(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 1786965 B
(45)授权公告日 2010.05.26

(21)申请号 200510132372.0

(22)申请日 2005.12.21

(73)专利权人 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
专利权人 北京北大方正技术研究院有限公司
北京大学

(72)发明人 舒文兵 吴於茜 肖建国

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 田明 王达佐

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

(56)对比文件
CN 1536483 A,2004.10.13, 全文.
CN 1435780 A,2003.08.13, 全文.
孙承志,关毅. 基于统计的网页正文信息抽取方法的研究. 中文信息学报 18 5.2004,18(5),19-20.

审查员 李倩

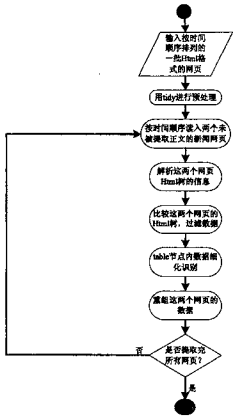
权利要求书 2 页 说明书 7 页 附图 2 页

(54)发明名称

一种新闻网页正文信息的提取方法

(57)摘要

本发明涉及一种新闻网页正文信息的提取方法,属于网页信息分析处理技术领域。现有技术中,通常采用包装器来抽取网页中感兴趣的数据,而包装器是根据一定的信息模式识别知识从特定的信息源中按固定规则抽取相关内容,并以特定形式加以表示的,包装器所需的信息模式识别知识的获取是一个费时费力且需要较高智能的工作。本发明所述的方法以堆栈数据结构,把网页数据的层次结构信息转化为用向量表达,构建和解析 Html 树,然后将 Html 树的各层次数据做对比,进行数据过滤,细化、识别,和数据重组,提取所需的数据信息。采用本发明所述的方法,适用于长期从一固定站点抓取由模版生成的新闻网页中的新闻信息,速度快,准确性高。



CN 1786965 B

1. 一种新闻网页正文信息的提取方法，包括以下步骤：

(1) 对网页进行规范化预处理，使之符合 Html 语言标准，然后依据 Html 语言中的 <table> 和 标记，解析所有新闻网页的 Html 数据，得到 Html 树；

(2) 将由相同模版生成的 Html 树的各层次数据做对比，把坐标相同，所包含的有效信息也相同的 table 节点或 div 节点剔除；

(3) 将 Html 树中各层次的 table 节点内的数据进行细化识别，区分出标题信息和内容信息；

(4) 重组处理后的 Html 树中各个节点内的数据，提取所需的数据信息。

2. 如权利要求 1 所述的一种新闻网页正文信息的提取方法，其特征是：步骤 (1) 中解析所有新闻网页的 Html 数据，构建 Html 树时，采用如下方法：

1) 初始化一个空数组 T，用于保存 Html 树中的各个 table 结构体；

所述的 table 结构体用来表示 table 节点，形式如下：

struct Table

{

此 table 节点的坐标；

此 table 节点所包含的信息；

}；

上述 table 节点的坐标即 table 节点在整个 Html 树中的位置用一个向量来表示，即每一个 table 节点均与一个向量 $v = (n_1, n_2, n_3, \ldots, n_k)$ 相对应，v 的第 i 个分量 n_i 的含义是 Html 树中第 i 层的第 n_i 个节点；

2) 初始化一个栈，设从栈底到栈顶元素依次标记为 a[0]，a[1]，a[2]，a[3]，...，且 0 = a[0] = a[1] = a[2] = a[3] = ... ；并设置一个栈元素指针 p，指向栈顶元素，由于初始时栈内没有元素，可假设 p 指向一个虚拟元素 a[-1]；

3) 扫描待处理的 Html 文档，如果遇到 <table> 标记，即遇到一个新的 table 节点时，将栈元素指针 p 向上移一格，然后将栈元素指针 p 指向的元素的值加 1，设此时栈元素指针 p 指向的栈元素为 a[k]，那么 table 节点 A 的坐标就是从栈底元素 a[0] 到 a[k] 所构成的序列，即向量 (a[0]，a[1]，a[2]，...，a[k])，由此得到 table 节点 A 的坐标；

4) 如果遇到 </table> 节点，即一个 table 节点结束时，将栈元素指针 p 向下移一格，此时构造一个新 table 结构体，把当前 table 节点的坐标和所包含的信息存于此 table 结构体中，然后把此结构体添加到数组 T 的末尾位置；

