(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 110110252 B
(45)授权公告日 2021.01.15

(21)申请号 201910416413.0

(22)申请日 2019.05.17

(65)同一申请的已公布的文献号
申请公布号 CN 110110252 A

(43)申请公布日 2019.08.09

(73)专利权人 北京市博汇科技股份有限公司
地址 100000 北京市海淀区丰贤中路7号(孵化楼)5层501

(72)发明人 白冰 关靖霖 李国华

(74)专利代理机构 北京超凡宏宇专利代理事务所(特殊普通合伙) 11463
代理人 徐彦圣

(51)Int.Cl.
G06F 16/955(2019.01)
G06F 16/958(2019.01)
G06F 16/738(2019.01)
G06F 16/783(2019.01)
G06F 40/242(2020.01)
G06F 40/289(2020.01)

(56)对比文件
CN 108334508 A,2018.07.27
US 6862731 B1,2005.03.01
CN 102156737 A,2011.08.17
US 2012290918 A1,2012.11.15
CN 102332028 A,2012.01.25
叶利华."视频标签检测与识别".《制造业自动化》.2011,
A. Stefanidis等."Summarizing video datasets in the spatiotemporal domain".《Proceedings 11th International Workshop on Database and Expert Systems Applications》.2002,

审查员 张思洋

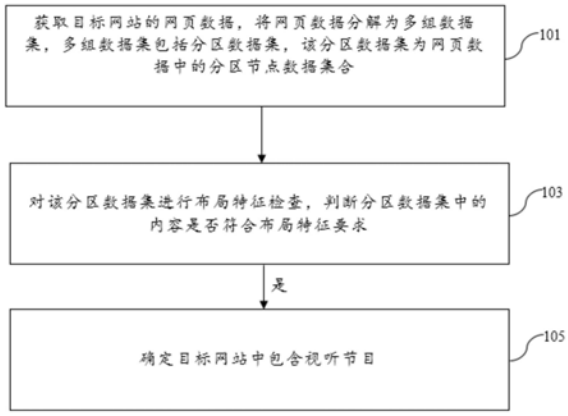
权利要求书2页 说明书9页 附图6页

(54)发明名称

一种视听节目识别方法、装置及存储介质

(57)摘要

本申请提供了一种视听节目识别方法、装置及存储介质，该方法包括获取目标网站的网页数据，将网页数据分解为多组数据集，多组数据集包括分区数据集，该分区数据集为网页数据中的分区节点数据集合；对分区数据集进行布局特征检查，判断分区数据集中的内容是否符合布局特征要求；若分区数据集中的内容符合布局特征要求，则确定目标网站中包含视听节目，具有准确率高的优点。



CN 110110252 B

1.一种视听节目识别方法,其特征在于,所述方法包括:

获取目标网站的网页数据,将所述网页数据分解为多组数据集,所述多组数据集包括分区数据集,所述分区数据集为所述网页数据中的分区节点数据集合;

对所述分区数据集进行布局特征检查,判断所述分区数据集中的内容是否符合布局特征要求;

若所述分区数据集中的内容符合布局特征要求,则确定所述目标网站中包含视听节目。

2.根据权利要求1所述方法,其特征在于,所述对所述分区数据集进行布局特征检查,判断所述分区数据集中的内容是否符合布局特征要求,包括:

去除所述分区数据集中的HTML标签内容;

提取所述分区数据集中每个分区节点的位置信息,并根据所述位置信息以及所述位置信息对应的分区节点构建布局特征,所述布局特征包括特征位置;

对所述特征位置中包含的预设格式的数据信息进行标记;

判断所述分区数据集的标记数据个数占所述分区数据集的总体个数的比率是否在预设的阈值范围内;

若是,则确定所述分区数据集中的内容符合布局特征要求。

3.根据权利要求1所述方法,其特征在于,在所述将所述网页数据分解为多组数据集之后,所述方法还包括:

所述多组数据集还包括链接数据集,所述链接数据集为所述网页数据中的链接节点数据集合;

对所述链接数据集进行链接文本特征检查,判断所述链接数据集中的内容是否符合链接文本特征要求;

若所述链接数据集中的内容符合链接文本特征要求,则确定所述目标网站中包含视听节目。

4.根据权利要求3所述方法,其特征在于,所述对所述链接数据集进行链接文本特征检查,判断所述链接数据集中的内容是否符合链接文本特征要求,包括:

去除所述链接数据集中的HTML标签内容;

对所述链接数据集中剩余的文本内容进行分词,根据分词后的文本内容建立词典;

获取所述词典中的词语出现预设判别词的统计结果,所述统计结果包括词频、出现个数以及对应词语的长度;

判断所述统计结果是否在预设的阈值范围内,若是,则确定所述链接数据集中的内容符合链接文本特征要求。

5.根据权利要求1所述方法,其特征在于,在判断所述分区数据集中的内容是否符合布局特征要求之后,所述方法还包括:

若所述分区数据集中的内容不符合布局特征要求,则对所述网页数据的所有内容进行视频文件检查以及播放器检查,判断所述网页数据的所有内容是否符合视频文件检查要求以及播放器检查要求;

若所述网页数据的所有内容符合视频文件检查要求和/或所述网页数据的所有内容符合播放器检查要求,则确定所述目标网站中包含视听节目。

6.根据权利要求5所述方法，其特征在于，对网页数据的所有内容进行视频文件检查，判断所述网页数据的所有内容是否符合视频文件检查要求，包括：

对所述网页数据所有内容的各个节点进行视频文件检查；

判断每个节点内容是否包含视频类型的文件；

若存在任意一个节点内容包含视频类型的文件，则确定所述网页数据的所有内容符合视频文件检查要求。

7.根据权利要求5所述方法，其特征在于，对网页数据的所有内容进行播放器检查，判断所述网页数据的所有内容是否符合播放器检查要求，包括：

对所述网页数据所有内容逐条进行播放器检查；

判断所述网页数据所有内容中是否存在播放器链接和/或引入播放器文件；

若存在播放器链接和/或引入播放器文件，则确定所述网页数据的所有内容符合播放器检查要求。

8.根据权利要求1所述方法，其特征在于，在所述对所述分区数据集进行布局特征检查之前，所述方法还包括：

判断所述分区数据集是否适用于布局特征检查；

若是，则对所述分区数据集进行布局特征检查。

9.一种视听节目识别装置，其特征在于，所述装置包括：

获取模块，用于获取目标网站的网页数据；

分解模块，用于将所述网页数据分解为多组数据集，所述多组数据集包括分区数据集，所述分区数据集为所述网页数据中的分区节点数据集合；

检查模块，用于对所述分区数据集进行布局特征检查；

判断模块，用于判断所述分区数据集中的内容是否符合布局特征要求；

确定模块，用于在所述判断模块判断所述分区数据集中的内容符合布局特征要求之后，确定所述目标网站中包含视听节目。

10.一种非暂态计算机可读存储介质，其特征在于，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令使计算机执行如权利要求1-8中任意一项所述的方法。

一种视听节目识别方法、装置及存储介质

技术领域

[0001] 本申请涉及互联网监测技术领域，主要涉及一种视听节目识别方法、装置及存储介质。

背景技术

[0002] 目前，判断互联网网站是否包含视听节目的方式都是通过关键词对比来进行判别，然而通过关键词进行对比实现判别的方式存在着准确率低的问题。

发明内容

[0003] 本申请的目的在于提供一种视听节目识别方法、装置及存储介质，用于解决现有技术中关键词对比存在的准确率低的问题。

[0004] 为了实现上述目的，本申请提供了以下技术方案如下：

[0005] 第一方面：本申请提供了一种视听节目识别方法，所述方法包括：获取目标网站的网页数据，将所述网页数据分解为多组数据集，所述多组数据集包括分区数据集，所述分区数据集为所述网页数据中的分区节点数据集合；对所述分区数据集进行布局特征检查，判断所述分区数据集中的内容是否符合布局特征要求；若所述分区数据集中的内容符合布局特征要求，则确定所述目标网站中包含视听节目。

