(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 103927325 A
(43)申请公布日 2014.07.16

(21)申请号 201410093939.7

(22)申请日 2014.03.13

(71)申请人 中国联合网络通信集团有限公司
地址 100033 北京市西城区金融大街 21 号

(72)发明人 贾卷群

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 栗若木 白莹

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

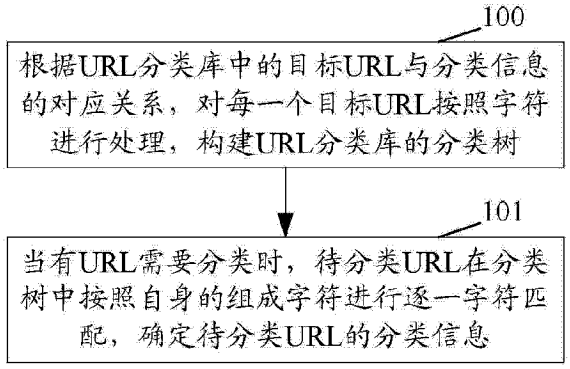
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种对 URL 进行分类的方法及装置

(57)摘要

本发明公开了一种对统一资源定位符(URL)进行分类的方法及装置,包括根据 URL 分类库中的目标 URL 与分类信息的对应关系,对每一个目标URL按照字符进行处理,构建URL分类库的分类树;当有URL需要分类时,待分类URL在分类树中按照自身的组成字符进行逐一字符匹配,确定待分类URL的分类信息。本发明方法中,在建立分类树时,以目标URL字符的ASCII码的数值作为Tree中的节点数组的下标来构建URL树,即采用树型结构存储目标URL分类库。这样,在进行URL分类时,只需要对待分类URL做一次逐一字符的访问,即可完成分类过程,提高了分类效率并实现了快速分类,进而实现了对万亿级的上网记录数据中用户访问的URL的快速分类。



CN 103927325 A

1. 一种对统一资源定位符 URL 进行分类的方法，其特征在于，包括：根据 URL 分类库中的目标 URL 与分类信息的对应关系，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树；

当有 URL 需要分类时，待分类 URL 在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类 URL 的分类信息。

2. 根据权利要求 1 所述的方法，其特征在于，该方法之前还包括：

建立所述 URL 分类库，URL 分类库中的内容包括：一个或一个以上目标 URL，各目标 URL 对应的分类信息。

3. 根据权利要求 2 所述的方法，其特征在于，所述对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树包括：

依次取出所述目标 URL 中的字符，以取出的字符的 ASCII 码作为分类树中的节点数组 node[] 的下标；

访问所述 node[ASCII] 的节点，直至目标 URL 的最后一个字符，则将所述对应的分类信息存储到节点数组 node[ASCII] 的目标 obj 域。

4. 根据权利要求 3 所述的方法，其特征在于，所述确定待分类 URL 的分类信息包括：

依次取出所述待分类 URL 中一个字符，以该字符的 ASCII 码作为所述分类树中的节点数组 node[] 的下标，访问所述分类树的节点数组 node[ASCII] 的节点，如果其对应的目标 obj 域已存有分类信息，则确定所述目标 obj 域中的分类信息为待分类 URL 的分类信息。

5. 根据权利要求 4 所述的方法，其特征在于，如果所述目标 obj 域为空，且已经访问到所述待分类 URL 的最后一个字符，该方法还包括：确定所述分类信息为空。

6. 一种对 URL 进行分类的方法，其特征在于，至少包括构建模块、分类模块，其中，

构建模块，用于根据 URL 分类库中的目标 URL 与分类信息的对应关系，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树；

分类模块，用于当有 URL 需要分类时，待分类 URL 在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类 URL 的分类信息。

一种对 URL 进行分类的方法及装置

技术领域

[0001] 本发明涉及互联网技术，尤指一种对统一资源定位符（URL，Uniform Resource Locator）进行分类的方法及装置。

背景技术

[0002] 目前，在对上网记录数据中用户访问的某个 URL 进行分类时，需要将该待分类 URL 的整个 URL 与分类库中的各目标 URL 逐一进行匹配，如果能够匹配上，则返回匹配上的分类库中的目标 URL 对应的分类信息。以分类库中有 100 个目标 URL 为例，一次分类过程的平均匹配次数会在 50 次以上，而最大匹配系数则会是 100 次。

[0003] 在现有对 URL 进行分类的方法中，对 URL 的匹配是：将组成待分类 URL 的整个字符串与分类库中的各目标 URL 逐一进行比较，效率不高；而且，当分类库中的目标 URL 数量增加时，匹配时间将大幅增加。