5) 如果遇到其它字符，设栈元素指针 p 指向的栈元素为 a[k]，那么当前正在扫描的 table 节点的坐标就是从栈底元素 a[0] 到 a[k] 所构成的序列，即向量 (a[0]，a[1]，a[2]，...，a[k])，把此字符添加到坐标为 (a[0]，a[1]，a[2]，...，a[k]) 的 table 节点所包含的信息里。

6) 如果还没有扫描到 Html 文档末尾，则继续扫描，转入第 3) 步，否则结束，返回保存了 Html 树层次信息的数组 T。

3. 如权利要求 1、2 所述的一种新闻网页正文信息的提取方法，其特征是：步骤 (2) 中过滤数据，删除不需要的数据信息时，采用如下的方法：

设 C 和 D 是由相同模板生成的两个发布时间相邻的新闻网页，

1) 经过步骤 (1) 后得到网页 C 的结构体数组为 T_1 ;

2) 经过步骤 (1) 后得到网页 D 的结构体数组为 T_2 ;

3) 遍历 T_1 中每个 table 结构体，对 T_1 中每个结构体，设为 S_1 并进行如下操作 :

a) 遍历 T_2，在 T_2 中找到与 S_1 坐标值相同的结构体，设为 S_2 ;

b) 判断 S_1 包含的信息是否与 S_2 包含的信息中相同，链接文字除外，如果相同，则在 T_1 中删除 S_1，在 T_2 中删除 S_2。

4. 如权利要求 1、2 所述的一种新闻网页正文信息的提取方法，其特征是 :步骤 (3) 中将 Html 树中各层次的 table 节点内的数据进行细化识别，区分出标题信息和内容信息时，采用如下的方法 :

1) 对 table 节点内的结构体，判断该结构体信息中有没有标题元素 ;

2) 如果该结构体的标题元素多于 1 个，那么取第一个作为本结构体的标题，如果没有标题元素，说明本 table 结构体标题为空。

5. 如权利要求 3 所述的一种新闻网页正文信息的提取方法，其特征是 :步骤 (3) 中将 Html 树中各层次的 table 节点内的数据进行细化识别，区分出标题信息和内容信息时，采用如下的方法 :

1) 对 table 节点内的结构体，判断该结构体信息中有没有标题元素 ;

2) 如果该结构体的标题元素多于 1 个，那么取第一个作为本结构体的标题，如果没有标题元素，说明本 table 结构体标题为空。

6. 如权利要求 1 所述的一种新闻网页正文信息的提取方法，其特征是 :

步骤 (4) 中重组处理后的 Html 树中各个节点内的数据时，采用如下方法 :

1) 初始化一个空字符串 S ;

2) 遍历 table 结构体数组 T 中每个 table 结构体，把每个 table 结构体包含的信息添加到 S 中 ;

3) 删除 S 中的 Html 标记，删除 Html 标记后的 S_1 即为所需提取的新闻网页的正文内容。

7. 如权利要求 5 所述的一种新闻网页正文信息的提取方法，其特征是 :

步骤 (4) 中重组处理后的 Html 树中各个节点内的数据时，采用如下方法 :

1) 初始化一个空字符串 S ;

2) 遍历 table 结构体数组 T 中每个 table 结构体，把每个 table 结构体包含的信息添加到 S 中 ;

3) 删除 S 中的 Html 标记，删除 Html 标记后的 S_1 即为所需提取的新闻网页的正文内容。

一种新闻网页正文信息的提取方法

技术领域

[0001] 本发明属于网页信息分析处理技术领域，具体涉及一种新闻网页正文信息的提取方法。

背景技术

[0002] 互联网的飞速发展使网络即 Web 上的信息量每天都以惊人的速度增加，许多企业常常需要各种信息，通常会从网络上大规模搜集信息，因而海量信息的采集成为每个企业都要关心的问题。因为目前的信息处理技术都是针对纯文本格式的内容的，而 Web 上的信息主要是以静态 Html 形式存在的，如何把 Web 上采集的 Html 形式的信息转换成有利用价值的文本格式的信息，方便后续的信息处理，成为亟待解决的技术问题。

[0003] Web 上信息的重要表现形式就是新闻，每天，各大门户网站都会新增大量的各种新闻，如何采集这些新闻信息，就成为 Web 信息采集的重要问题。通常，一个新闻网页中，除了包括主要新闻的内容（通常称之为网页正文）外，还包括大量的与新闻内容无关的信息（比如广告、网页导航信息、版权信息等，为方便，下面将这些与新闻无关的信息统称为广告），如何从新闻网页中准确提取新闻，去除与新闻信息无关的广告等其他信息，并最大程度地避免由网页改版所带来原网页抓取方法失效的问题也正是目前需要解决的技术问题。

