(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 109165373 B
(45)授权公告日 2022.04.22

(21)申请号 201811073868.9
(22)申请日 2018.09.14
(65)同一申请的已公布的文献号
申请公布号 CN 109165373 A
(43)申请公布日 2019.01.08
(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地信息产业基地创业路6号
(72)发明人 杨帆 戴超男
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王宝筠
(51)Int.Cl.
G06F 16/958(2019.01)
G06F 16/955(2019.01)

(56)对比文件
CN 102890681 A,2013.01.23
US 7765236 B2,2010.07.27
CN 103136358 A,2013.06.05
CN 104572934 A,2015.04.29

审查员 陈学元

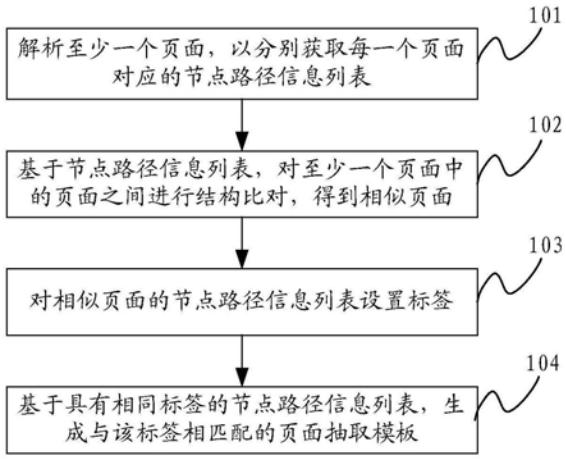
权利要求书3页 说明书12页 附图7页

(54)发明名称

一种数据处理方法及装置

(57)摘要

本申请公开了一种数据处理方法及装置,方法包括:解析至少一个页面,以分别获取每一个所述页面对应的节点路径信息列表;基于所述节点路径信息列表,对所述至少一个页面中的页面之间进行结构比对,得到相似页面;对所述相似页面的节点路径信息列表设置标签;基于具有相同标签的节点路径信息列表,生成与该标签相匹配的页面抽取模板。



CN 109165373 B

1.一种数据处理方法,包括:

解析页面,以分别获取每一个所述页面对应的节点路径信息列表;

基于所述节点路径信息列表,对所述页面中的页面之间进行结构比对,得到相似页面;

对所述相似页面的节点路径信息列表设置标签;

基于具有相同标签的节点路径信息列表,生成与该标签相匹配的页面抽取模板;

对所述相似页面的节点路径信息列表设置标签,包括:

根据所述相似页面的节点路径信息列表,确定待抽取的目标内容;

基于所述目标内容,设置所述相似页面的节点路径信息列表的标签,包括:对所述目标内容进行细化分类,根据分类结果设置所述相似页面的节点路径信息列表的标签;

所述基于具有相同标签的节点路径信息列表,生成与该标签相匹配的页面抽取模板包括:

对具有相同标签的节点路径信息列表,提取所述节点路径信息列表中与其标签存在关联关系的强特征词及其特征属性;

基于所述强特征词及其特征属性,生成所述页面抽取模板的特征词典;

解析所述强特征词的词义,以得到所述强特征词的同义词;

将所述同义词加入所述页面抽取模板的特征词典中;

对具有相同标签的节点路径信息列表进行合并;

基于合并后的节点路径信息列表,生成与所述标签相匹配的页面抽取模板;其中,生成的页面抽取模板中包含多个节点路径信息,所述节点路径信息用于信息抽取,且含有强特征词的节点路径信息的优先级高于其他节点路径信息的优先级。

2.根据权利要求1所述的方法,其特征在于,基于所述节点路径信息列表对所述页面中的页面之间进行结构比对,包括:

对所述页面中的两个页面的节点路径信息列表执行以下操作:

基于所述两个页面的节点路径信息列表,分别获得所述两个页面中第一页面和第二页面的树结构根节点和相应的子树;

基于所述两个页面的树结构根节点比对相同的判断,在所述第一页面的各子树中,确定分别与所述第二页面中各子树相似度最高的子树,以组成子树对;

获得所述子树对中两个子树的相似度值,并获得所述子树对中属于所述第一页面的子树的预设权重;

基于所述子树对的相似度值及所述预设权重,获得所述第一页面和所述第二页面之间的总相似度值;

基于所述总相似度值高于预设阈值的判断,确定所述第一页面和所述第二页面为相似页面。

3.根据权利要求1或2所述的方法,其特征在于,还包括:

获得所述页面中的页面内容;

对所述页面中的页面之间进行页面内容所属类别以及结构比对,得到相似页面。

4.根据权利要求1所述的方法,其特征在于,对具有相同标签的节点路径信息列表进行合并,包括:

在具有相同标签的节点路径信息列表中,对于提取到所述强特征词的节点路径信息列

表，设置预设的标记符号；

对所述节点路径信息列表之间依次按照节点路径信息列表中的节点次序进行一一比对，得到比对结果；

基于所述比对结果，将节点比对完全相同的节点路径信息列表合并为一个节点路径信息列表，将存在一个不同的节点的节点路径信息列表合并为一个节点路径信息列表，并将所述不同的节点以所述标记符号替代。

5.根据权利要求4所述的方法，其特征在于，在对所述节点路径信息列表之间依次按照节点路径信息列表中的节点次序进行一一比对之前，所述方法还包括：

利用所述标记符号，对所述节点路径信息列表进行简化；

具体为：

对设置有所述标记符号的节点路径信息列表的节点路径信息，至少保留所述节点路径信息中的树结构节点名称及用于填充所述强特征词的信息；

对没有设置所述标记符号的节点路径信息列表的节点路径信息，保留所述节点路径信息中的树结构节点名称。

6.根据权利要求1或2所述的方法，其特征在于，还包括：

响应于接收到的页面抽取请求，获得待抽取的目标页面以及所述目标页面的抽取标签；

在所述抽取标签对应的目标页面抽取模板中，优先使用含有强特征词的节点路径信息对所述目标页面进行页面数据抽取，得到抽取结果；

基于所述抽取结果中没有抽取到对应数据的判断，获得所述目标页面的节点路径信息列表并生成相应的页面抽取模板。

7.一种数据处理装置，包括：

页面解析单元，用于解析获得到的页面，以分别获取每一个所述页面对应的节点路径信息列表；

相似比对单元，用于基于所述节点路径信息列表，对所述页面中的页面之间进行结构比对，得到相似页面；

标签设置单元，用于对所述相似页面的节点路径信息列表设置标签；

模板生成单元，用于基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板；

对所述相似页面的节点路径信息列表设置标签，包括：

根据所述相似页面的节点路径信息列表，确定待抽取的目标内容；

基于所述目标内容，设置所述相似页面的节点路径信息列表的标签，包括：对所述目标内容进行细化分类，根据分类结果设置所述相似页面的节点路径信息列表的标签；

所述基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板包括：

对具有相同标签的节点路径信息列表，提取所述节点路径信息列表中与其标签存在关联关系的强特征词及其特征属性；

基于所述强特征词及其特征属性，生成所述页面抽取模板的特征词典；

解析所述强特征词的词义，以得到所述强特征词的同义词；

将所述同义词加入所述页面抽取模板的特征词典中；

对具有相同标签的节点路径信息列表进行合并；

基于合并后的节点路径信息列表，生成与所述标签相匹配的页面抽取模板；其中，生成的页面抽取模板中包含多个节点路径信息，所述节点路径信息用于信息抽取，且含有强特征词的节点路径信息的优先级高于其他节点路径信息的优先级。