[0006] 上述方案设计的方法，通过对分区数据集进行布局特征检查，判断其是否满足特定的布局特征要求，使得对网页中是否包含视听节目的识别更加准确。

[0007] 在第一方面的可选实施方式中，所述对所述分区数据集进行布局特征检查，判断所述分区数据集中的内容是否符合布局特征要求，包括：去除所述分区数据集中HTML标签内容；提取所述分区数据集中每个分区节点的位置信息，并根据所述位置信息以及所述位置信息对应的分区节点构建布局特征，所述布局特征包括特征位置；对所述特征位置中包含预设格式的数据信息进行标记；判断所述分区数据集的标记数据个数占所述分区数据集的总体个数的比率是否在预设的阈值范围内；若是，则所述分区数据集中的内容符合布局特征要求。

[0008] 上述方案设计的方法，通过对四个特征位置布局的标记数据进行统计来进行视听节目的判断，使得对视听节目的识别更加准确。

[0009] 在第一方面的可选实施方式中，在所述将所述网页数据分解为多组数据集之后，所述方法还包括：所述多组数据集还包括链接数据集，所述链接数据集为所述网页数据中的链接节点数据集合；对所述链接数据集进行链接文本特征检查，判断所述链接数据集中的内容是否符合链接文本特征要求；若所述链接数据集中的内容符合链接文本特征要求，则确定所述目标网站中包含视听节目。

[0010] 上述方案设计的方法，在布局特征检查的基础上加入链接文本特征检查，增加了特征检查方式，使得对网页数据内视听节目的识别更加准确。

[0011] 在第一方面的可选实施方式中，所述对所述链接数据集进行链接文本特征检查，

判断所述链接数据集中的内容是否符合链接文本特征要求，包括：去除所述链接数据集中HTML标签内容；对所述链接数据集中剩余的文本内容进行分词，根据分词后的文本内容建立词典；获取所述词典中的词语出现预设判别词的统计结果，所述统计结果包括词频、出现个数以及对应词语的长度；判断所述统计的结果是否在预设的阈值范围内，若是，则所述链接数据集中的内容符合链接文本特征要求。

[0012] 上述方案设计的方法，描述了实现链接文本特征检查的具体方式，通过对网页数据中词语出现判别词的统计结果来进行特征判断，使得对视听节目的判断更加准确。

[0013] 在第一方面的可选实施方式中，在判断所述分区数据集中的内容是否符合布局特征要求之后，所述方法还包括：若所述分区数据集中的内容不符合布局特征要求，则对所述网页数据的所有内容进行视频文件检查以及播放器检查，判断所述网页数据的所有内容是否符合视频文件检查要求以及播放器检查要求；若所述网页数据的所有内容符合视频文件检查要求和/或所述网页数据的所有内容符合播放器检查要求，则确定所述目标网站中包含视听节目。

[0014] 上述方案设计的方法，在布局特征要求不符合时，则进行对网页数据的所有内容进行后续的判断，使得判断网页是否包含视听节目的准确率大大提高，并且后续方法具有极强的通用性。

[0015] 在第一方面的可选实施方式中，对网页数据的所有内容进行视频文件检查，判断所述网页数据的所有内容是否符合视频文件检查要求，包括：对所述网页数据所有内容的各个节点进行视频文件检查；判断每个节点内容是否包含视频类型的文件；若存在任意一个节点内容包含视频类型的文件，则所述网页数据的所有内容符合视频文件检查要求。

[0016] 在第一方面的可选实施方式中，对网页数据的所有内容进行播放器检查，判断所述网页数据的所有内容是否符合播放器检查要求，包括：对所述网页数据所有内容逐条进行播放器检查；判断所述网页数据所有内容中是否存在播放器链接和/或引入播放器文件；若存在，则所述网页数据的所有内容符合播放器检查要求。

[0017] 在第一方面的可选实施方式中，在所述对所述分区数据集进行布局特征检查之前，所述方法还包括：判断所述分区数据集是否适用于布局特征检查；若是，则对所述分区数据集进行布局特征检查。

[0018] 第二方面：本申请提供一种视听节目识别装置，所述装置包括：获取模块，用于获取目标网站的网页数据；分解模块，用于将所述网页数据分解为多组数据集，所述多组数据集包括分区数据集，所述分区数据集为所述网页数据中的分区节点数据集合；检查模块，用于对所述分区数据集进行布局特征检查；判断模块，用于判断所述分区数据集中的内容是否符合布局特征要求；确定模块，用于在所述判断模块所述分区数据集中的内容符合布局特征要求之后，确定所述目标网站中包含视听节目。

[0019] 上述方案设计的装置，通过对分区数据集进行布局特征检查，判断其是否满足特定的布局特征要求，使得对网页中是否包含视听节目的识别更加准确。

[0020] 在第二方面的可选实施方式中，所述装置还包括：去除模块，用于去除所述分区数据集中HTML标签内容；提取模块，用于提取所述分区数据集中每个分区节点的位置信息，并根据所述位置信息以及所述位置信息对应的分区节点构建布局特征，所述布局特征包括特征位置；标记模块，用于对所述特征位置中包含预设格式的数据信息进行标记；所述判断模

块，还用于判断所述分区数据集的标记数据个数占所述分区数据集的总体个数的比率是否在预设的阈值范围内；所述确定模块，还用于在所述判断模块判断所述分区数据集的标记数据个数占所述分区数据集的总体个数的比率在预设的阈值范围内，确定所述分区数据集中的内容符合布局特征要求。

[0021] 在第二方面的可选实施方式中，所述多组数据集还包括链接数据集，所述链接数据集为所述网页数据中的链接节点数据集合；所述检查模块，还用于对所述链接数据集进行链接文本特征检查；所述判断模块，还用于判断所述链接数据集中的内容是否符合链接文本特征要求；所述确定模块，还用于在所述判断模块判断所述链接数据集中的内容符合链接文本特征要求时，确定所述目标网站中包含视听节目。

[0022] 在第二方面的可选实施方式中，所述去除模块，还用于去除所述链接数据集中HTML标签内容；分词模块，用于对所述链接数据集中剩余的文本内容进行分词；构建模块，用于根据分词后的文本内容建立词典；所述获取模块，还用于获取所述词典中的词语出现预设判别词的统计结果，所述统计结果包括词频、出现个数以及对应词语的长度；所述判断模块，还用于判断所述统计的结果是否在预设的阈值范围内；所述确定模块，还用于在所述判断模块判断所述统计的结果在预设的阈值范围内时，确定所述链接数据集中的内容符合链接文本特征要求。

[0023] 在第二方面的可选实施方式中，所述检查模块，还用于在所述确定模块确定所述分区数据集中的内容不符合布局特征要求之后，对所述网页数据的所有内容进行视频文件检查以及播放器检查；所述判断模块，还用于判断所述网页数据的所有内容是否符合视频文件检查要求以及播放器检查要求；所述确定模块，还用于在所述判断模块判断所述网页数据的所有内容符合视频文件检查要求和/或所述媒体数据集中的内容符合播放器检查要求时，确定所述目标网站中包含视听节目。