[0004] 目前网络上绝大部分新闻信息都来自于重要的门户网站，而这些网站的新闻网页往往都是由模版后台生成的，其风格和样式在某段时间内是相同的。目前互联网上的网页绝大部分是用 HTML 语言编写的。Html 语言提供的标记主要是用来控制网页内容的显示格式的，如 <table>，<tr>，<td>，<th> 是用来绘制表格的；<li>，<ol>， 是用来表示列表的，这些标记的使用没有什么规律，网页设计人员可以随便设计。但是不同种类的数据一般是放在不同的显示单元中的。经过实际分析各大网站的新闻网页，结果显示需要提取的新闻类网页中的正文信息绝大部分是存在于 Html 标记“<table>”和“ ”之中的。

[0005] 传统的网页数据提取方法，是通过包装器来提取网页中感兴趣的数据的。包装器根据信息模式识别知识从固定的信息源中抽取相关内容，并以固定形式加以表示。早期，最简单的包装器是通过人工分析欲提取信息的目标网页的结构特征，然后编写有针对性的软件来实现的，这种方法人工干预大，代价很高；后来又引进了一些模式识别的算法，但至今，包装器所需的信息模式识别知识的获取还是一个费时费力且需要较高智能的工作，因此，目前网页数据抽取研究工作的热点之一就是探索简易的获得构造一个包装器所需规则的有效方法。目前利用包装器的系统有 TSIMMIS 系统，XWRAP 系统等。

[0006] TSIMMIS 系统中的包装器需要人工来书写数据抽取规则。规则被放在专门的文件中，规则的形式是 [variables，source，pattern]。其中，variables 保存抽取结果，source 保存输入，pattern 保存了数据在 source 中的模式信息；variables 可以用作后面的规则的 source，文件中最后一个规则执行结束后，variables 中保存了最后的抽取结果。这种需要人工书写规则的方法，费时、费力，而且容易出错，不易维护。

[0007] XWRAP 系统中的包装器采用了半自动化的方法来获取数据抽取规则。它提供了友

好的人机交互界面，用户可以根据系统的引导来完成数据抽取规则的编写，最终，系统生成一个针对特定数据源的用 java 语言编写的包装器。在进行数据抽取之前，XWRAP 系统会对网页进行检查，修正其中的不符合规范的语法错误和标记，并把网页解析成一棵树。

[0008] 上面介绍的几种包装器都是针对某一个固定网页架构来按固定的规则或模式来抽取数据，有比较大的局限性。由于网页结构的复杂性及不规范性，并且一旦网页改版，网页架构改变，原先适用的包装器就不能再适用了，这是包装器的严重缺点，即一个包装器的实现一般只能针对一个信息源。如上所述，目前的网页数据抽取工具，都需要针对特定的数据源来编写对应的包装器或抽取规则。所以，如果信息是来自很多信息源，就需要很多包装器，这样包装器的生成及维护就成了一种复杂的工作。对于网络上大量存在的结构风格各不相同的新闻类网页的正文信息抽取这样的任务来说，使用包装器的代价是很大的。

[0009] 《基于统计的网页正文信息抽取方法的研究》（中文信息学报，第 18 卷，第 5 期）公开了一种新闻类网页正文抽取方法。该方法根据新闻类网页的正文大部分存储在 table 中的特点，首先对网页进行规范化预处理，然后根据 HTML 标记把网页表示成一棵树，再找到 HTML 树中包含的所有 table 节点，去掉 HTML 标记，得到不含 HTML 标记的字符串。如果得到的字符串中所含有的中文字符的数量大于预先设定的阈值，则把该 table 节点作为候选。最后，对每个 table 节点按照由它得到的字符串的长度进行降序排序，排在前面的 table 节点便是需要抽取的正文信息。该方法具有以下不足之处。

[0010] (1) 采用该方法抽取的正文信息不完整：因为新闻类网页的正文信息不仅存在与 table 中，而且也存在与 div 中；此外，新闻的信息不仅包括正文信息，而且也包括标题信息，对于标题信息的抽取，该方法并未涉及。

[0011] (2) 采用该方法抽取的正文信息不够准确，效率也不高：因为选取候选 table 节点的方法中阈值的设定很难把握，阈值的大小对于正文信息的抽取影响很大，因此如果阈值设定不合适，则抽取的正文信息将很不准确；即使选取了合适的阈值，仅仅通过将字符串中含有的中文字符的数量大于阈值的 table 节点就作为候选，这种提取正文信息的方法是不够准确的。此外，阈值的设定需要通过大量的试验，从而也影响了提取的效率。