一种数据处理方法及装置

技术领域

[0001] 本申请涉及页面抽取技术领域，尤其涉及一种数据处理方法及装置。

背景技术

[0002] 目前，在相同类型网站中进行结构化信息抽取时，通常采用构建模板的方式对网页信息进行抽取。

[0003] 但是现有的抽取模板配置中无法适用于不同网站网页的信息抽取，由此降低信息抽取的普遍适用性。

发明内容

[0004] 有鉴于此，本申请提供一种数据处理方法及装置，用以解决现有技术中的页面抽取模板无法对不同网站的网页进行信息抽取，导致信息抽取适用性较低的技术问题。

[0005] 本申请提供了一种数据处理方法，包括：

[0006] 解析至少一个页面，以分别获取每一个所述页面对应的节点路径信息列表；

[0007] 基于所述节点路径信息列表，对所述至少一个页面中的页面之间进行结构比对，得到相似页面；

[0008] 对所述相似页面的节点路径信息列表设置标签；

[0009] 基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板。

[0010] 上述方法，优选地，基于所述节点路径信息列表对所述至少一个页面中的页面之间进行结构比对，包括：

[0011] 对所述至少一个页面中的两个页面的节点路径信息列表执行以下操作：

[0012] 基于所述两个页面的节点路径信息列表，分别获得所述两个页面中第一页面和第二页面的树结构根节点和相应的子树；

[0013] 基于所述两个页面的树结构根节点比对相同的判断，在所述第一页面的各子树中，确定分别与所述第二页面中各子树相似度最高的子树，以组成子树对；

[0014] 获得所述子树对中两个子树的相似度值，并获得所述子树对中属于所述第一页面的子树的预设权重；

[0015] 基于所述子树对的相似度值及所述预设权重，获得所述第一页面和所述第二页面之间的总相似度值；

[0016] 基于所述总相似度值高于预设阈值的判断，确定所述第一页面和所述第二页面为相似页面。

[0017] 上述方法，优选地，还包括：

[0018] 获得所述至少一个页面中的页面内容；

[0019] 对所述至少一个页面中的页面之间进行页面内容所属类别以及结构比对，得到相似页面。

[0020] 上述方法，优选地，对所述相似页面的节点路径信息列表设置标签，包括：

[0021] 根据所述相似页面的节点路径信息列表，确定待抽取的目标内容；

[0022] 基于所述目标内容，设置所述相似页面的节点路径信息列表的标签。

[0023] 上述方法，优选地，还包括：

[0024] 对具有相同标签的节点路径信息列表，提取所述节点路径信息列表中与其标签存在关联关系的强特征词及其特征属性；

[0025] 基于所述强特征词及其特征属性，生成所述页面抽取模板的特征词典；

[0026] 解析所述强特征词的词义，以得到所述强特征词的同义词；

[0027] 将所述同义词加入所述页面抽取模板的特征词典中。

[0028] 上述方法，优选地，基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板，包括：

[0029] 对具有相同标签的节点路径信息列表进行合并；

[0030] 基于合并后的节点路径信息列表，生成与所述标签相匹配的页面抽取模板；其中，生成的页面抽取模板中包含多个节点路径信息，所述节点路径信息用于信息抽取，且含有强特征词的节点路径信息的优先级高于其他节点路径信息的优先级。

[0031] 上述方法，优选地，对具有相同标签的节点路径信息列表进行合并，包括：

[0032] 在具有相同标签的节点路径信息列表中，对于提取到所述强特征词的节点路径信息列表，设置预设的标记符号；

[0033] 对所述节点路径信息列表之间依次按照节点路径信息列表中的节点次序进行一一比对，得到比对结果；

[0034] 基于所述比对结果，将节点比对完全相同的节点路径信息列表合并为一个节点路径信息列表，将存在一个不同的节点的节点路径信息列表合并为一个节点路径信息列表，并将所述不同的节点以所述标记符号替代。

[0035] 上述方法，优选地，在对所述节点路径信息列表之间依次按照节点路径信息列表中的节点次序进行一一比对之前，所述方法还包括：

[0036] 利用所述标记符号，对所述节点路径信息列表进行简化；

[0037] 具体为：

[0038] 对设置有所述标记符号的节点路径信息列表的节点路径信息，至少保留所述节点路径信息中的树结构节点名称及用于填充所述强特征词的信息；

[0039] 对没有设置所述标记符号的节点路径信息列表的节点路径信息，保留所述节点路径信息中的树结构节点名称。

[0040] 上述方法，优选地，还包括：

[0041] 响应于接收到的页面抽取请求，获得待抽取的目标页面以及所述目标页面的抽取标签；

[0042] 在所述抽取标签对应的目标页面抽取模板中，优先使用含有强特征词的节点路径信息对所述目标页面进行页面数据抽取，得到抽取结果；

[0043] 基于所述抽取结果中没有抽取到对应数据的判断，获得所述目标页面的节点路径信息列表并生成相应的页面抽取模板。

[0044] 本申请还提供了一种数据处理装置，包括：

[0045] 页面解析单元，用于解析至少一个页面，以分别获取每一个所述页面对应的节点

路径信息列表；

[0046] 相似比对单元，用于基于所述节点路径信息列表，对所述至少一个页面中的页面之间进行结构比对，得到相似页面；

[0047] 标签设置单元，用于对所述相似页面的节点路径信息列表设置标签；

[0048] 模板生成单元，用于基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板。

[0049] 从上述技术方案可以看出，本申请公开的一种数据处理方法及装置，通过在解析出各种页面的节点路径信息列表之后，基于这些节点路径信息列表对这些不同的页面之间相似度分类，从而对相似页面进行标签设置进而生成该标签下的页面抽取模板，以便于对相应的页面进行信息抽取。可见，本申请中对于不同网站的网页通过相似度分类，得到结构内容相似的网页后，再生成相应的页面抽取模板，从而实现不同网站网页的数据抽取，该信息抽取的方案能够适用于对不同网站的页面信息抽取，不限于某一种结构或内容的网站页面，从而提高了信息抽取的普遍实用性。

附图说明

[0050] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0051] 图1为本申请实施例一提供的一种数据处理方法的流程图；

[0052] 图2为本申请实施例一的部分流程图；

[0053] 图3、图4及图5分别为本申请实施例的示例图；

[0054] 图6、图7、图8及图9分别为本申请实施例一的另一部分流程图；

[0055] 图10为本申请实施例二提供的一种数据处理装置的结构示意图；

[0056] 图11及图12分别为本申请实施例的另一示例图。

具体实施方式

[0057] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0058] 参考图1，为本申请实施例一提供的一种数据处理方法的实现流程图，该方法适用于对不同网站即不同结构或内容的网站的页面进行页面抽取模板的构建，进而用于对页面信息进行抽取。本实施例中的方法可以运行在具有计算能力的计算机或服务器中。

[0059] 具体的，本实施例中的方法可以包括有以下步骤：

[0060] 步骤101：解析至少一个页面，以分别获取每一个页面对应的节点路径信息列表。

[0061] 其中，本实施例中的页面可以包括有一个页面网站上的页面，也可以包括有多个页面网站上的页面，而不同页面网站上的页面在结构和内容上可以相同(或相似)或者不同。例如，购物网站、新闻网站及广告网站上的页面在结构和内容上都不相同。

[0062] 需要说明的是，本实施例中在对页面进行解析之前，肯定是获得到这些页面的，如从数据库中读取这些页面或者利用网络爬虫等工具在网站上实时爬取到这些页面。

[0063] 其中，本实施例中通过对页面解析，得到页面对应的节点路径信息列表可以理解为页面的xpath(XML Path Language)列表，其中，xpath列表使用路径表达式表征页面结构及结构内容。

[0064] 具体的，本实施例中可以通过对页面构建页面的树结构，进而解析树结构来生成xpath列表。例如，本实施例中使用第三方库(例如lxml)解析一个网站或多个网站上的超文本HTML(HyperText Markup Language)页面并构建文档对象模型DOM(Document Object Model)树，进而解析DOM树，以形成相应页面的完整xpath列表。

[0065] 步骤102：基于节点路径信息列表，对至少一个页面中的页面之间进行结构比对，得到相似页面。

[0066] 其中，本实施例中可以对至少一个页面中的任意两个页面分别进行结构比对，从而确定哪些页面属于相似页面，哪些页面并不是相似页面。

[0067] 需要说明的是，本实施例中的相似页面可以理解为：页面之间的相似度值大于一定阈值的页面才成为相似页面，而页面之间的相似度值可以为页面之间的结构相似度值和/或内容相似度值，也就是说，两个页面之间为相似页面是指：两个页面之间结构和/或页面相似。

[0068] 步骤103：对相似页面的节点路径信息列表设置标签。

[0069] 其中，本实施例中的标签可以为提取页面内容中的字符作为标签，如页面内容中的关键词作为标签；或者标签可以位于页面内容中的字符相关联的字符作为标签，如与页面内容中的关键词近似的词作为标签，等等。