[0004] 由于互联网产业的迅速发展，用户上网记录数据的数量也呈现爆发式的增长，这样，要实现对万亿级的上网记录数据中用户访问的 URL 的快速分类，成为了一个难题。

发明内容

[0005] 为了解决上述技术问题，本发明提供了一种对 URL 进行分类的方法及装置，能够提高效率并实现快速分类。

[0006] 为了达到本发明目的，本发明提供了一种对 URL 进行分类的方法，包括：根据 URL 分类库中的目标 URL 与分类信息的对应关系，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树；

[0007] 当有 URL 需要分类时，待分类 URL 在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类 URL 的分类信息。

[0008] 该方法之前还包括：

[0009] 建立所述 URL 分类库，URL 分类库中的内容包括：一个或一个以上目标 URL，各目标 URL 对应的分类信息。

[0010] 所述对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树包括：

[0011] 依次取出所述目标 URL 中的字符，以取出的字符的 ASCII 码作为分类树中的节点数组 node[] 的下标；

[0012] 访问所述 node[ASCII] 的节点，直至目标 URL 的最后一个字符，则将所述对应的分类信息存储到节点数组 node[ASCII] 的目标 obj 域。

[0013] 所述确定待分类 URL 的分类信息包括：

[0014] 依次取出所述待分类 URL 中一个字符，以该字符的 ASCII 码作为所述分类树中的节点数组 node[] 的下标，访问所述分类树的节点数组 node[ASCII] 的节点，如果其对应的目标 obj 域已存有分类信息，则确定所述目标 obj 域中的分类信息为待分类 URL 的分类信息。

[0015] 如果所述目标 obj 域为空，且已经访问到所述待分类 URL 的最后一个字符，该方法还包括：确定所述分类信息为空。

[0016] 本发明还公开了一种对 URL 进行分类的方法，至少包括构建模块、分类模块，其中，

[0017] 构建模块，用于根据 URL 分类库中的目标 URL 与分类信息的对应关系，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树；

[0018] 分类模块，用于当有 URL 需要分类时，待分类 URL 在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类 URL 的分类信息。

[0019] 与现有技术相比，本发明包括根据 URL 分类库中的目标 URL 与分类信息的对应关系，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树；当有 URL 需要分类时，待分类 URL 在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类 URL 的分类信息。本发明方法中，在建立分类树时，以目标 URL 字符的 ASCII 码的数值作为 Tree 中的节点数组 node[] 数组的下标来构建 URL 树，即采用树型结构存储目标 URL 分类库。这样，在进行 URL 分类时，只需要对待分类 URL 做一次逐一字符的访问，即可完成分类过程，提高了分类效率并实现了快速分类，进而实现了对万亿级的上网记录数据中用户访问的 URL 的快速分类。

[0020] 本发明的其它特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点可通过在说明书、权利要求书以及附图中所特别指出的结构来实现和获得。

附图说明

[0021] 附图用来提供对本发明技术方案的进一步理解，并且构成说明书的一部分，与本申请的实施例一起用于解释本发明的技术方案，并不构成对本发明技术方案的限制。

[0022] 图 1 为本发明对 URL 进行分类的方法的流程图；

[0023] 图 2 为本发明构建 URL 分类库的分类树的示意图；

[0024] 图 3 为本发明对 URL 进行分类的装置的组成结构示意图。

具体实施方式

[0025] 为使本发明的目的、技术方案和优点更加清楚明白，下文中将结合附图对本发明的实施例进行详细说明。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互任意组合。

[0026] 在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行。并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0027] 图 1 为本发明对 URL 进行分类的方法的流程图，如图 1 所示，包括：

[0028] 步骤 100：根据 URL 分类库中的目标 URL 与分类信息的对应关系，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树。

[0029] 本步骤的目的在于，在分类库中的目标 URL 与分类信息之间，按照目标 URL 中的字符，建立树形的对应关系。本步骤具体实现包括：

[0030] 首先，建立一个 URL 分类库，URL 分类库中的内容包括：一个或一个以上目标 URL，以及各目标 URL 对应的分类信息。比如：

[0031] 目标 URL 为“www.baidu.com”，对应的分类信息为“百度”；目标 URL 为“www.sina.com”，对应的分类信息为“新浪”，等等。