发明内容

[0012] 针对现有技术中存在的缺陷，本发明的目的是提供一种新闻网页正文信息的提取方法，该方法对于那些新闻网页的正文均存在于“<table>”或“ ”之中的数据信息而言，能够实现对由各种不同结构的模版生成的一系列新闻网页的内容的自动提取，能够提高网页信息提取的效率、完整性和准确率。

[0013] 为达到以上目的，本发明采用的技术方案是：一种新闻网页正文信息的提取方法，包括以下步骤：

[0014] (1) 对网页进行规范化预处理，使之符合 Html 语言标准，然后依据 Html 语言中的 <table> 和 标记，解析所有新闻网页的 Html 数据，得到 Html 树；

[0015] (2) 将由从同一站点抓取的由模版生成的并且时间相邻的两个网页的 Html 树的各层次数据做对比，把坐标相同，所包含信息也相同的 table 节点或 div 节点剔除；

[0016] (3) 将 Html 树中各层次的 table 节点内的数据进行细化识别，区分出标题信息和内容信息；

[0017] (4) 重组处理后的 Html 树中各个节点内的数据，提取所需的数据信息。

[0018] 更进一步，为使本发明具有更好的效果，步骤 (1) 中解析所有新闻网页的 Html 数据，构建 Html 树时，采用如下方法：

[0019] 1) 初始化一个空数组 T，用于保存 Html 树中的各个 table 结构体；

[0020] 所述的 table 结构体用来表示 table 节点，形式如下：

[0021] struct Table

[0022] {

[0023] 此 table 节点的坐标；

[0024] 此 table 节点所包含的信息；

[0025] }；

[0026] 上述 table 节点的坐标即 table 节点在整个 Html 树中的位置用一个向量来表示，即每一个 table 节点均与一个向量 $v = (n_1, n_2, n_3, \ldots, n_k)$ 相对应，v 的第 i 个分量 n_i 的含义是 Html 树中第 i 层的第 n_i 个节点；

[0027] 2) 初始化一个栈，设从栈底到栈顶元素依次标记为 a[0]，a[1]，a[2]，a[3]，...，且 0 = a[0] = a[1] = a[2] = a[3] = ... ；并设置一个栈元素指针 p，指向栈顶元素，由于初始时栈内没有元素，可假设 p 指向一个虚拟元素 a[-1]；

[0028] 3) 扫描待处理的 Html 文档，如果遇到 <table> 标记，即遇到一个新的 table 节点时，将栈元素指针 p 向上移一格，然后将栈元素指针 p 指向的元素的值加 1，设此时栈元素指针 p 指向的栈元素为 a[k]，那么 table 节点 A 的坐标就是从栈底元素 a[0] 到 a[k] 所构成的序列，即向量 (a[0]，a[1]，a[2]，...，a[k])，由此得到 table 节点 A 的坐标；

[0029] 4) 如果遇到 </table> 节点，即一个 table 节点结束时，将栈元素指针 p 向下移一格，此时构造一个新 table 结构体，把当前 table 节点的坐标和所包含的信息存于此 table 结构体中，然后把此结构体添加到数组 T 的末尾位置；

[0030] 5) 如果遇到其它字符，设栈元素指针 p 指向的栈元素为 a[k]，那么当前正在扫描的 table 节点的坐标就是从栈底元素 a[0] 到 a[k] 所构成的序列，即向量 (a[0]，a[1]，a[2]，...，a[k])，把此字符添加到坐标为 (a[0]，a[1]，a[2]，...，a[k]) 的 table 节点所包含的信息里。