[0070] 而该标签设置的含义至少在于：在同一标签下的节点路径信息列表所对应的页面为相似页面，不同标签下的节点路径信息列表所对应的页面不属于相似页面。

[0071] 具体的，本实施例中可以相似页面的节点路径信息列表按照其信息内容名称及类型进行细化分类，进而基于细化分类的结果来生成标签，并将标签设置给相似页面的节点路径信息列表中。

[0072] 步骤104：基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板。

[0073] 其中，本实施例中可以通过对具有相同标签的节点路径信息列表进行处理，来生成页面抽取模板，例如，选择这些相同标签的节点路径信息列表中与其他节点路径信息列表相似度最高的一个节点路径信息列表作为页面抽取模板；或者基于这些相同标签的节点路径信息列表进行组合或整合来生成该标签下相匹配的页面抽取模板，等等。

[0074] 由上述方案可知，本申请实施例一提供的一种数据处理方法，通过在解析出各种页面的节点路径信息列表之后，基于这些节点路径信息列表对这些不同的页面之间相似度分类，从而对相似页面进行标签设置进而生成该标签下的页面抽取模板，以便于对相应的页面进行信息抽取。可见，本申请中对于不同网站的网页通过相似度分类，得到结构内容相似的网页后，再生成相应的页面抽取模板，从而实现不同网站网页的数据抽取，该信息抽取的方案能够适用于对不同网站的页面信息抽取，不限于某一种结构或内容的网站页面，从而提高了信息抽取的普遍实用性。

[0075] 在一种实现方式中，图1中步骤102中具体可以对至少一个页面中的任两个页面的节点路径信息列表执行以下如图2中的操作，以便于确定两个页面之间是否为相似页面，从而解析出所有需要进行信息抽取的页面中的相似页面，如图2中所示：

[0076] 步骤201：基于两个页面的节点路径信息列表，分别获得两个页面中第一页面和第二页面的树结构根节点和相应的子树。

[0077] 如图3中所示，对第一页面和第二页面的节点路径信息列表所表征的树结构进行构建，得到第一页面树结构和第二页面树结构，进而获得第一页面和第二页面各自的树结构根节点和相应的子树。

[0078] 需要说明的是，在图3中只展示了两个页面的树结构示例，即仅展示了树结构中的各节点和节点位置，并没有将节点路径信息列表中的其他信息进行展示，但并不代表节点路径信息列表中除了树结构不包含其他信息。

[0079] 步骤202：比对两个页面的树结构根节点是否相同，如果相同，则执行步骤203。

[0080] 其中，本实施例中可以将两个页面的树结构根节点是否完全一致或近似一致进行比对，如在节点路径信息列表中表征为同一根目录文件夹名称，如果两个页面的树结构根节点完全一致或者近似一致，那么执行步骤203。

[0081] 步骤203：在第一页面的各子树中，确定分别与第二页面中各子树相似度最高的子树，以组成子树对。

[0082] 如图4中所示，对于第一页面中的子树a1、子树a2、子树a3，确定分别与第二页面中的子树b1、子树b2、子树b3相似度最高的子树，以组成子树对，例如，与b1相似度最高的是a2，与b2相似度最高的是a1，与b3相似度最高的是a3，此时，b1和a2组成子树对，b2和a1组成子树对，b3和a3组成子树对。需要说明的是，第二页面中的各子树在第一页面中的子树中相似度最高的子树可能是相同也可能不同，例如，与b1相似度最高的是a2，与b2相似度最高的也是a2，而与b3相似度最高的是a1，此时，b1和a2组成子树对，b2和a2组成子树对，b3和a1组成子树对。

[0083] 步骤204：获得子树对中两个子树的相似度值，并获得子树对中属于第一页面的子树的预设权重。

[0084] 其中，本实施例中可以采用迭代循环方案迭代进入步骤202实现对两个子树的相似度计算。具体的，本实施例中在获得字数对中两个子树的相似度值时，可以迭代进入步骤202中，继续以本实施例中的方案对两个子树的总相似度值进行计算，直到最终到达树结构中的叶子节点，比对叶子节点的相似度，如内容相似度，从而得到叶子节点之间的相似度值，之后，迭代回叶子节点的上一层父节点子树，计算父节点子树之间的总相似度值之后，继续迭代回上一层父节点子树，直到获得子树对中两个子树的相似度值，如图5中所示。

[0085] 其中，第一页面的子树的预设权重可以理解为：页面中各位置结构对用户所表征的重要性程度，属于页面的固有属性，不同页面位置结构的预设权重可能不同，如页面正文所对应的子树预设权重高于页面侧栏所对应的子树预设权重，等等。本实施例中可以在节点路径信息列表或者其相应的页面信息中获得该预设权重。

[0086] 步骤205：基于子树对的相似度值及预设权重，获得第一页面和第二页面之间的总相似度值。

[0087] 具体的，本实施例中可以将子树对的相似度值乘以该子树对中属于第一页面的子

树的预设权重，并加和，得到第一页面和第二页面的总相似度值。

[0088] 例如，第一页面中的子树a1、子树a2、子树a3的预设权重分别为：0.3、0.2和0.1；相应的，如果b1和a2组成子树对c1，b2和a1组成子树对c2，b3和a3组成子树对c3，那么将第一页面的总相似度值为：c1的相似度值*0.2+c2的相似度值*0.3+c3的相似度值*0.1；如果b1和a2组成子树对d1，b2和a2组成子树对d2，b3和a1组成子树对d3，那么将第一页面的总相似度值为：d1的相似度值*0.2+d2的相似度值*0.2+c3的相似度值*0.3。

[0089] 步骤206：判断总相似度值是否高于预设阈值，如果是，执行步骤207，否则，执行步骤208。

[0090] 其中，预设阈值可以根据需求如信息抽取的精度需求或者信息抽取的效率需求等来装置，例如，该预设阈值越高，信息抽取的精度越高，预设阈值越低，信息抽取的效率越高，用户可以根据自身需求来自由设置预设阈值，由此为用户带来更加自由的信息抽取服务模式。

[0091] 步骤207：确定第一页面和第二页面为相似页面。

[0092] 可见，本实施例中通过对第一页面和第二页面的树结构进行相似度比对，从而在总相似度值高于预设阈值时，就可以确定两个页面为相似页面。

[0093] 步骤208：确定第一页面和第二页面不是相似页面。

[0094] 需要说明的是，在步骤202中如果比对两个页面的树结构根节点相同，那么也可以执行步骤208，如图2中所示，可见，本实施例中首先判断两个页面的树结构根节点是不是相同，如果相同，那么继续通过子树计算来得总相似度值来确定是不是相似页面，而如果根节点不相同，那么可以直接确定两个页面不是相似页面。

[0095] 在一种实现方式中，图2所示的两个页面相似度比对是指对两个页面在页面结构上的相似度比对，而在本实施例中还需要对页面内容进行比对，相应的，本实施例中在对第一页面和第二页面进行结构比对的同时，还可以通过对第一页面和第二页面之间进行页面内容所属类别进行比对，由此，本实施例中还需要获得至少一个页面中各个页面的页面内容，从而通过对至少一个页面中的任意两个页面之间进行页面内容所属类别及结构进行比对，从而得到相似页面。而此时的相似页面是指两个页面在结构和内容上都是相同的或者相似度值高于一定阈值的页面。

[0096] 具体的，本实施例中可以采用图2中所示的方案对页面之间的结构进行比对，可以通过对页面标题或隐藏主题等内容进行分析，以对各页面的页面内容所属类别进行确定，从而实现内容比对，得到内容所属类别上的相似度值。进一步的，本实施例中在对页面进行内容所属类别及结构进行比对之后，可以根据内容所属类别及结构所占的权重来再次计算页面之间的页面相似度值，例如，页面内容所属类别的权重为0.5，结构的权重为0.5(或者页面内容所属类别的权重为0.4，结构的权重为0.6，等等)，将页面内容所属类别上的相似度值和结构上的相似度值分别乘以各自权重后加和，得到页面相似度值，从而确定页面之间是否为相似页面。