[0024] 第三方面：本申请提供一种电子设备，包括：处理器，以及分别与处理器连接的存储器和通信模块，所述存储器存储有所述处理器可执行的机器可读指令，所述通信模块用于与外部设备进行通信传输；当所述计算设备运行时，所述处理器执行所述机器可读指令，以执行时执行第一方面、第一方面的任一可选的实现方式中的所述方法。

[0025] 第四方面：本申请提供一种非暂态计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行第一方面、第一方面的任一可选的实现方式中的所述方法。

[0026] 第五方面：本申请提供一种计算机程序产品，所述计算机程序产品在计算机上运行时，使得计算机执行第一方面、第一方面的任一可选的实现方式中的所述方法。

[0027] 本申请的其他特征和优点将在随后的说明书阐述，并且，部分地从说明书中变得显而易见，或者通过实施本申请实施例而了解。本申请的目的和其他优点可通过在所写的说明书、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

[0028] 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。通过附图所示，本申请的上述及其它目的、特征和优势将更加清晰。在全部附图中相同的附图标记指示相同的部分。并未刻意按实际尺寸等比例缩放绘制附图，重点在于示出本申请的主旨。

[0029] 图1是本申请第一实施例提供的视听节目识别方法第一流程图；

[0030] 图2是本申请第一实施例提供的视听节目识别方法第二流程图；

[0031] 图3是本申请第一实施例提供的视听节目识别方法第三流程图；

[0032] 图4是本申请第一实施例提供的视听节目识别方法第四流程图；

[0033] 图5是本申请第一实施例提供的视听节目识别方法第五流程图；

[0034] 图6是本申请第一实施例提供的视听节目识别方法第六流程图；

[0035] 图7是本申请第一实施例提供的视听节目识别方法第七流程图；

[0036] 图8是本申请第二实施例提供的视听节目识别装置结构示意图；

[0037] 图9是本申请第三实施例提供的电子设备结构示意图。

具体实施方式

[0038] 为使本申请实施方式的目的、技术方案和优点更加清楚，下面将结合本申请实施方式中的附图，对本申请实施方式中的技术方案进行清楚、完整地描述，显然，所描述的实施方式是本申请一部分实施方式，而不是全部的实施方式。基于本申请中的实施方式，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施方式，都属于本申请保护的范围。因此，以下对在附图中提供的本申请的实施方式的详细描述并非旨在限制要求保护的本申请的范围，而是仅仅表示本申请的选定实施方式。基于本申请中的实施方式，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施方式，都属于本申请保护的范围。

[0039] 在本申请的描述中，需要理解的是，术语“中心”、“长度”、“宽度”、“厚度”、“上”、“下”、“前”、“后”、“左”、“右”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本申请和简化描述，而不是指示或暗示所指的设备或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本申请的限制。

[0040] 此外，术语“第一”、“第二”等仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”等的特征可以明示或者隐含地包括一个或者更多个该特征。在本申请的描述中，“多个”的含义是两个或两个以上，除非另有明确具体的限定。

[0041] 在本申请中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”、“固定”等术语应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或成一体；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通或两个元件的相互作用关系。对于本领域的普通技术人员而言，可以根据具体情况理解上述术语在本申请中的具体含义。

[0042] 在本申请中，除非另有明确的规定和限定，第一特征在第二特征之“上”或之“下”可以包括第一和第二特征直接接触，也可以包括第一和第二特征不是直接接触而是通过它们之间的另外的特征接触。而且，第一特征在第二特征“之上”、“上方”和“上面”包括第一特征在第二特征正上方和斜上方，或仅仅表示第一特征水平高度高于第二特征。第一特征在

第二特征“之下”、“下方”和“下面”包括第一特征在第二特征正下方和斜下方，或仅仅表示第一特征水平高度小于第二特征。

[0043] 第一实施例

[0044] 如图1所示，本申请提供一种视听节目识别方法，该方法包括：

[0045] 步骤101：获取目标网站的网页数据，将网页数据分解为多组数据集，多组数据集包括分区数据集，该分区数据集为网页数据中的分区节点数据集合，转到步骤103。

[0046] 步骤103：对该分区数据集进行布局特征检查，判断分区数据集中的内容是否符合布局特征要求，转到步骤105。

[0047] 步骤105：若分区数据集中的内容符合布局特征要求，则确定目标网站中包含视听节目。

[0048] 其中，步骤101中获取目标网站的网页数据可为获取目标网站首页的网页数据，也可以不是目标网站首页的网页数据，抓取网页数据的多种方式，例如get方式或post方式等。

[0049] 另外，步骤101中获取的分区数据集的方式也就是在多组数据集中提取分区节点的数据集，也就是包含特定布局特征的 标签及其上下文的信息数据集。

[0050] 对于步骤103中的布局特征可理解为节点区域中预设的特定位置中的内容。

[0051] 上述方案设计的方法，通过对分区数据集进行布局特征检查，判断其是否满足特定的布局特征要求，使得对网页中是否包含视听节目的识别更加准确。

[0052] 可选地，如图2所示，在步骤103中的对该分区数据集进行布局特征检查之前，该方法还包括：

[0053] 步骤102：判断该分区数据集是否适用于布局特征检查，若是，则转到步骤103。

[0054] 对于步骤102，其具体可为提取该分区数据集中的每一条数据，获取该 标签下的<img>和<a>标签数量，如果 中的<img>和<a>标签配对数超过某一个阈值，则认为该分区数据集适用于布局特征检查。

[0055] 可选地，如图3所示，对于步骤103中的对该分区数据集进行布局特征检查，判断分区数据集中的内容是否符合布局特征要求，包括：

[0056] 步骤1031：去除分区数据集中的HTML标签内容，转到步骤1032。

[0057] 步骤1032：提取分区数据集中每个分区节点的位置信息，并根据位置信息以及位置信息对应的分区节点构建布局特征，该布局特征包括特征位置，转到步骤1033。

[0058] 步骤1033：对特征位置中包含预设格式的数据信息进行标记；判断分区数据集的标记数据个数占分区数据集的总体个数的比率是否在预设的阈值范围内，转到步骤1034。

[0059] 步骤1034：若分区数据集的标记数据个数占分区数据集的总体个数的比率在预设的阈值范围内，则确定分区数据集中的内容符合布局特征要求。

[0060] 对于步骤1032中的位置信息可包括与父节点的相对位置信息或者与页面的绝对位置信息。

[0061] 对于上述步骤1031-1034，其具体方案可通过以下方式实现：

[0062] 提取每一条 中取出HTML后相应的位置信息，并与该 构建成字典组。如图4所示，特征包括 左上、左下、右上、右下四部分。

[0063] 对于出现在该 的四个位置的内容进行判别，对于包含特定信息的数据进行

标记，例如，包含时间格式信息“00:00”或包含描述电视剧剧集的信息“第N集”或者包含某些特定词“N人观看”“N次播放”等。

[0064] 统计分区数据集中被标记的数据占分区数据集总数据的比率，当比率在预设的阈值范围内的时候，则表示分区数据集中的内容符合布局特征要求，那么该网站就包含了视听节目，也就说明该网站包含了视听节目。