[0031] 6) 如果还没有扫描到 Html 文档末尾，则继续扫描，转入第 3) 步，否则结束，返回保存了 Html 树层次信息的数组 T。

[0032] 更进一步，为使本发明具有更好的效果，步骤 (2) 中过滤数据，删除不需要的数据信息时，采用如下的方法：

[0033] 设 C 和 D 是由相同模板生成的两个发布时间相邻的新闻网页，

[0034] 1) 经过步骤 (1) 后得到网页 C 的结构体数组为 T_1；

[0035] 2) 经过步骤 (1) 后得到网页 D 的结构体数组为 T_2；

[0036] 3) 遍历 T_1 中每个 table 结构体，对 T_1 中每个结构体，设为 S_1 进行如下操作：

[0037] a) 遍历 T_2，在 T_2 中找到与 S_1 坐标值相同的结构体，设为 S_2；

[0038] b) 判断 S_1 包含的信息是否与 S_2 包含的信息中相同（链接文字除外），则在 T_1 中删除 S_1，在 T_2 中删除 S_2。

[0039] 更进一步，为使本发明具有更好的效果，步骤 (3) 中将 Html 树中各层次的 table

节点内的数据进行细化识别，区分出标题信息和内容信息时，采用如下的方法：

[0040] 1) 对 table 节点内的结构体，判断该结构体信息中有没有标题元素；

[0041] 2) 如果该结构体的标题元素多于 1 个，那么取第一个作为本结构体的标题，如果没有标题元素，说明本 table 结构体标题为空。

[0042] 更进一步，为使本发明具有更好的效果，步骤 (4) 中重组处理后的 Html 树中各个节点内的数据时，采用如下方法：

[0043] 1) 初始化一个空字符串 S；

[0044] 2) 遍历 table 结构体数组 T 中每个 table 结构体，把每个 table 结构体包含的信息添加到 S 中；

[0045] 3) 删除 S 中的 Html 标记，删除 Html 标记后的 S_1 即为所需提取的新闻网页的正文内容。

[0046] 本发明的效果在于：采用本发明所述的方法，能够处理从通过模板来生成网页的新闻站点的信息采集任务，能够迅速自动提取目标新闻网页的正文内容，即使网页改版，也不需要重新编写程序，人工干预大大降低，从而极大地提高了网页信息提取的效率、完整性和准确率。

[0047] 本发明之所以具有以上效果，是由于本发明所述的方法采用了一种的新的解析 Html 树的方法，可以高效准确地知道 Html 中每个 table 节点的坐标和所包含的信息；如果网页改版，也能迅速的解析新模版的树形结构信息，然后比较由新模版生成的网页，仍能准确抽取新闻正文信息。

附图说明

[0048] 图 1 是本发明的流程图；

[0049] 图 2 是本发明具体实施方式中解析 Html 树的流程图。

具体实施方式

[0050] 下面结合实施例及附图，进一步阐明本发明所述方法。

[0051] 以从新浪新闻的体育频道抓取下来的按时间顺序排列好的 1000 个新闻网页中提取正文信息为例，如图 1 所示，一种新闻网页正文信息的提取方法，包括以下步骤：

[0052] (1) 对 1000 个网页用第三方网页净化工具（比如可以使用 tidy 工具），进行规范化预处理，使之符合 Html 语言标准，然后依据 Html 语言中的 <table> 和 标记，解析所有新闻网页的 Html 数据，得到 Html 树；

[0053] 解析所有新闻网页的 Html 数据，构建 Html 树时，采用如下方法：

[0054] 由于在本发明中，Html 标记 <table> 和 作用是相同，因此本发明以 <table> 为例来阐述， 的情形完全类同于 <table>。以如下的 Html 片段为例（如上所述，只标出所关心的 <table> 节点，// 是注释），阐明本发明所涉及的述语：

[0055] <table>// 第一个 <table> 节点开始

[0056] Text1

[0057] <table>// 第二个 <table> 节点开始

[0058] Text2

[0059] <table>// 第三个 <table> 节点开始

[0060] Text3

[0061] </table>// 第三个 <table> 节点结束

[0062] Text4

[0063] </table>// 第二个 <table> 节点结束

[0064] <table>// 第四个 <table> 节点开始

[0065] Text4

[0066] </table>// 第四个 <table> 节点结束

[0067] </table>// 第一个 <table> 节点结束

[0068] 将每个 table 开始符（以 <table> 为标志）和结束符（以 </table> 为标志）之间的 Html 内容作为一个 table 节点，那么从上面的片段可以看出，每个 table 节点里面还可以嵌套其它 table 节点，比如第三个 table 节点里面就嵌套在第二个 table 节点里面。

[0069] 如果一个 table 节点 A 嵌套在另一个 table 节点 B 里面，那么 A 叫做 B 的子节点，B 叫做 A 的父节点。

[0070] 将位于一个 table 节点 A 开始符和结束符之间，且不位于此节点任何子节点开始符和结束符之间的 Html 内容叫做 A 包含的信息。

[0071] 将一个 table 节点所对应的向量称为此 table 节点在 Html 树中的坐标。

[0072] 上述 Html 片段中，第二个 table 节点包含的信息为 Text2 和 Text4，第三个 table 节点包含的信息为 Text3。

[0073] 用直观的形式表达 Html 树状层次的嵌套信息，即利用一个向量来表示所关心的 table 节点在整个 Html 树中的位置。每一个 table 节点均与一个向量 $v = (n_1, n_2, n_3, \ldots, n_k)$ 相对应，v 的第 i 个分量 n_i 的含义是 Html 树中第 i 层的第 n_i 个节点。如果一个 table 节点对应向量是 (1,2,3)，那么就说明此 table 节点是 Html 树第一层第一个 table 节点的第二个子节点的第三个子节点。