[0097] 在一种实现方式中，本实施例中步骤103在对相似页面的节点路径信息列表设置标签时，具体可以通过以下方式实现，如图6中所示：

[0098] 步骤601：根据相似页面的节点路径信息列表，确定待抽取的目标内容。

[0099] 具体的，本实施例中可以通过对节点路径信息列表xpath中的文件夹名称或者文

件名称、文件属性、文件类型等内容进行识别，从而确定出可能需要抽取的信息作为目标内容。

[0100] 步骤602：基于目标内容，设置相似页面的节点路径信息列表的标签。

[0101] 具体的，本实施例中可以对以上目标内容进行细化分类后，根据分类结果来确定合适的标签来设置给节点路径信息列表。例如，节点路径信息列表中国/河南/郑州/高新区，与节点信息列表中国/河南/郑州/会展中心，作为相似页面对节点路径信息列表中的各种信息如文件夹名称、文件名称、文件属性、文件类型等内容进行识别，确定为目标内容，从这些目标内容中进行内容细化分类，可以设置“中国郑州”作为这两个相似页面的节点路径信息列表的标签。

[0102] 在一种实现方式中，本实施例中在步骤104进行页面抽取模板生成时，可以首先对具有相同标签的节点路径信息列表生成该标签下的页面抽取模板的特征词典，该特征词典中可以包含有与标签相关联的强特征词及其特征属性，还可以包含有强特征词的同义词等，具体的，本实施例中可以通过以下方式获得特征词典，如图7中所示：

[0103] 步骤701：对具有相同标签的节点路径信息列表，提取这些节点路径信息列表中与其标签存在关联关系的强特征词及其特征属性。

[0104] 其中，强特征词可以为节点路径信息列表中的文件名称、文件夹名称等字词，强特征词可以是包含其所在节点路径信息的标签的字词，也可以是与该标签相似度达到一定阈值的字词，或者与该标签在内容、概念或含义上存在相应的关联关系的字词等。相应的，强特征词的特征属性可以为：文件名称或文件夹名称所对应的文件类型属性，如class属性、css属性等。本实施例中通过对节点路径信息列表中的内容进行解析来提取这些强特征词及其特征属性。

[0105] 步骤702：基于强特征词及其特征属性，生成页面抽取模板的特征词典。

[0106] 其中，本实施例中可以将这些强特征词及其特征属性进行分类整合，以得到词集合，作为后续生成的页面抽取模板的特征词典。

[0107] 步骤703：解析强特征词的词义，以得到强特征词的同义词，并将同义词加入特征词典中。

[0108] 例如，强特征词“台风”，其同义词“强台风”、“热带风暴”等，将这些同义词加入到特征词典中，而在特征词典中，强特征词与其同义词之间具有对应关系。

[0109] 基于以上实现，本实施例中步骤104具体可以通过以下方式实现，如图8中所示：

[0110] 步骤801：对具有相同标签的节点路径信息列表进行合并。

[0111] 具体的，本实施例中可以首先在具有相同标签的节点路径信息列表中，对于提取到强特征词的节点路径信息列表，设置预设的标记符号，如通配符“*”等，用以标记该节点路径信息列表中提取到了强特征词，而其他没有提取到强特征词的节点路径信息列表并不设置标记符号；

[0112] 之后，对节点路径信息列表之间依次按照节点路径信息列表中的节点次序进行一一比对，得到比对结果，该比对结果能够表征节点路径信息列表中各个节点是否对应相同，有哪些节点不同，等等；

[0113] 之后，基于以上比对结果，将节点比对完全相同的节点路径信息列表合并为一个节点路径信息列表，例如保留其中一个，删除另一个节点路径信息列表；而对于存在一个不

同的节点的节点路径信息列表之间也合并为一个节点路径信息列表，并将其中那个不同的节点以标记符号替代，例如，删除其中一个，将另一个节点路径信息列表中与删除的那个节点路径信息列表中存在不同的那个节点以标记符号替代；当然，对于存在两个或多个不同的节点的节点路径信息列表则认为是不同的，不能作为相似进行合并处理。

[0114] 例如，本实施例中在进行节点一一比对的两个页面的节点路径信息列表在节点层级上尽量是相似或相同的，比如都是5层节点或者3层节点，所以就以xpath中“/”隔开的节点进行一一比对，如果完全相同，合并为一个，如果只有一个节点不同，那么也把这两个xpath合并成一个，同时把这合并的xpath中相区别的那个节点用标记符号如通配符“*”代替；如果有多个不同节点，就不认为是可以合并的xpath，不做处理。

[0115] 另外，为了降低数据的计算量，在本实施例中对具有相同标签的节点路径信息列表进行合并时，在对节点路径信息列表之间依次按照节点路径信息列表中的节点次序进行一一比对之前，可以首先对节点路径信息列表进行简化，例如，可以利用标记符号，对节点路径信息列表进行简化。

[0116] 具体的，本实施例中在对节点路径信息列表进行简化时，具体可以通过以下方式实现：

[0117] 对于设置有标记符号的节点路径信息列表的节点路径信息，至少保留节点路径信息中的树结构节点名称及用于填充强特征词的信息；

[0118] 而对于没有设置标记符号的节点路径信息列表的节点路径信息，可以只保留节点路径信息中的树结构节点名称。

[0119] 需要说明的是，如果节点路径信息对应的树结构节点包含某个目标内容如表格内容等，可以保留节点路径信息的节点名称及表格序号信息等，而如果节点路径信息对应的树结构节点不包含目标内容，那么就只保留节点路径信息的节点名称。

[0120] 步骤802：基于合并后的节点路径信息列表，生成与标签相匹配的页面抽取模板。

[0121] 其中，本实施例中可以将合并后的节点路径信息列表直接作为页面抽取模板，在本实施例中得到的页面抽取模板中可以包含有多个节点路径信息，而这些节点路径信息可以用于后续对页面进行信息抽取，而且在节点路径信息中，含有强特征词的节点路径信息的优先级在信息抽取时高于其他不含有强特征词的节点路径信息的优先级，例如，在后续进行页面信息抽取时，优先使用含有强特征词的节点路径信息对页面进行信息抽取，即可以理解为优先使用特征词典中的强特征词进行信息抽取。

[0122] 在一种实现方式中，本实施例中在得到页面抽取模板之后，还可以包括以下步骤，如图9中所示：

[0123] 步骤901：响应于接收到的页面抽取请求，获得待抽取的目标页面以及目标页面的抽取标签。

[0124] 其中，页面抽取请求中可以包含有待抽取的目标页面的页面标识或者页面地址等，以表征待抽取的目标页面；本实施例中可以通过对目标页面的页面标识或页面内容或主题内容等信息进行解析，以得到目标页面的抽取标签。

[0125] 步骤902：在抽取标签对应的目标页面抽取模板中，优先使用含有强特征词的节点路径信息对目标页面进行页面数据抽取得到抽取结果。

[0126] 例如，本实施例中通过抽取标签在各种页面抽取模板中找到具有与该抽取标签相

同的标签的目标页面抽取模板，进而利用该目标页面抽取模板对目标页面进行页面数据抽取，具体的，可以优先使用含有强特征词的节点路径信息进行页面数据抽取，如果无法抽取到合适的信息，再利用目标页面抽取模板的特征词典中的所有或者部分强特征词进行页面数据抽取，如果还是没有抽取到合适的信息，可以考虑使用特征词典中的强特征词的同义词进行页面数据抽取，如果还是没有抽取到合适的信息，那么最后再使用不含有强特征词的节点路径信息进行页面信息抽取，最终得到抽取结果。

[0127] 另外，在利用强特征词进行页面数据抽取而无法得到合适的信息时，还可以使用强特征词的特征属性进行学习训练，例如，对某个文件属性如class属性下的文件名称、文件夹名称等的信息进行抽取训练等，从而抽取到页面中相应的信息。

[0128] 而如果最后仍然没有抽取到合适的信息，那么本实施例中可以对目标页面及其抽取标签标记缺省，并通过获得该目标页面的节点路径信息列表来重新生成相应的页面抽取模板，合并到已经生成的页面抽取模板中，以便于对该目标页面或者更多的其他页面进行信息抽取。