[0032] 然后，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树，具体包括：依次取出目标 URL 的字符，以取出的字符的 ASCII 码作为分类树（Tree）中的节点数组 node[] 的下标，从根节点开始访问 node[ASCII] 的节点，直至目标 URL 的最后一个字符，则将分类信息存储到 node[ASCII] 的目标（obj）域中；如果未到目标 URL 的最后一个字符，接着取出下一个字符，对 *tree 指向的子树重复上述访问过程，直至目标 URL 的最后一个字符，如图 2 所示。其中，分类树包括两类数据结构即 tree 和 node[]，具体地，在 tree 节点中包含一个 node[] 数组，node[] 数组中的每一个节点中有两个域，一个域是 *tree 指向本节点的子树的指针，子树结构和根节点相同；另一个域是 obj，用于存储 URL 的分类信息。按照本步骤建立出的分类树，本发明分类表中的目标 URL 中的每一个字符，对应分类树中的一层即子树。

[0033] 以目标 URL 为“www.baidu.com”，对应的分类信息为“百度”；目标 URL 为“www.sina.com”，对应的分类信息为“新浪”；目标 URL 为“www.sohu.com”，对应的分类信息为“搜狐”为例，通过本步骤按照目标 URL 中的字符，建立树形的对应关系是：

[0034] 第一层子树对应 URL 中的第一个 w，第二层子树对应 URL 中的第二个 w，第三层子树对应 URL 中的第三个 w，这三层子树是上述两个 URL 的公共的子树；第四层子树有两个分支，即对应 b 的第四层子树一和对应 s 的公共的第四层子树二；第五层子树有三个分支，即对应 a 的第五层子树一、对应 i 的第五层子树二，以及对应 o 的第五层子树三，以此类推，对于目标 URL 为“www.baidu.com”的最后一层子树对应 m，且其目标 obj 域的值为“百度”，对于目标 URL 为“www.sina.com”的最后一层子树对应 m，且其目标 obj 域的值为“新浪”，对于目标 URL 为“www.sohu.com”的最后一层子树对应 m，且其目标 obj 域的值为“搜狐”。

[0035] 从本步骤的具体实现可以清楚地看到，本发明以目标 URL 中的每一个字符的 ASCII 码作为 Tree 中的 node[] 数组的下标来构建 URL 分类库的 URL 树。按照本步骤对分类库中的每一个目标 URL 进行上述处理后，将会构建出一个 URL 分类库的分类树。本步骤强调的是，将 URL 分类库建立为一个树结构，而对于树的建立属于本领域技术人员的惯用技术手段，并不用于限定本发明的保护范围，这里不再赘述。

[0036] 需要说明的是，步骤 100 并不是每次分类都要执行的，而是预先通过步骤 100 所述的方法建立了一个目标 URL 分类树。需要进行 URL 分类时，按照字符在建立好的目标 URL 分类树中进行查找即可。

[0037] 需要说明的是，如果目标 URL 分类树需要扩充，也只需按照步骤 100 所述的方法进行添加即可。

[0038] 步骤 101：当有 URL 需要分类时，待分类 URL 在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类 URL 的分类信息。

[0039] 本步骤具体包括：依次取出待分类 URL 中一个字符，以该字符的 ASCII 码作为分类树中的 node[] 数组的下标，访问分类树的 node[ASCII] 的节点，如果其对应的 obj 域中已存储有分类信息，则确定 obj 域中的分类信息为待分类 URL 的分类信息并返回，即获得了待

分类 URL 的分类信息；

[0040] 如果 obj 域为空，并且已经访问到待分类 URL 的最后一个字符，那么，分类处理结束，同时返回分类信息为空；

[0041] 如果 obj 域为空，但是未访问到待分类 URL 的最后一个字符，那么，对 *tree 指向的子树继续进行上述访问。

[0042] 还以目标 URL 为“www.baidu.com”，对应的分类信息为“百度”；目标 URL 为“www.sina.com”，对应的分类信息为“新浪”，目标 URL 为“www.sohu.com”，对应的分类信息为“搜狐”为例，而且已通过步骤 100 建立了目标 URL 分类树。假设，此时待分类 URL 为“www.sohu.com”，那么，按照步骤 101，确定待分类 URL 的分类信息具体包括：

[0043] 逐一取出 www.sohu.com 中的字符，在木匾 URL 分类树中一层一层地匹配，按照步骤 100 中的实施例中建立的目标 URL 分类树，匹配会经过建立好的第一层子树、第二层子树、第三层子树、公共第四层子树二、第五层子树三，直至最后一层子树，并获得对应的目标 obj 域的值为“搜狐”，这样既可的值待分类 URL 为 www.sohu.com 对应的分类信息为“搜狐”。

[0044] 本发明方法中，在建立分类树时，以目标 URL 中的每一个字符的 ASCII 码作为 Tree 中的 node[] 数组的下标来构建 URL 树，即采用树型结构存储目标 URL 分类库。这样，在进行 URL 分类时，只需要对待分类 URL 做一次逐一字符的访问，即可完成分类过程，提高了分类效率并实现了快速分类，进而实现了对万亿级的上网记录数据中用户访问的 URL 的快速分类。