[0074] 上述 Html 片段中第三个和第四个 table 节点的坐标分别为 (1,1,1) 和 (1,2)。

[0075] 采用结构体的形式来表示 table 节点，形式如下：

[0076] struct Table

[0077] {

[0078] 此 table 节点的坐标；

[0079] 此 table 节点所包含的信息；

[0080] }；

[0081] 将 Html 文档转换为各个 table 节点的结构体时，采用如下方法：

[0082] 1) 初始化一个空数组 T，用于保存各个 table 结构体；

[0083] 2) 初始化一个栈，设从栈底到栈顶元素依次标记为 a[0]，a[1]，a[2]，a[3]，...，且 0 = a[0] = a[1] = a[2] = a[3] = ... ；并设置一个栈元素指针 p，指向栈顶元素。由于初始时栈内没有元素，可假设 p 指向一个虚拟元素 a[-1]；

[0084] 3) 扫描待处理的 Html 文档，如果遇到 <table> 标记，即遇到一个新的 table 节点时，将栈元素指针 p 向上移一格，然后将栈元素指针 p 指向的元素的值加 1，设此时栈元素指针 p 指向的栈元素为 a[k]，那么 table 节点 A 的坐标就是从栈底元素 a[0] 到 a[k] 所构成

的序列，即向量 (a[0]，a[1]，a[2]，...，a[k])，由此得到 table 节点 A 的坐标；

[0085] 4) 如果遇到 </table> 节点，即一个 table 节点结束时，将栈元素指针 p 向下移一格，此时构造一个新 table 结构体，把当前 table 节点的坐标和所包含的信息存于此 table 结构体中，然后把此结构体添加到数组 T 的末尾位置；

[0086] 5) 如果遇到其它字符，设栈元素指针 p 指向的栈元素为 a[k]，那么当前正在扫描的 table 节点的坐标就是从栈底元素 a[0] 到 a[k] 所构成的序列，即向量 (a[0]，a[1]，a[2]，...，a[k])，把此字符添加到坐标为 (a[0]，a[1]，a[2]，...，a[k]) 的 table 节点所包含的信息里。

[0087] 6) 如果还没有扫描到 Html 文档末尾，则继续扫描，转入第 3) 步，否则结束，返回保存了 Html 树层次信息的数组 T。

[0088] (2) 将由相同模版生成的 Html 树的各层次数据做对比，过滤数据，删除不需要的数据信息；

[0089] 在本实施例中，首先将所有网页按时间顺序排序，设网页集合为 S，从网页集合 S 中取出时间相邻的两个网页 W_1，W_2；解析网页 W_1，W_2 的 Html 树，得到每个网页中 table 节点的坐标和其所包含的信息；比较 W_1，W_2 的 Html 树，过滤数据，删除不需要的信息，具体采用如下方法：

[0090] 1) 经过步骤 (1) 后得到网页 W_1 的结构体数组为 T_1；

[0091] 2) 经过步骤 (1) 后得到网页 W_2 的结构体数组为 T_2；

[0092] 3) 遍历 T_1 中每个 table 结构体，对 T_1 中每个结构体，设为 S_1 进行如下操作：

[0093] a) 遍历 T_2，在 T_2 中找到与 S_1 坐标值相同的结构体，设为 S_2；

[0094] b) 判断 S_1 包含的信息是否与 S_2 包含的信息中相同（链接文字除外），则在 T_1 中删除 S_1，在 T_2 中删除 S_2。

[0095] (3) 将 Html 树中各层次的 table 节点内的数据进行细化识别，区分出标题信息和内容信息；

[0096] 经过步骤 (2) 后，不需要的广告信息已经被删除，但是还需要对未被过滤的 table 结构体进行内容的细化识别，识别出标题信息和内容信息，通常新闻的标题一般都以大号黑体形式出现，这在 Html 中，是通过 <th>，<b>，，<h1>，<h2> 等标记实现的，这些元素可称之为标题元素。因此可以采取以下具体步骤，来实现 table 结构体内容的细化识别。