[0129] 需要说明的是，在本实施例中得到抽取结果之后，由于存在数据对齐的缘故，可能会在抽取结果中存在信息冗余的情况，因此，本实施例中在得到抽取结果之后，可以对抽取结果进行进一步清洗，如数据冗余处理将重复的数据删除等，以得到更加精确的抽取结果。

[0130] 参考图10，为本申请实施例二提供的一种数据处理装置的结构示意图，该装置适用于对不同网站即不同结构或内容的网站的页面进行页面抽取模板的构建，进而用于对页面信息进行抽取。本实施例中的装置可以运行在具有计算能力的计算机或服务器中。

[0131] 具体的，本实施例中的装置可以包括以下结构：

[0132] 页面解析单元1001，用于解析至少一个页面，以分别获取每一个页面对应的节点路径信息列表。

[0133] 其中，本实施例中的页面可以包括有一个页面网站上的页面，也可以包括有多个页面网站上的页面，而不同页面网站上的页面在结构和内容上可以相同(或相似)或者不同。例如，购物网站、新闻网站及广告网站上的页面在结构和内容上都不相同。

[0134] 需要说明的是，本实施例中在对页面进行解析之前，肯定是获得到这些页面的，如从数据库中读取这些页面或者利用网络爬虫等工具在网站上实时爬取到这些页面。

[0135] 其中，本实施例中通过对页面解析，得到页面对应的节点路径信息列表可以理解为页面的xpath(XML Path Language)列表，其中，xpath列表使用路径表达式表征页面结构及结构内容。

[0136] 具体的，本实施例中可以通过对页面构建页面的树结构，进而解析树结构来生成xpath列表。例如，本实施例中使用第三方库(例如lxml)解析一个网站或多个网站上的超文本HTML(HyperText Markup Language)页面并构建文档对象模型DOM(Document Object Model)树，进而解析DOM树，以形成相应页面的完整xpath列表。

[0137] 相似比对单元1002，用于基于节点路径信息列表，对至少一个页面中的页面之间进行结构比对，得到相似页面。

[0138] 其中，本实施例中可以对至少一个页面中的任意两个页面分别进行结构比对，从而确定哪些页面属于相似页面，哪些页面并不是相似页面。

[0139] 需要说明的是，本实施例中的相似页面可以理解为：页面之间的相似度值大于一

定阈值的页面才成为相似页面，而页面之间的相似度值可以为页面之间的结构相似度值和/或内容相似度值，也就是说，两个页面之间为相似页面是指：两个页面之间结构和/或页面相似。

[0140] 标签设置单元1003，用于对相似页面的节点路径信息列表设置标签。

[0141] 其中，本实施例中的标签可以为提取页面内容中的字符作为标签，如页面内容中的关键词作为标签；或者标签可以位于页面内容中的字符相关联的字符作为标签，如与页面内容中的关键词近似的词作为标签，等等。

[0142] 而该标签设置的含义至少在于：在同一标签下的节点路径信息列表所对应的页面为相似页面，不同标签下的节点路径信息列表所对应的页面不属于相似页面。

[0143] 具体的，本实施例中可以相似页面的节点路径信息列表按照其信息内容名称及类型进行细化分类，进而基于细化分类的结果来生成标签，并将标签设置给相似页面的节点路径信息列表中。

[0144] 模板生成单元1004，用于基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板。

[0145] 其中，本实施例中可以通过对具有相同标签的节点路径信息列表进行处理，来生成页面抽取模板，例如，选择这些相同标签的节点路径信息列表中与其他节点路径信息列表相似度最高的一个节点路径信息列表作为页面抽取模板；或者基于这些相同标签的节点路径信息列表进行组合或整合来生成该标签下相匹配的页面抽取模板，等等。

[0146] 由上述方案可知，本申请实施例二提供的一种数据处理装置，通过在解析出各种页面的节点路径信息列表之后，基于这些节点路径信息列表对这些不同的页面之间相似度分类，从而对相似页面进行标签设置进而生成该标签下的页面抽取模板，以便于对相应的页面进行信息抽取。可见，本实施例中对于不同网站的网页通过相似度分类，得到结构内容相似的网页后，再生成相应的页面抽取模板，从而实现不同网站网页的数据抽取，该信息抽取的方案能够适用于对不同网站的页面信息抽取，不限于某一种结构或内容的网站页面，从而提高了信息抽取的普遍实用性。

[0147] 基于以上实现方案，以下对本实施例中在针对结构化数据抽取进行模板构建进行具体的抽取时的示例进行举例说明，如图11中所示：

[0148] 步骤1101.对网页生成的DOM树进行结构相似度对比和内容的对比，为同类型且结构相似的网页进行分类。

[0149] 具体的，本实施例中可以使用第三方库(例如lxml)解析HTML页面并构建DOM树，解析DOM树形成该页面的完整XPath列表。得到的树状XPath列表示例如图12，其中，本实施例中没有展示出强特征词和特征属性，仅展示节点以及节点位置。本实施例中首先根据页面标题或隐藏主题进行内容的分类，以确定在内容上的相似页面。为了判断两个或多个网页的结构相似性，对比两个网页树的根节点是否相同，如果不同则相似度为0，停止计算；如果相同，则继续接下来的计算。针对每个子树，从另一个子树集合中选取与之相似度最大的子树作为匹配对象，该相似度为相似度参考值。以子树的节点作为权重，计算所有子树的总参考值，得到两棵树的整体相似度。当满足相似度阈值的情况下，判定两个网页是结构相似的网页。本实施例中的判定方案适用于不同网站的网页界面，对于相同网站的网页，可以使用正则匹配进行判断。

[0150] 步骤1102.解析内容上同类型且结构相似的网页，并按照需要抽取的数据名称和类型整理解析到的XPath路径并标记标签。

[0151] 例如，本实施例中将至少部分内容相同类型且结构相似的HTML页面作为一个领域的页面样本，根据页面内容，确定可能需要抽取的信息，并解析他们对应的XPath路径，进而将待抽取数据按照名称以及类型进行细化分类并标注标签，并把对应XPath路径统一到同一个标签下。为了与部分抽取结果进行对照，相同标签可以保留部分抽取到的原始标记数据或者数据特征。

[0152] 步骤1103.针对相同标签的XPath路径，提取路径中与标签接近的强特征词以及特征属性，并弱化这些特征所在节点，用通配符替代。

[0153] 其中，本实施例中可以在上一步骤1102解析XPath路径时，对于表格类或者在页面上已有对应名称或类型的内容，优先寻找通过该名称可以获取到的XPath路径。按照节点遍历标签下的XPath路径，如文字等，当存在包含与标签接近的文字或代表抽取对象类型的属性时，用‘*’通配符匹配当前节点和特征词位置，并记录相应的文字或属性，及强特征词及其特征属性。

[0154] 步骤1104.整理并归纳强特征词以及特征属性形成特征词典，分析词意与词性，得到其他可能的同义词并加入词典备选。

[0155] 其中，本实施例中整理归纳遍历得到的强特征词，得到初步的特征词典。因为当前抽取对象主要为中文或英文网页，使用中文或英文近义词工具包寻找相似度较高的近义词，并加入词典备选；对于强特征词是短语或者词语组合的情况而言，需要先进行分词，选择切分后词性为名词的单词，寻找相似度较高的近义词，并加入词典备选。如果寻找到的同义词在特征词典中已经存在，则不再添加；如果仅包含在词典中的某个短语内，则继续添加。

[0156] 步骤1105.基于步骤1103中的处理，依照设计好的对照表简化并合并同标签的XPath路径，使可能含有强特征词以及特征属性的路径可以被优先选择，其他备选路径为较低级别，并保存为该标签下的模板。

[0157] 其中，本实施例中首先简化xpath路径，具体的，对于没有用通配符替代的节点，仅保留该节点名称，部分表格相关节点保留至序号部分；针对使用通配符替代的节点，保留节点名称以及可用于填充强特征词的部分。缩减完成所有节点后，得到初步简化的XPath路径。

[0158] 之后，本实施例中合并同标签的xpath路径，具体的，依次按照节点对比同一标签下的路径，合并完全相同的条目；如果存在两条以上XPath表达式之间仅有一个节点不同，则将它们等同看作完全相同的表达式，并用通配符替代唯一不同的节点。