[0065] 上述方案设计的方法，通过对四个特点位置布局的标记数据进行统计来进行视听节目的判断，使得对视听节目的识别更加准确。

[0066] 可选地，如图4所示，步骤101中的多组数据集还包括链接数据集，该链接数据集为网页数据中的链接节点数据集合，在步骤101之后，该方法还包括：

[0067] 步骤107：对链接数据集进行链接文本特征检查，判断链接数据集中的内容是否符合链接文本特征要求，转到步骤109。

[0068] 步骤109：若链接数据集中的内容符合链接文本特征要求，则确定目标网站中包含视听节目。

[0069] 这里需要说明的是，在步骤101之后的进行布局特征检查的步骤103-步骤105以及进行链接文本特征检查的步骤107-109的顺序可为以下顺序：可先进行布局特征检查，再进行链接文本特征检查；也可先进行链接文本特征检查，再进行布局特征检查；也可以同时进行布局特征检查和链接文本特征检查。

[0070] 上述方案设计的方法，在布局特征检查的基础上加入链接文本特征检查，增加了特征检查方式，使得对网页数据内视听节目的识别更加准确。

[0071] 其中，如图5所示，对于步骤107中的对链接数据集进行链接文本特征检查，判断链接数据集中的内容是否符合链接文本特征要求，包括：

[0072] 步骤1071：去除链接数据集中的HTML标签内容，转到步骤1073。

[0073] 步骤1073：对链接数据集中剩余的文本内容进行分词，根据分词后的文本内容建立词典，转到步骤1075。

[0074] 步骤1075：获取词典中的词语出现预设判别词的统计结果，该统计结果包括词频、出现个数以及对应词语的长度，转到步骤1077。

[0075] 步骤1077：判断该统计的结果是否在预设的阈值范围内，转到步骤1079。

[0076] 步骤1079：若该统计的结果在预设的阈值范围内，则确定该链接数据集中的内容符合链接文本特征要求。

[0077] 对于上述步骤1075，其具体实现方式可为：对词典中出现的词依据判别词进行词频、出现个数以及对应词语的长度的统计，其中判别词可分为两类：

[0078] 第一类：“电影”、“电视剧”、“直播”、“点播”、“视频”、“片花”、“综艺”、“剧集”、“动漫”、“影视”、“纪录片”、“美剧”、“日剧”、“韩剧”、“港剧”、“脱口秀”、“网络剧”、“花絮”、“片库”以及“预告片”等。

[0079] 第二类：“电视剧”、“点播”、“直播”、“视频”、“视频”、“预告片”、“视频新闻”、“新闻视频”以及“原创视频”等。

[0080] 对于步骤1077，其具体可为：对第一类判别词，进行长度、词频以及出现次数进行评估，对于以上3个条件均满足超过阈值要求的，判别为包含视听节目；对于第二类判别词，进行长度和次数进行评估，对于在预设的阈值范围内的，判别为包含视听节目。

[0081] 上述方案设计的方法，描述了实现链接文本特征检查的具体方式，通过对网页数据中词语出现判别词的统计结果来进行特征判断，使得对视听节目的判断更加准确。

[0082] 应当理解的是，在其他实施例中，判别词除了上述两类，还可以包括另外类别。

[0083] 可选地，如图6所示，在步骤107之前，该方法还包括：

[0084] 步骤106：判断该链接数据集是否适用于链接文本特征检查，若是，则转到步骤107。

[0085] 对于步骤106，其具体实现方式可为：

[0086] 提取链接数据集中的每一条数据，判断每一条数据中的<a>标签下是否包含<img>标签，将计算结果保存。如果不包含<img>标签的<a>所占比例超过设定的阈值，那么判定该链接数据集适用于文本特征检查。

[0087] 可选地，如图7所示，在步骤103判断分区数据集中的内容是否符合布局特征要求之后，该方法包括：

[0088] 步骤111：若分区数据集中的内容不符合布局特征要求，则对网页数据的所有内容进行视频文件检查以及播放器检查，转到步骤113。

[0089] 步骤113：判断网页数据的所有内容是否符合视频文件检查要求以及播放器检查要求，转到步骤115。

[0090] 步骤115：若网页数据的所有内容符合视频文件检查要求和/或网页数据的所有内容符合播放器检查要求，则确定目标网站中包含视听节目。

[0091] 这里需要说明的是，对于步骤111-115的执行前提可以是：在分区数据集的内容不符合布局特征要求之后；如果方案中包含有链接数据集时，则执行前提可以是，在链接数据集的内容不符合链接文本特征要求以及分区数据集的内容也不符合布局特征要求之后，也就是说在两种情况都不符合要求之后才执行步骤111-步骤115。

[0092] 上述方案设计的方法，在布局特征要求不符合时，则进行对网页数据的所有内容进行后续的判断，使得判断网页是否包含视听节目的准确率大大提高，并且后续方法具有极强的通用性。

[0093] 其中，对于判断网页数据的所有内容是否符合视频文件检查要求，包括：

[0094] 对网页数据所有内容的各个节点进行视频文件检查；判断每个节点内容是否包含视频类型的文件；若存在任意一个节点内容包含视频类型的文件，则网页数据的所有内容符合视频文件检查要求。

[0095] 对于上述方案，其具体实现方式可如下：

[0096] 沿DOM树对网页数据所有内容逐个节点检查是否包含mp4,flv,m4v,m3u8,wmv,ts等类型的文件，当存在任意一个节点内容包含上述任意一种类型的文件，则网页数据的所有内容符合视频文件检查要求，则判定该网页包含视听节目。

[0097] 对于判断网页数据的所有内容是否符合播放器检查要求，包括：

[0098] 对所述网页数据所有内容逐条进行播放器检查；判断所述网页数据所有内容中是否存在播放器链接和/或引入播放器文件；若存在，则所述网页数据的所有内容符合播放器检查要求。

[0099] 对于上述方案，其具体实现方式可如下：

[0100] 沿DOM树对网页数据所有内容逐条检查页面中是否带有video.js或者

ckplayer.js等引用，以及，沿DOM树对网页数据所有内容逐条检查页面中带有video标签，source标签(类型相当于video的视频类型)，当存在任意一条内容包含上述任意一种类型时，则网页数据的所有内容符合播放器检查要求，则判定该网页包含视听节目。

[0101] 第二实施例

[0102] 如图8所示，本申请提供一种视听节目识别装置，该装置包括：

[0103] 获取模块201，用于获取目标网站的网页数据。

[0104] 分解模块202，用于将网页数据分解为多组数据集，多组数据集包括分区数据集，该分区数据集为网页数据中的分区节点数据集合。

[0105] 检查模块203，用于对分区数据集进行布局特征检查。

[0106] 判断模块204，用于判断分区数据集中的内容是否符合布局特征要求。

[0107] 确定模块205，用于在判断模块204分区数据集中的内容符合布局特征要求之后，确定目标网站中包含视听节目。

[0108] 上述方案设计的装置，通过对分区数据集进行布局特征检查，判断其是否满足特定的布局特征要求，使得对网页中是否包含视听节目的识别更加准确。

[0109] 在第二实施例的可选实施方式中，所述装置还包括：

[0110] 去除模块206，用于去除分区数据集中HTML标签内容。

[0111] 提取模块207，用于提取分区数据集中每个分区节点的位置信息，并根据位置信息以及位置信息对应的分区节点构建布局特征，该布局特征包括特征位置。

[0112] 标记模块208，用于对特征位置中包含预设格式的数据信息进行标记。

[0113] 判断模块204，还用于判断所述分区数据集的标记数据个数占所述分区数据集的总体个数的比率是否在预设的阈值范围内。

[0114] 确定模块205，还用于在判断模块204判断分区数据集的标记数据个数占分区数据集的总体个数的比率在预设的阈值范围内，确定分区数据集中的内容符合布局特征要求。