[0097] 1) 对 table 节点内的结构体，判断该结构体信息中有没有标题元素；

[0098] 2) 如果该结构体的标题元素多于 1 个，那么取第一个作为本结构体的标题，如果没有标题元素，说明本 table 结构体标题为空。

[0099] (4) 重组处理后的 Html 树中各个节点内的数据，提取所需的数据信息。

[0100] 由步骤 (1) 得到的 table 结构体数组 T 经过步骤 (2) 和步骤 (3) 的处理后，数组 T 里面的每个结构体的信息都已经被识别了，下面要做的就是把这些数组 T 里面的每个 table 结构体所包含的信息合并起来，可采用如下方法：

[0101] 1) 初始化一个空字符串 S；

[0102] 2) 遍历 table 结构体数组 T 中每个 table 结构体，把每个 table 结构体包含的信息添加到 S 中；

[0103] 3) 删除 S 中的 Html 标记，删除 Html 标记后的 S_1 即为所需提取的新闻网页的正文内容。

[0104] 试验效果证明，本方抓取新闻网页的准确率很高，在存在改版情况下，仍能达到 98％以上的准确率，而且时间效率高。

[0105] 本发明所述的方法并不限于具体实施方式中所述的实施例，本领域技术人员根据本发明的技术方案得出其他的实施方式，同样属于本发明的技术创新范围。

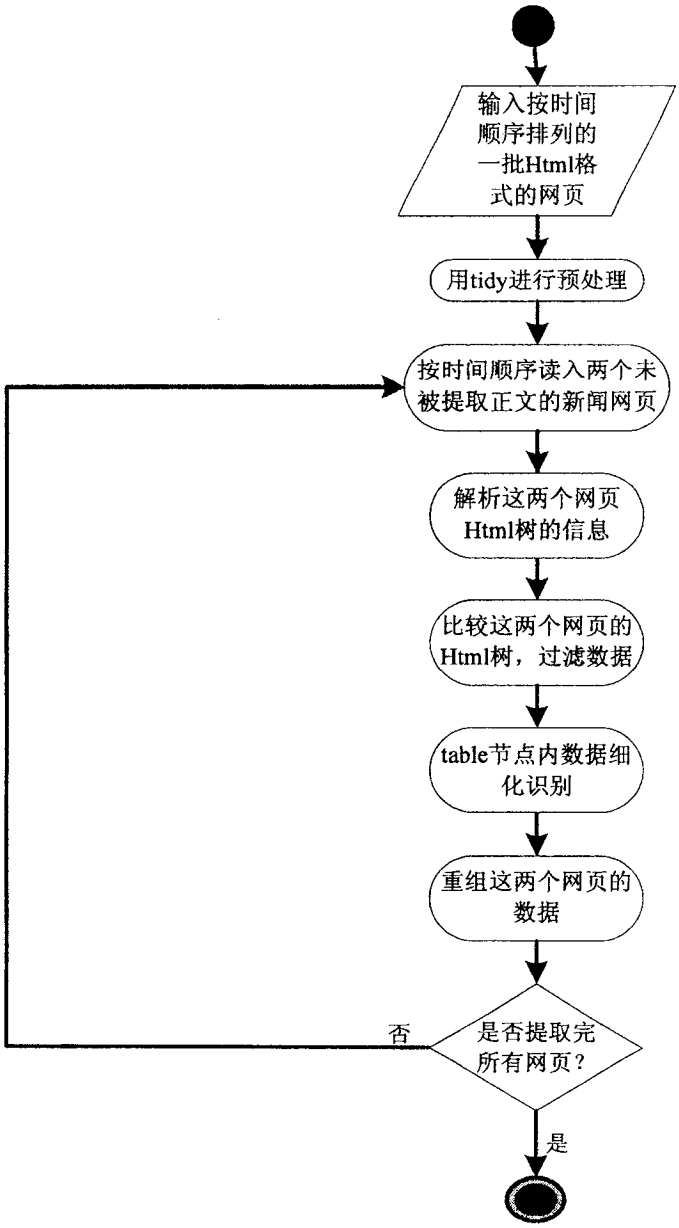

输入按时间顺序排列的一批Html格式的网页
用tidy进行预处理
按时间顺序读入两个未被提取正文的新闻网页
解析这两个网页Html树的信息
比较这两个网页的Html树，过滤数据
table节点内数据细化识别
重组这两个网页的数据
是否提取完所有网页？
否
是