[0159] 步骤1106.选择待抽取标签下的模板抽取同类网页，优先选择含有强特征词的路径，如果未匹配，则替换特征辞典中的同义词；当匹配完词典中所有同义词后，开始选择没有特征属性的XPath路径，并将抽取到的文本与标记数据对比。

[0160] 其中，本实施例中在需要对页面进行信息抽取时，针对类型相似且包含待抽取信息的目标网页，可以在确定目标网页的待抽取标签之后，根据需求选择待抽取标签下的页面抽取模板，进而优先选择模板中含有强特征词的路径和对应特征词典进行抽取；如果未匹配即没有抽取到合适的信息，则替换特征辞典中的同义词进行再次匹配。如果出现匹配

完词典中所有同义词，都无法抽取到结果的情况，则开始选择使用不带有特征属性的XPath路径进行匹配，并将抽取到的文本与原始标记数据对比，保留合理的信息作为抽取到的结果。

[0161] 需要说明的是，本实施例中的页面抽取模板可以对页面进行批量抽取。而在批量抽取的过程中，针对相同网站内的网页，如果已经使用某一条XPath路径成功抽取到了所需内容，则记录下当前标签对应的XPath路径，并在之后的抽取中优先使用这些路径进行抽取。需要注意的是，可能同一个页面有多个可以使用同一个标签抽取的数据，则因为数据对齐的缘故，可能出现抽取到信息冗余的情况，需要对抽取结果进行进一步清洗。

[0162] 步骤1107.如果没有匹配到待抽取数据，则标记缺省，并将该网页整体解析并加入合并模板。

[0163] 具体的，本实施例在批量抽取中，如果使用少量标签没有匹配到待抽取数据，则标记缺省，并继续抽取。而未匹配到的标签可以通过已抽取到的同类型网站的数据进行互补或手动添加相应的可抽取XPath路径。如果多数待抽取标签没有匹配到对应信息，则需要该网页整体解析作为种子，重复模板生成步骤，得到对应的XPath路径。

[0164] 可见，本实施例中对于不同网站的网页，可以通过计算相似度来进行分类，得到结构内容相似的网页；而在解析样本充足的情况下，对大部分相似的网页，其差异部分可以使用条件表达式进行微处理，模板通用性较强，有较为广泛的应用范围；同时，本实施例中使用自定义标签来归类特征词以及特征属性，利于归类有多种表达方式的相同含义字段；另外，本实施例中使用路径中原有词以及同义词组成特征词典，对不同网页的同类数据更具有普适性；而且仅保存待抽取的数据模板，不需要对无用数据进行整理。

[0165] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

[0166] 专业人员还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0167] 结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[0168] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

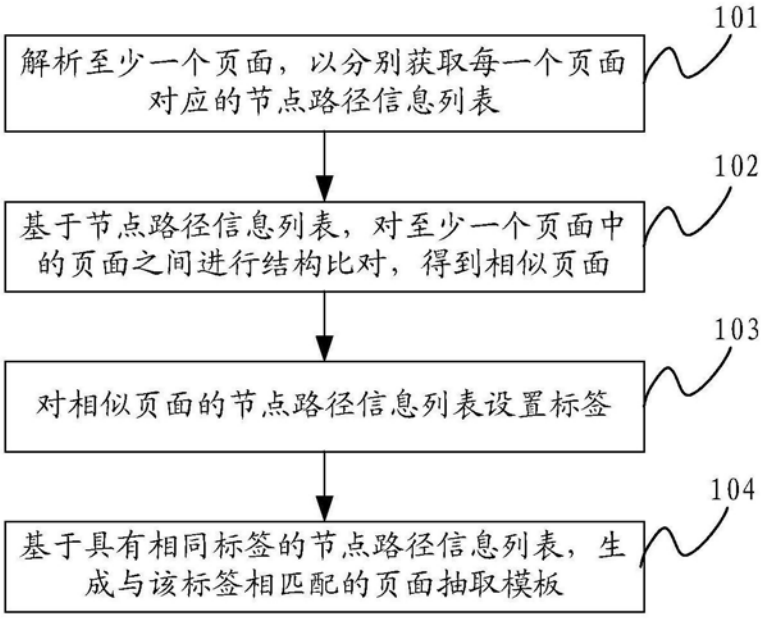

101
解析至少一个页面，以分别获取每一个页面对应的节点路径信息列表
102
基于节点路径信息列表，对至少一个页面中的页面之间进行结构比对，得到相似页面
103
对相似页面的节点路径信息列表设置标签
104
基于具有相同标签的节点路径信息列表，生成与该标签相匹配的页面抽取模板

图1

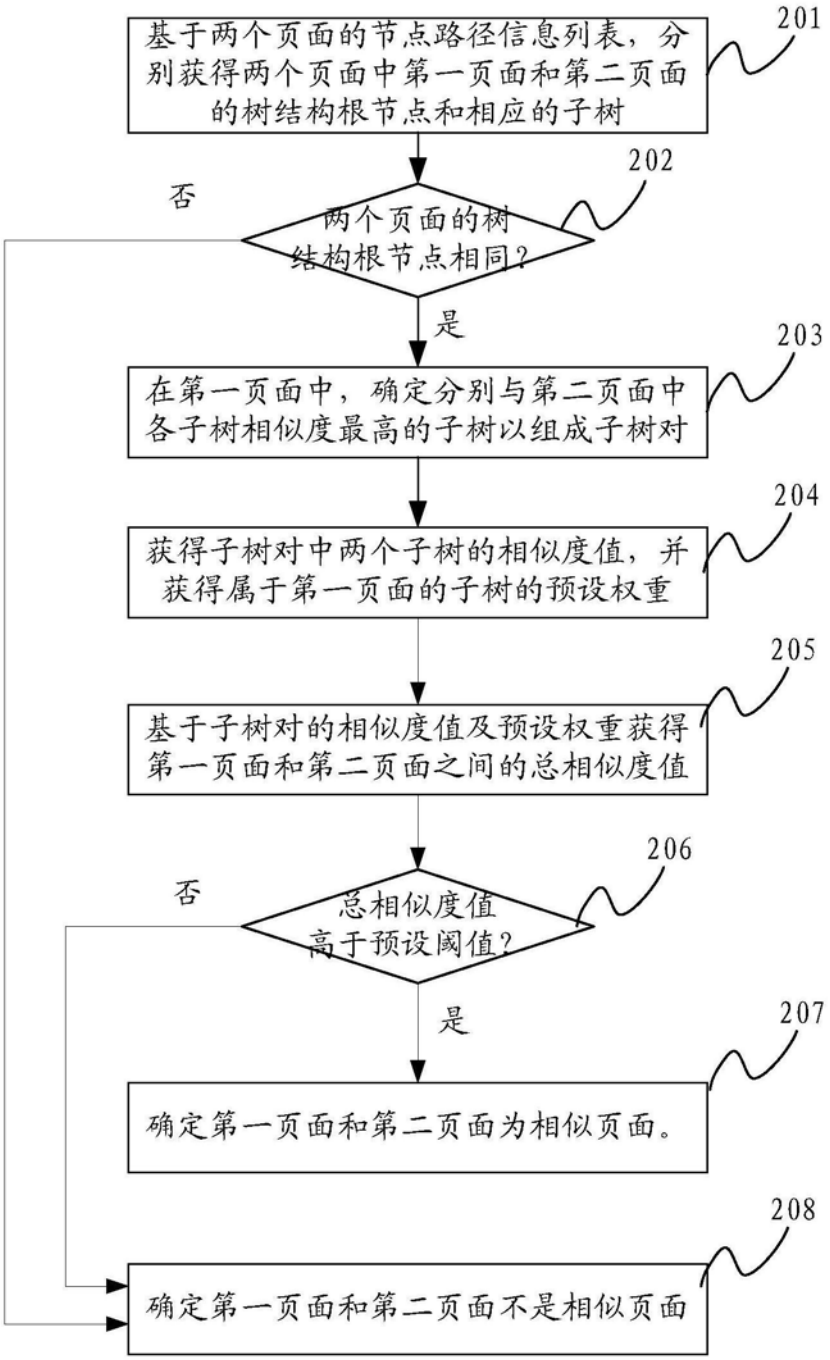

基于两个页面的节点路径信息列表，分别获得两个页面中第一页面和第二页面的树结构根节点和相应的子树
201
两个页面的树结构根节点相同？
202
否
是
在第一页面中，确定分别与第二页面中各子树相似度最高的子树以组成子树对
203
获得子树对中两个子树的相似度值，并获得属于第一页面的子树的预设权重
204
基于子树对的相似度值及预设权重获得第一页面和第二页面之间的总相似度值
205
总相似度值高于预设阈值？
206
否
是
确定第一页面和第二页面为相似页面。
207
确定第一页面和第二页面不是相似页面
208