[0115] 在第二实施例的可选实施方式中，多组数据集还包括链接数据集，该链接数据集为所述网页数据中的链接节点数据集合。

[0116] 检查模块203，还用于对链接数据集进行链接文本特征检查。

[0117] 判断模块204，还用于判断链接数据集中的内容是否符合链接文本特征要求。

[0118] 确定模块205，还用于在判断模块204判断链接数据集中的内容符合链接文本特征要求时，确定目标网站中包含视听节目。

[0119] 在第二方面的可选实施方式中，去除模块206，还用于去除链接数据集中HTML标签内容。

[0120] 分词模块209，用于对链接数据集中剩余的文本内容进行分词。

[0121] 构建模块210，用于根据分词后的文本内容建立词典。

[0122] 获取模块201，还用于获取词典中的词语出现预设判别词的统计结果，该统计结果包括词频、出现个数以及对应词语的长度。

[0123] 判断模块204，还用于判断统计的结果是否在预设的阈值范围内。

[0124] 确定模块205，还用于在判断模块204判断统计的结果在预设的阈值范围内时，确定链接数据集中的内容符合链接文本特征要求。

[0125] 在第二方面的可选实施方式中，检查模块203，还用于在确定模块确定所述分区数

据集中的内容不符合布局特征要求之后，对所述网页数据的所有内容进行视频文件检查以及播放器检查。

[0126] 判断模块204，还用于判断网页数据的所有内容是否符合视频文件检查要求以及播放器检查要求。

[0127] 确定模块205，还用于在判断模块204判断网页数据的所有内容符合视频文件检查要求和/或所述媒体数据集中的内容符合播放器检查要求时，确定目标网站中包含视听节目。

[0128] 第三实施例

[0129] 如图9所示，本申请提供一种电子设备，包括：处理器301，以及分别与处理器连接的存储器302和通信模块303，存储器302存储有处理器301可执行的机器可读指令，通信模块303用于与外部设备进行通信传输；当所述计算设备运行时，处理器301执行所述机器可读指令，以执行时执行第一实施例、第一实施例的任一可选的实现方式中的所述方法。

[0130] 本申请提供一种计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行第一实施例、第一实施例的任一可选的实现方式中的所述方法。

[0131] 本申请提供一种计算机程序产品，所述计算机程序产品在计算机上运行时，使得计算机执行第一实施例、第一实施例的任一可选的实现方式中的所述方法置。

[0132] 以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应所述以权利要求的保护范围为准。

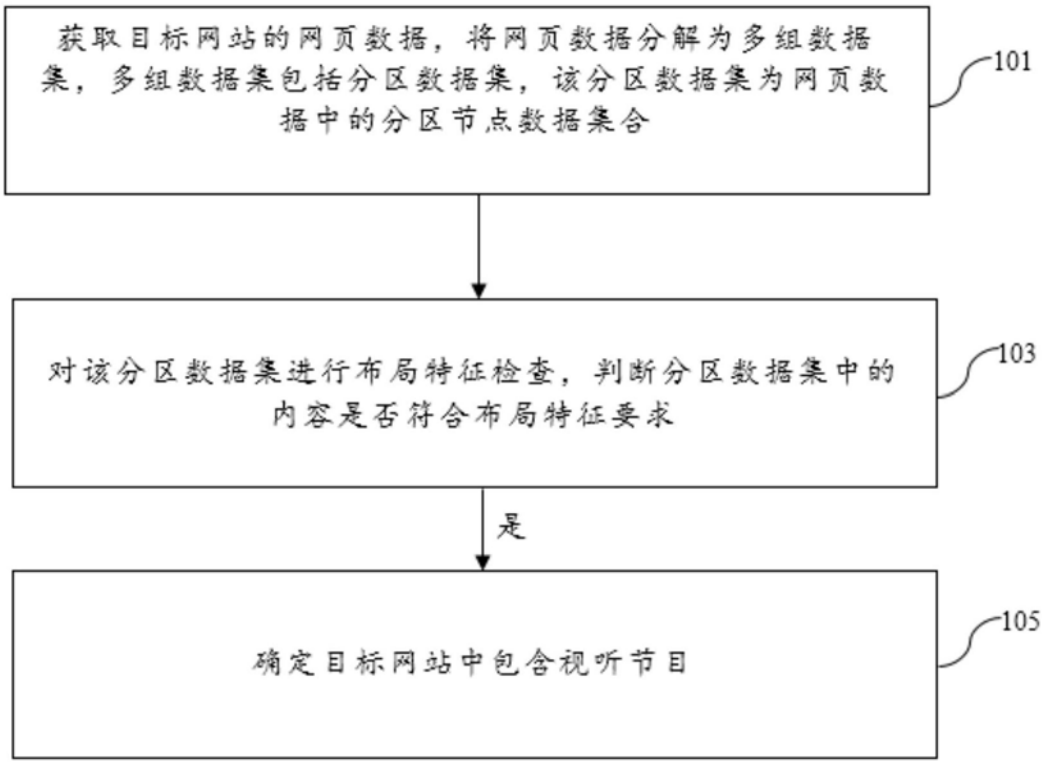

获取目标网站的网页数据，将网页数据分解为多组数据集，多组数据集包括分区数据集，该分区数据集为网页数据中的分区节点数据集合
101
对该分区数据集进行布局特征检查，判断分区数据集中的内容是否符合布局特征要求
103
是
确定目标网站中包含视听节目
105