图 1

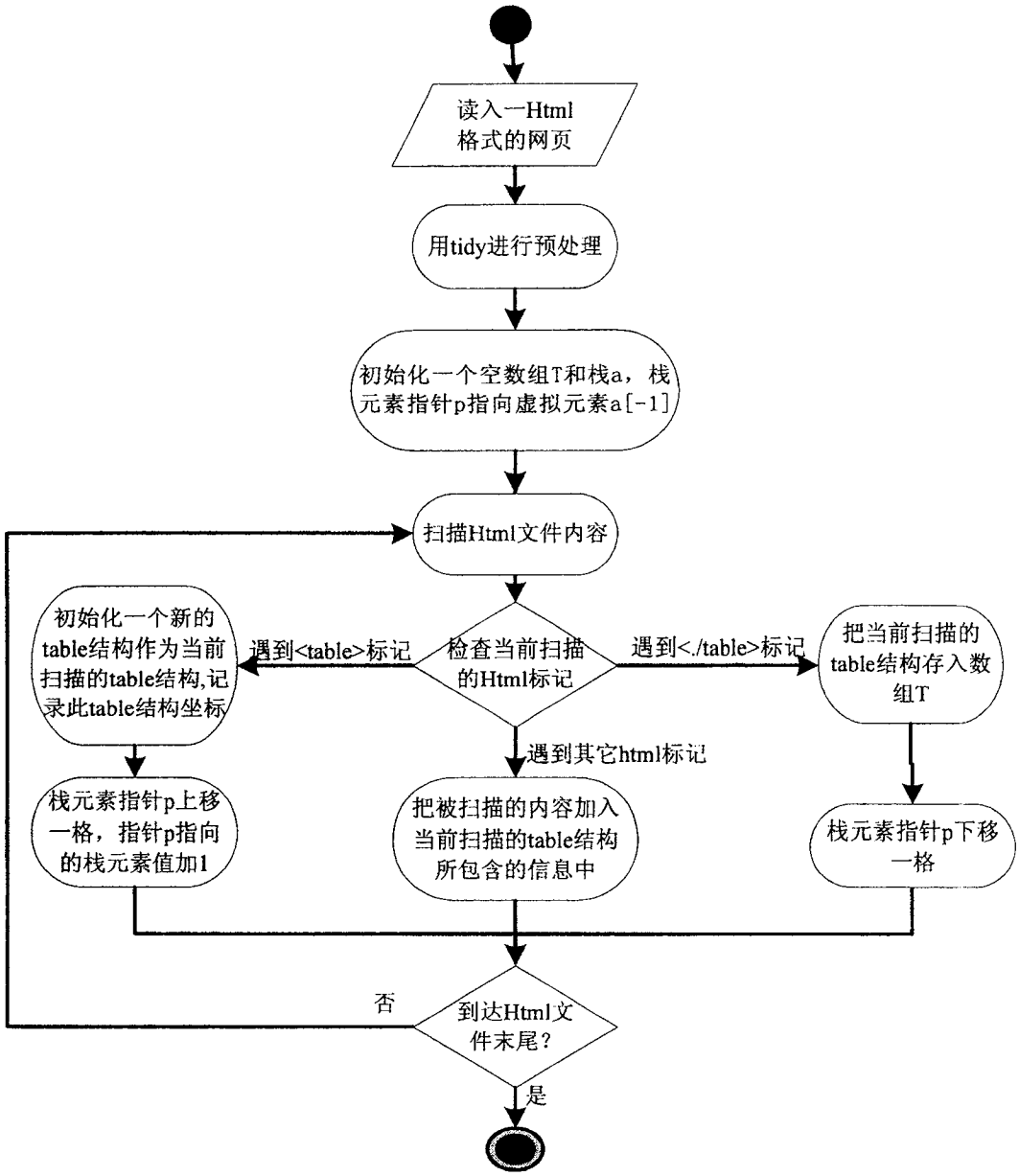

读入一Html格式的网页
用tidy进行预处理
初始化一个空数组T和栈a，栈元素指针p指向虚拟元素a[-1]
扫描Html文件内容
检查当前扫描的Html标记
遇到<table>标记
初始化一个新的table结构作为当前扫描的table结构,记录此table结构坐标
栈元素指针p上移一格，指针p指向的栈元素值加1
遇到</table>标记
把当前扫描的table结构存入数组T
栈元素指针p下移一格
遇到其它html标记
把被扫描的内容加入当前扫描的table结构所包含的信息中
到达Html文件末尾？
否
是

图 2