图2

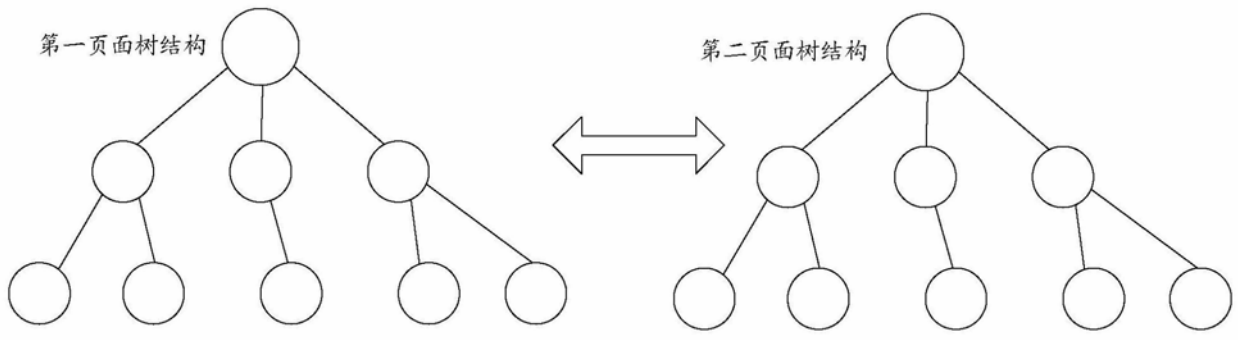

第一页面树结构
第二页面树结构

图3

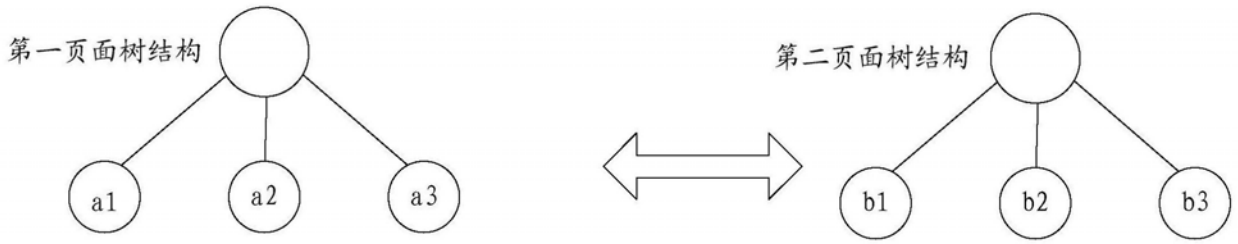

第一页面树结构
a1
a2
a3
第二页面树结构
b1
b2
b3

图4

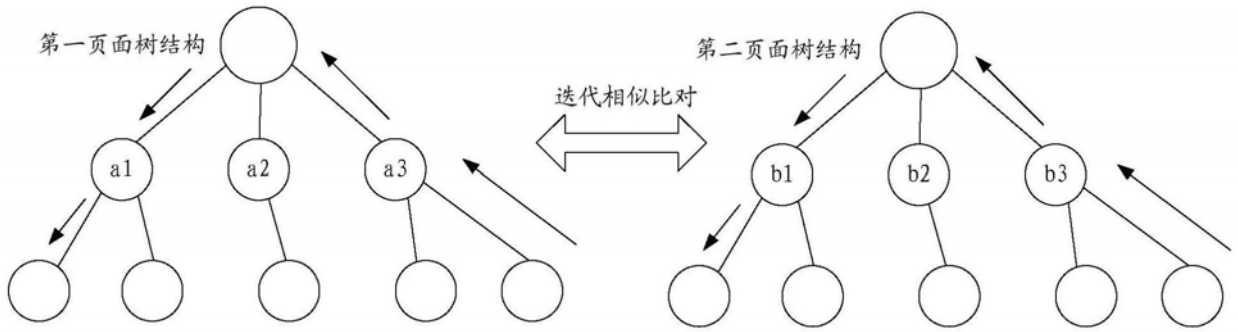

第一页面树结构
a1
a2
a3
迭代相似比对
第二页面树结构
b1
b2
b3

图5

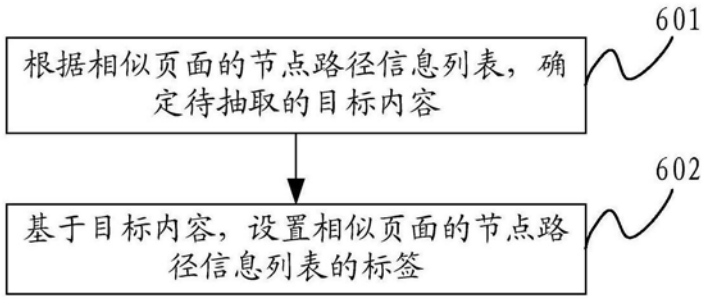

601
根据相似页面的节点路径信息列表，确定待抽取的目标内容
602
基于目标内容，设置相似页面的节点路径信息列表的标签

图6

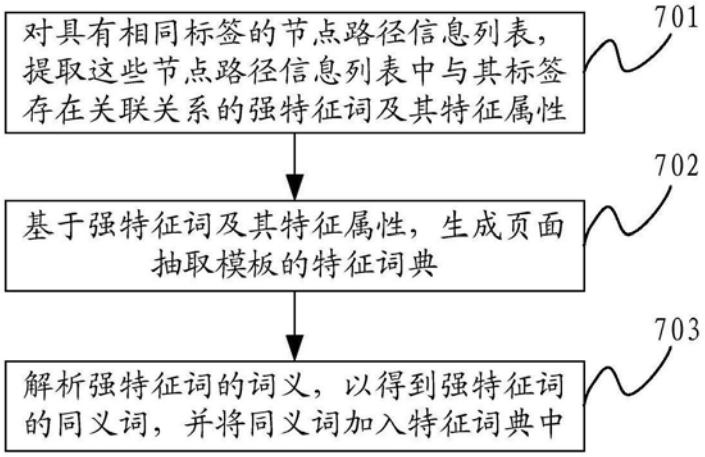

对具有相同标签的节点路径信息列表，提取这些节点路径信息列表中与其标签存在关联关系的强特征词及其特征属性
701
基于强特征词及其特征属性，生成页面抽取模板的特征词典
702
解析强特征词的词义，以得到强特征词的同义词，并将同义词加入特征词典中
703

图7

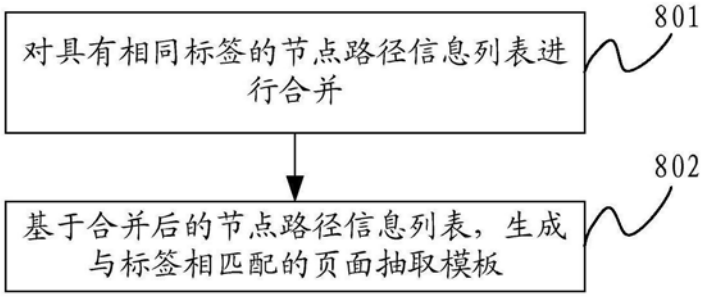

对具有相同标签的节点路径信息列表进行合并
801
基于合并后的节点路径信息列表，生成与标签相匹配的页面抽取模板
802

图8

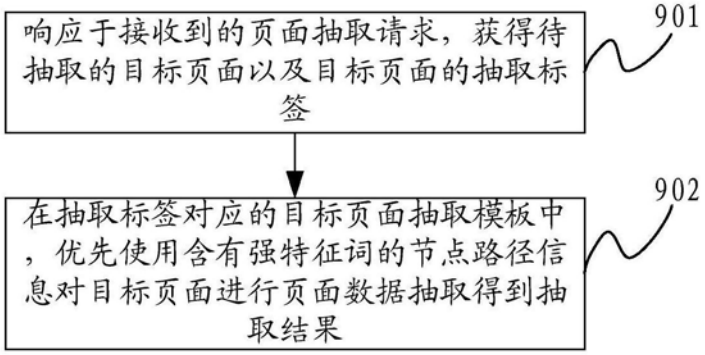

响应于接收到的页面抽取请求，获得待抽取的目标页面以及目标页面的抽取标签
901
在抽取标签对应的目标页面抽取模板中，优先使用含有强特征词的节点路径信息对目标页面进行页面数据抽取得到抽取结果
902