图1

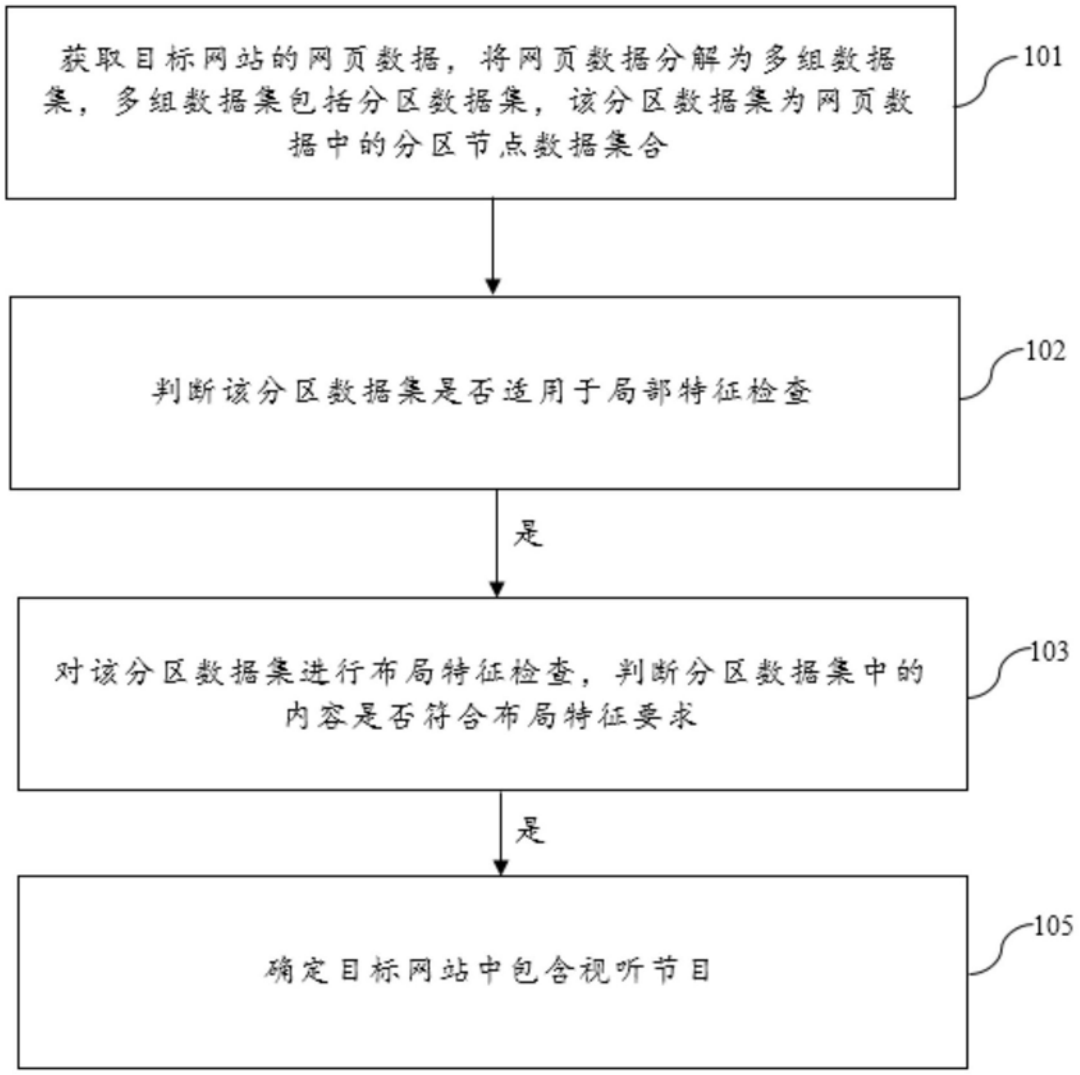

获取目标网站的网页数据，将网页数据分解为多组数据集，多组数据集包括分区数据集，该分区数据集为网页数据中的分区节点数据集合
101
判断该分区数据集是否适用于局部特征检查
102
是
对该分区数据集进行布局特征检查，判断分区数据集中的内容是否符合布局特征要求
103
是
确定目标网站中包含视听节目
105

图2

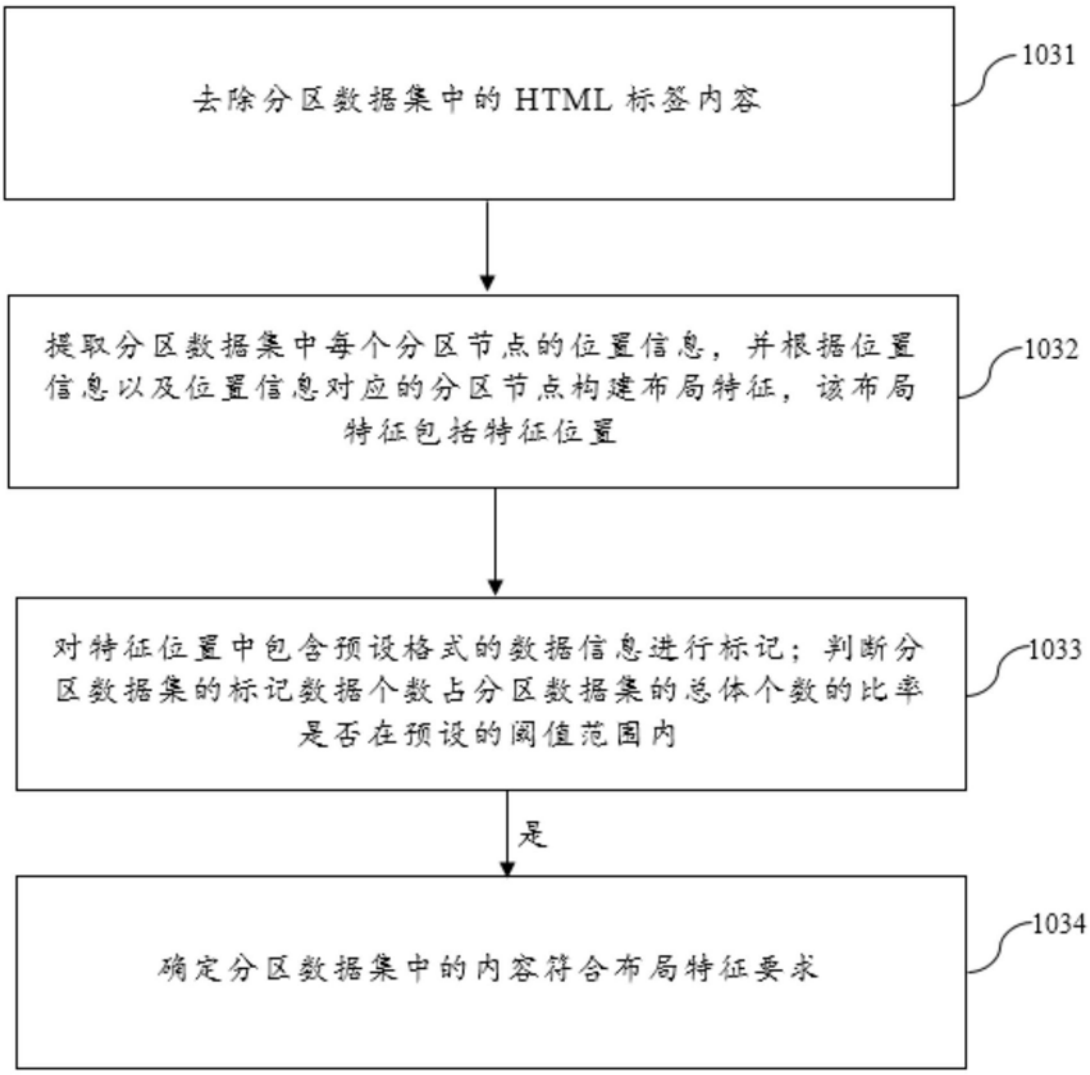

去除分区数据集中的HTML标签内容
1031
提取分区数据集中每个分区节点的位置信息，并根据位置信息以及位置信息对应的分区节点构建布局特征，该布局特征包括特征位置
1032
对特征位置中包含预设格式的数据信息进行标记；判断分区数据集的标记数据个数占分区数据集的总体个数的比率是否在预设的阈值范围内
1033
是
确定分区数据集中的内容符合布局特征要求
1034