[0045] 图 3 为本发明对 URL 进行分类的装置的组成结构示意图，如图 3 所示，至少包括构建模块、分类模块，其中，

[0046] 构建模块，用于根据 URL 分类库中的目标 URL 与分类信息的对应关系，对每一个目标 URL 按照字符进行处理，构建 URL 分类库的分类树；

[0047] 分类模块，用于当有 URL 需要分类时，待分类 URL 在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类 URL 的分类信息。

[0048] 虽然本发明所揭露的实施方式如上，但所述的内容仅为便于理解本发明而采用的实施方式，并非用以限定本发明。任何本发明所属领域内的技术人员，在不脱离本发明所揭露的精神和范围的前提下，可以在实施的形式及细节上进行任何的修改与变化，但本发明的专利保护范围，仍须以所附的权利要求书所界定的范围为准。

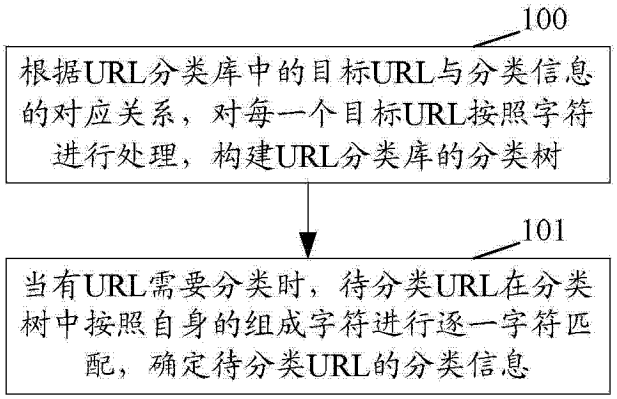

100
根据URL分类库中的目标URL与分类信息的对应关系，对每一个目标URL按照字符进行处理，构建URL分类库的分类树
101
当有URL需要分类时，待分类URL在分类树中按照自身的组成字符进行逐一字符匹配，确定待分类URL的分类信息

图 1

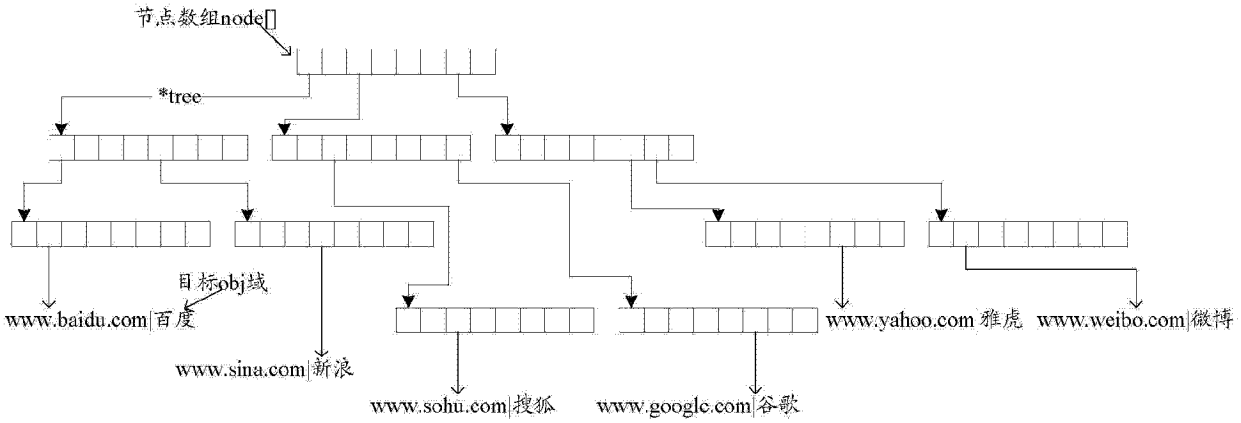

节点数组node[]
*tree
目标obj域
www.baidu.com|百度
www.sina.com|新浪
www.sohu.com|搜狐
www.google.com|谷歌
www.yahoo.com 雅虎
www.weibo.com|微博

图 2

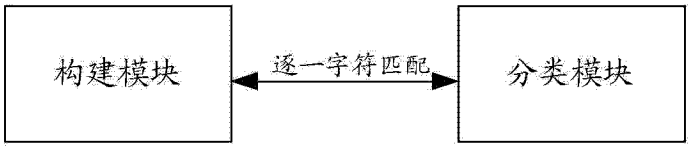

构建模块
逐一字符匹配
分类模块

图 3