图9

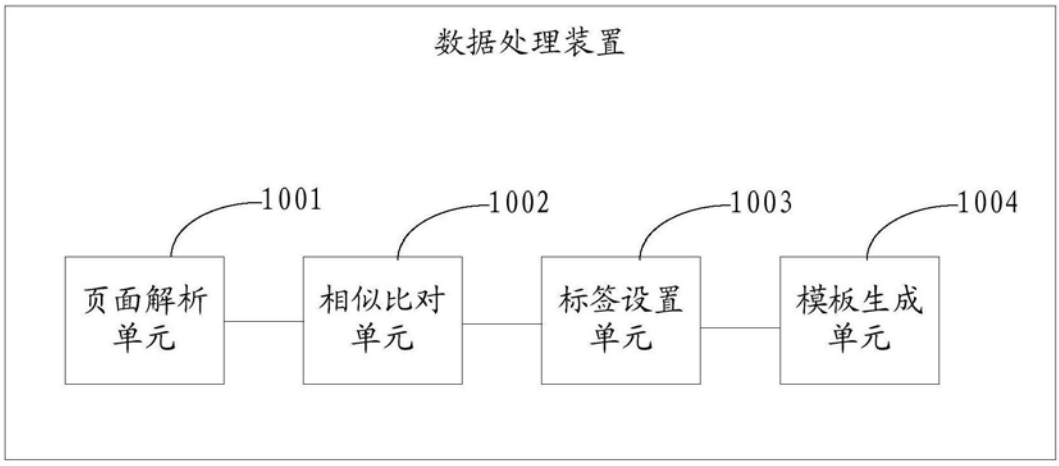


图10

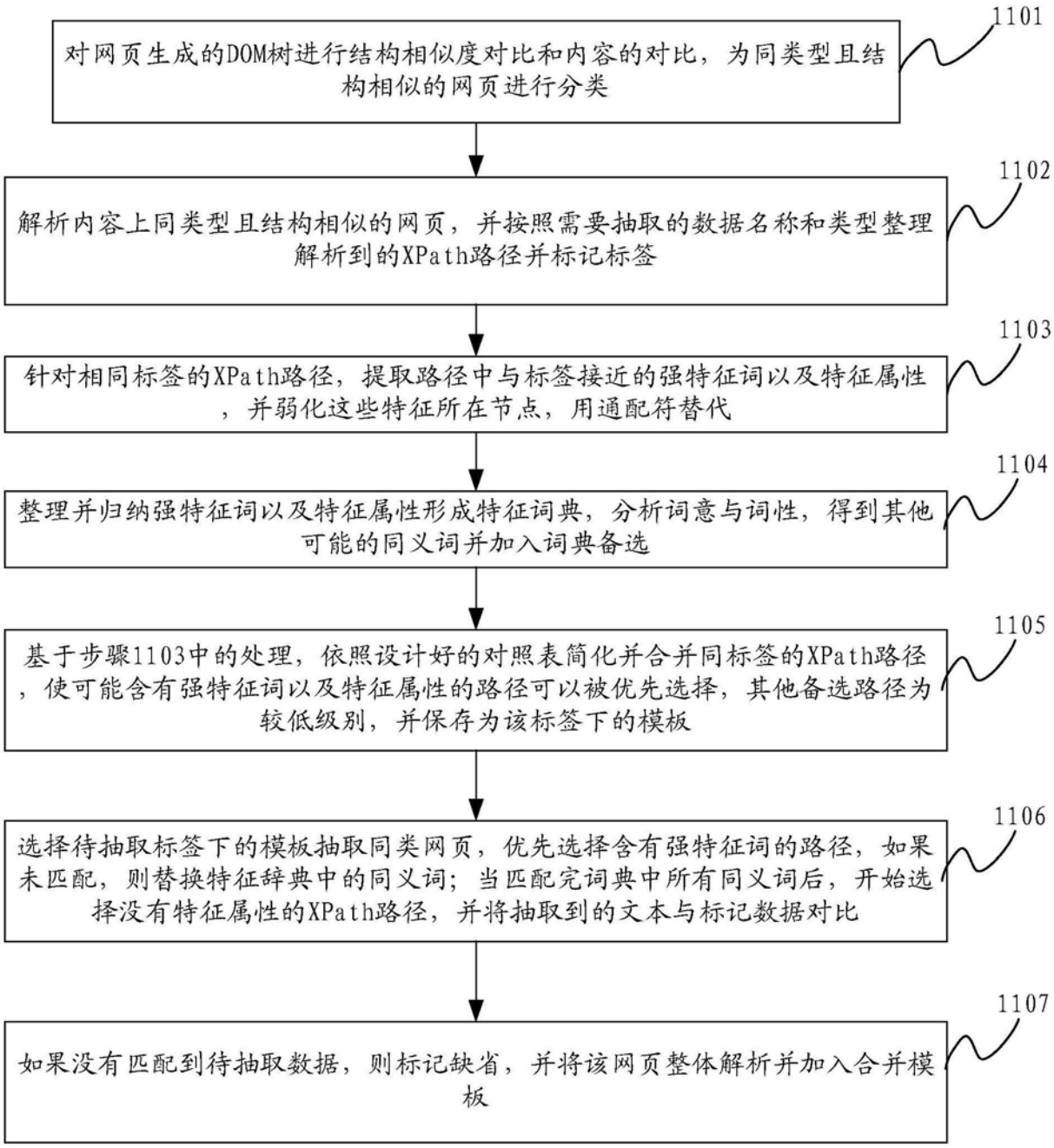

1101
对网页生成的DOM树进行结构相似度对比和内容的对比，为同类型且结构相似的网页进行分类
1102
解析内容上同类型且结构相似的网页，并按照需要抽取的数据名称和类型整理解析到的XPath路径并标记标签
1103
针对相同标签的XPath路径，提取路径中与标签接近的强特征词以及特征属性，并弱化这些特征所在节点，用通配符替代
1104
整理并归纳强特征词以及特征属性形成特征词典，分析词意与词性，得到其他可能的同义词并加入词典备选
1105
基于步骤1103中的处理，依照设计好的对照表简化并合并同标签的XPath路径，使可能含有强特征词以及特征属性的路径可以被优先选择，其他备选路径为较低级别，并保存为该标签下的模板
1106
选择待抽取标签下的模板抽取同类网页，优先选择含有强特征词的路径，如果未匹配，则替换特征辞典中的同义词；当匹配完词典中所有同义词后，开始选择没有特征属性的XPath路径，并将抽取到的文本与标记数据对比
1107
如果没有匹配到待抽取数据，则标记缺省，并将该网页整体解析并加入合并模板

图11

```
/html/body/div[1]/div[1]/div/div[1]/div
/html/body/div[1]/div[1]/div/div[1]/div/div[1]
/html/body/div[1]/div[1]/div/div[1]/div/a
/html/body/div[1]/div[1]/div/div[1]/div/form
/html/body/div[1]/div[1]/div/div[1]/div/form/span[1]
/html/body/div[1]/div[1]/div/div[1]/div/form/span[2]
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span/div
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span/div/span
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span/ul
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span/ul/li[1]
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span/ul/li[1]/a
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span/ul/li[2]
/html/body/div[1]/div[1]/div/div[1]/div/form/span[3]/span/ul/li[2]/a
```

图12