

图3

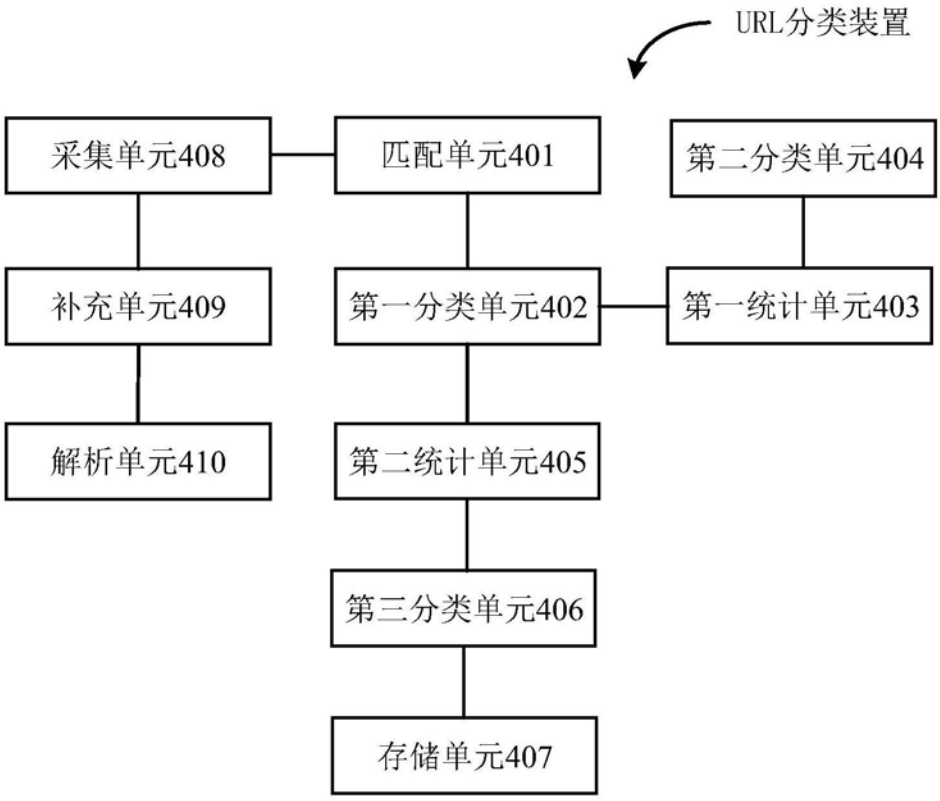

URL分类装置
采集单元408
匹配单元401
第二分类单元404
补充单元409
第一分类单元402
第一统计单元403
解析单元410
第二统计单元405
第三分类单元406
存储单元407

图4