图3

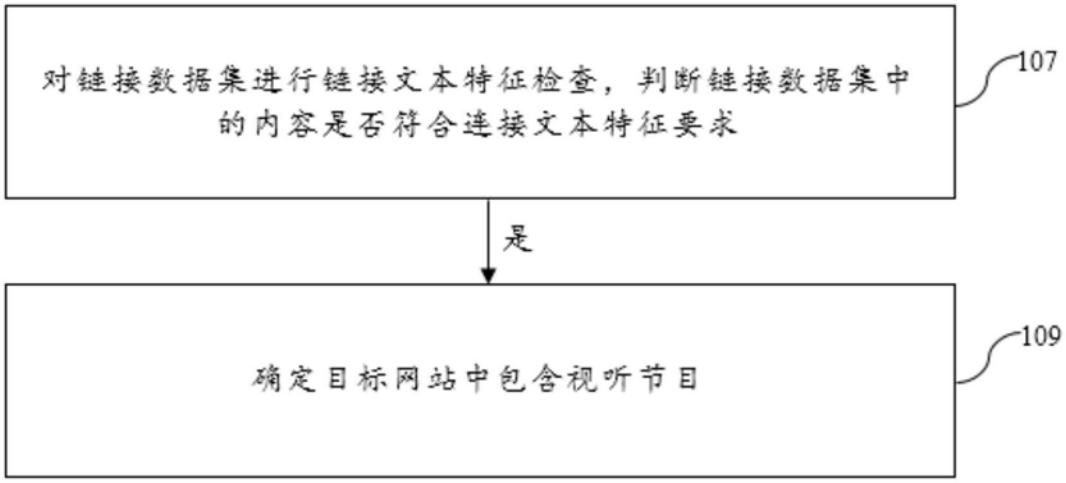

对链接数据集进行链接文本特征检查，判断链接数据集中的内容是否符合连接文本特征要求
107
是
确定目标网站中包含视听节目
109

图4

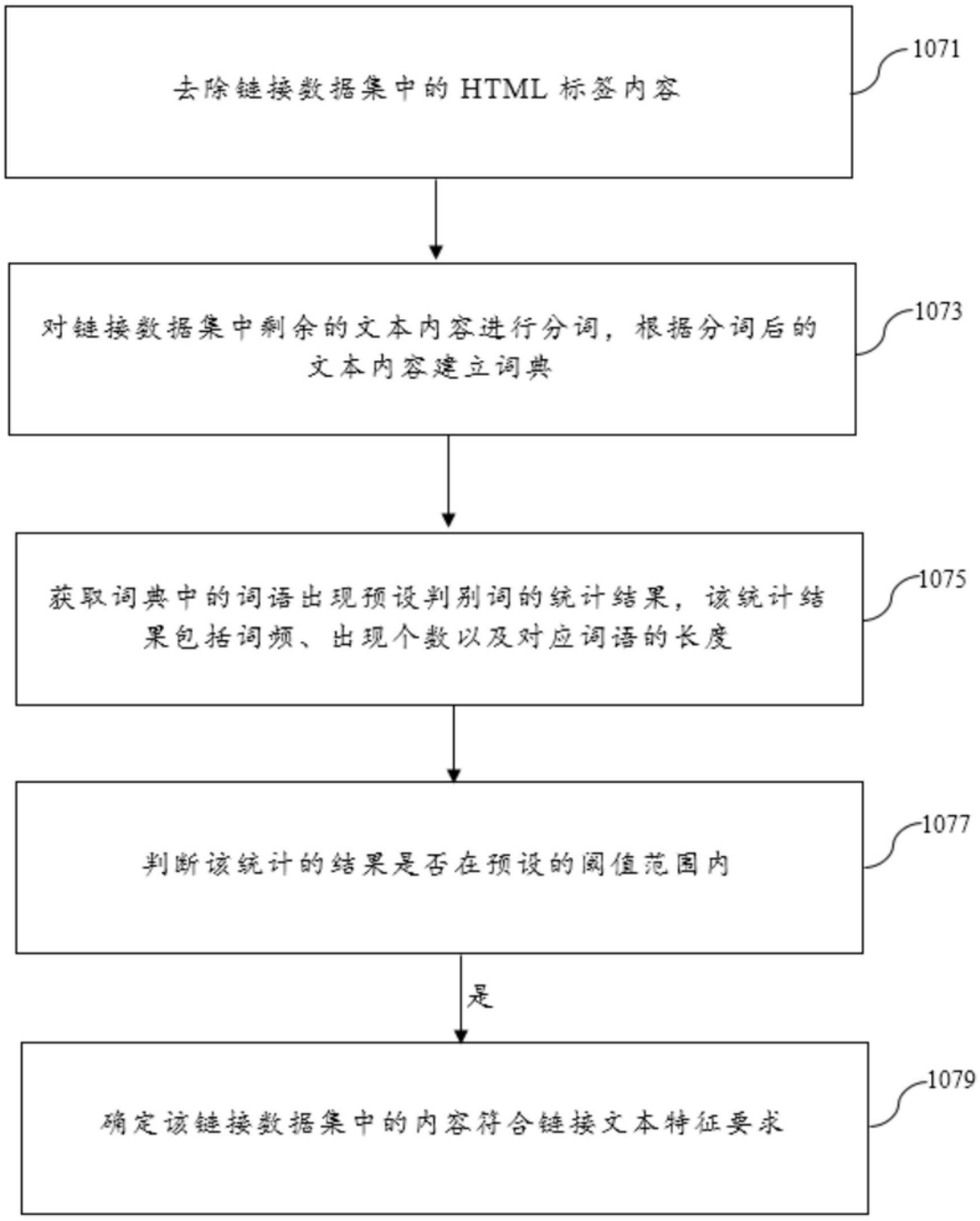

去除链接数据集中的HTML标签内容
1071
对链接数据集中剩余的文本内容进行分词，根据分词后的文本内容建立词典
1073
获取词典中的词语出现预设判别词的统计结果，该统计结果包括词频、出现个数以及对应词语的长度
1075
判断该统计的结果是否在预设的阈值范围内
1077
是
确定该链接数据集中的内容符合链接文本特征要求
1079

图5

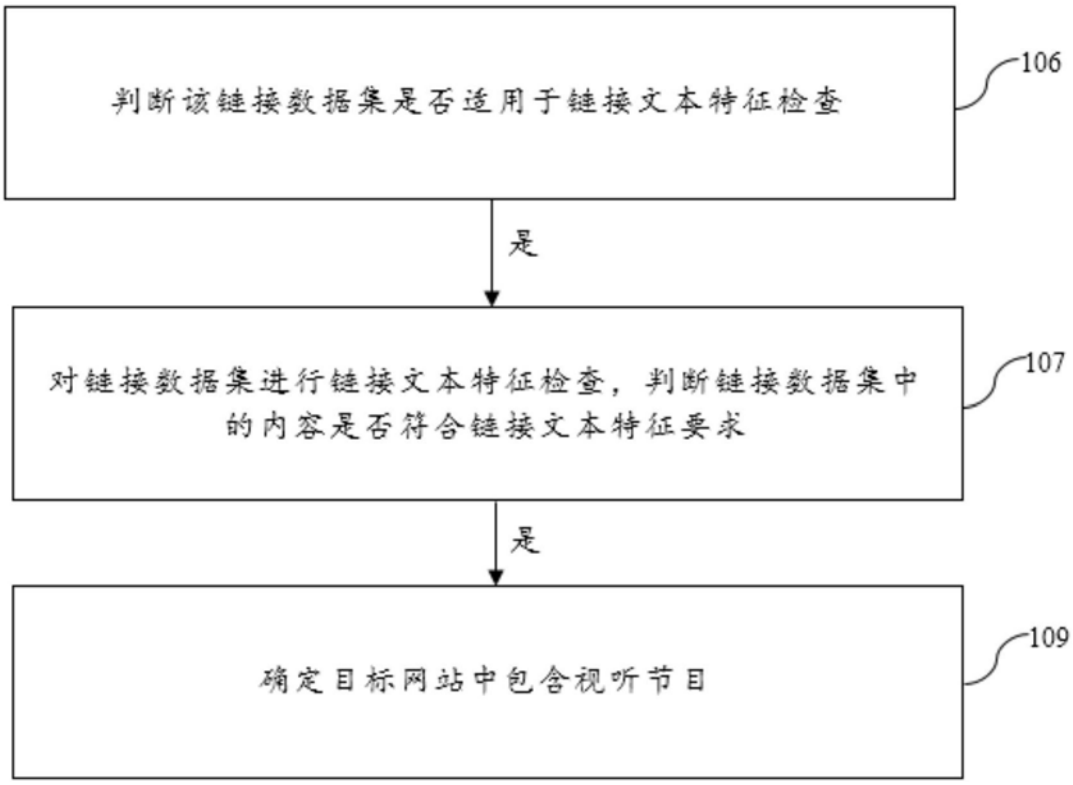

判断该链接数据集是否适用于链接文本特征检查
106
是
对链接数据集进行链接文本特征检查，判断链接数据集中的内容是否符合链接文本特征要求
107
是
确定目标网站中包含视听节目
109

图6

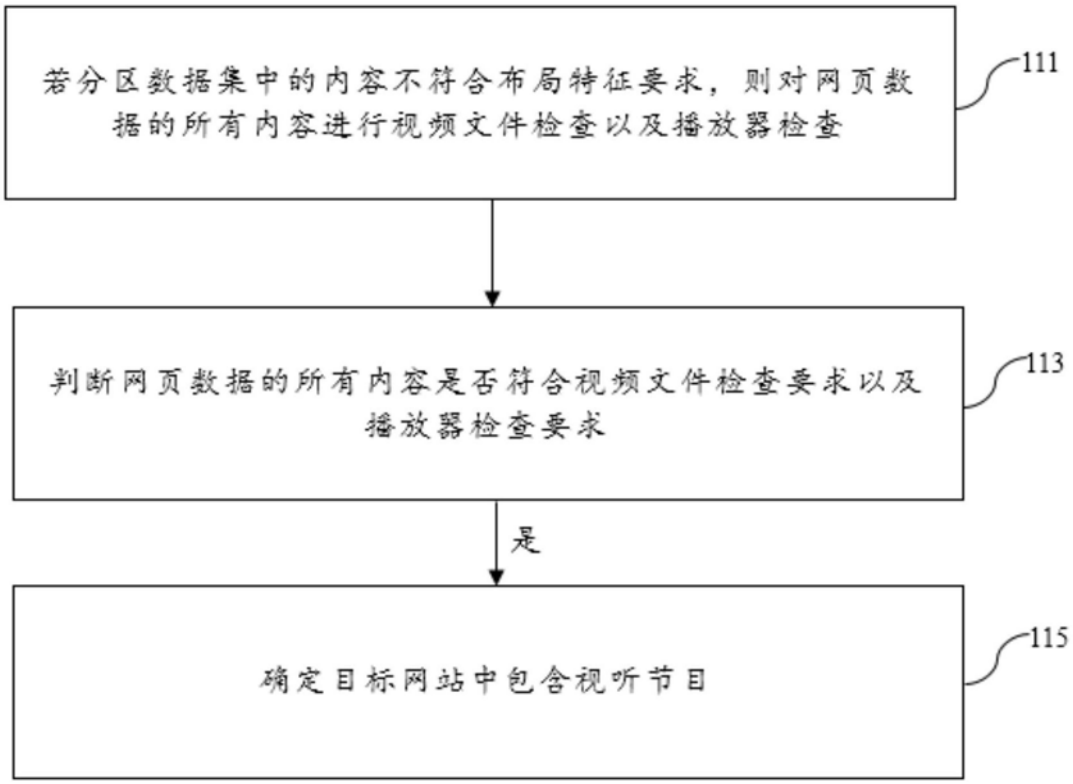

若分区数据集中的内容不符合布局特征要求，则对网页数据的所有内容进行视频文件检查以及播放器检查
111
判断网页数据的所有内容是否符合视频文件检查要求以及播放器检查要求
113
是
确定目标网站中包含视听节目
115

图7

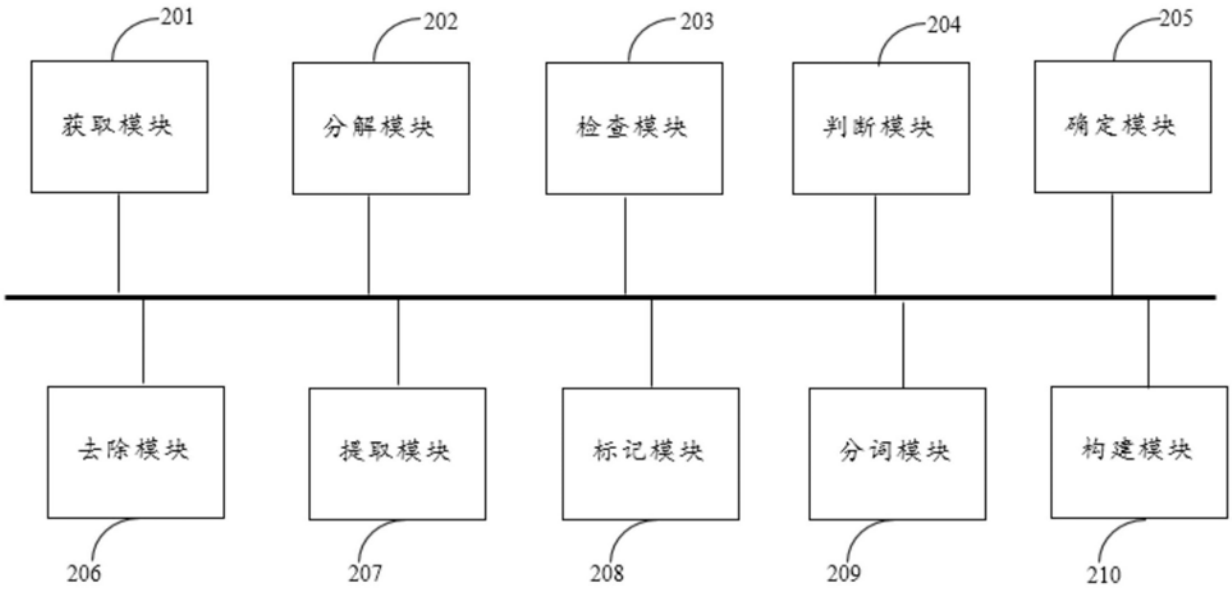

201
获取模块
202
分解模块
203
检查模块
204
判断模块
205
确定模块
去除模块
206
提取模块
207
标记模块
208
分词模块
209
构建模块
210

图8

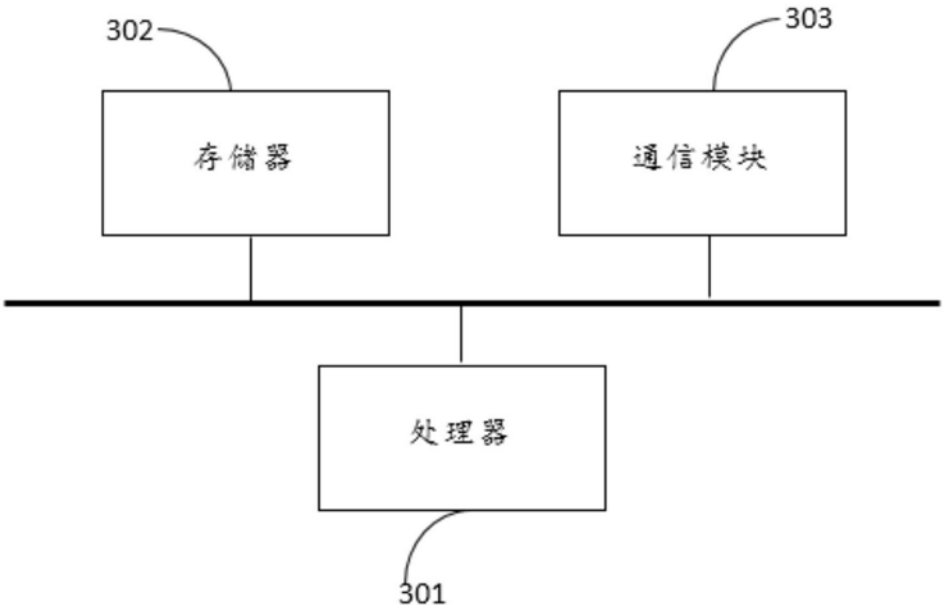

302
存储器
303
通信模块
处理器
301

图9