

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 6 月 18 日 (18.06.2020)



(10) 国际公布号
WO 2020/119064 A1

(51) 国际专利分类号:
G06F 16/958 (2019.01)

(21) 国际申请号: PCT/CN2019/092551

(22) 国际申请日: 2019 年 6 月 24 日 (24.06.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811526834.0 2018 年 12 月 13 日 (13.12.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 吴壮伟(WU, Zhuangwei); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市精英专利事务所(SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路 6009 号绿景广场 B 栋 20 层 B, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) **Title:** METHOD AND DEVICE FOR STORING INTERNET INFORMATION IN LINKED MANNER, COMPUTER APPARATUS AND STORAGE MEDIUM

(54) 发明名称: 互联网信息链式存储方法、装置、计算机设备及存储介质

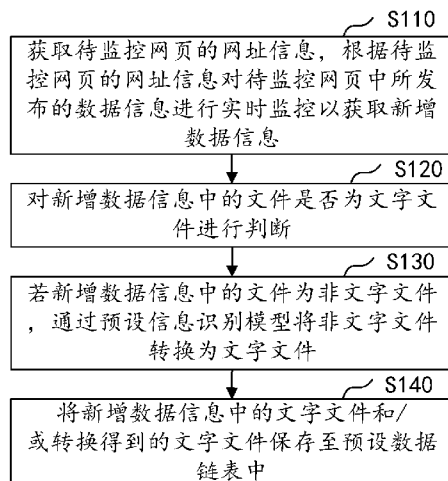


图 1

S110 ACQUIRE ADDRESS INFORMATION OF A WEBPAGE TO BE MONITORED, AND PERFORM, ACCORDING TO THE ADDRESS INFORMATION OF THE WEBPAGE, REAL-TIME MONITORING OF DATA INFORMATION PUBLISHED ON THE WEBPAGE SO AS TO ACQUIRE NEWLY ADDED DATA INFORMATION

S120 DETERMINE WHETHER A FILE IN THE NEWLY ADDED DATA INFORMATION IS A TEXT FILE

S130 IF THE FILE IS A NON-TEXT FILE, CONVERT THE NON-TEXT FILE TO A TEXT FILE BY MEANS OF A PRESET INFORMATION RECOGNITION MODEL

S140 STORE, IN A PRESET LINKED DATA LIST, A TEXT FILE IN THE NEWLY ADDED DATA INFORMATION AND/OR A TEXT FILE OBTAINED FROM CONVERSION

(57) **Abstract:** The present application discloses a method and device for storing Internet information in a linked manner, a computer apparatus and a storage medium. The method comprises: acquiring address information of a webpage to be monitored, and performing, according to the address information of the webpage, real-time monitoring of data information published on the webpage so as to acquire newly added data information; determining whether a file in the newly added data information is a text file; if the file is a non-text file, converting the non-text file to a text file by means of a preset information recognition model; and storing, in a preset linked data list, a text file in the newly added data information and/or a text file obtained from conversion.

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本申请公开了互联网信息链式存储方法、装置、计算机设备及存储介质。方法包括: 获取待监控网页的网址信息, 根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息; 对新增数据信息中的文件是否为文字文件进行判断; 若新增数据信息中的文件为非文字文件, 通过预设信息识别模型将非文字文件转换为文字文件; 将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

发明名称：互联网信息链式存储方法、装置、计算机设备及存储介质

本申请要求于2018年12月13日提交中国专利局、申请号为201811526834.0、申请名称为“互联网信息链式存储方法、装置、计算机设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，尤其涉及一种互联网信息链式存储方法、装置、计算机设备及存储介质。

背景技术

互联网中各网页上保存有海量的数据信息，且新增数据信息会逐渐更替网页中已保存的数据信息，造成网页中的数据信息发生更迭变化的情况，因而现有对互联网中的数据信息进行存储方法无法对互联网上已删除或已修改的数据信息进行获取，在司法实践中对互联网上所发布的相关数据信息进行取证存在极大的困难。因此，现有的数据信息存储方法无法获取已删除数据信息。

发明内容

本申请实施例提供了一种互联网信息链式存储方法、装置、计算机设备及存储介质，旨在解决现有技术中数据信息存储方法无法获取已删除数据信息的问题。

第一方面，本申请实施例提供了一种互联网信息链式存储方法，其包括：

获取待监控网页的网址信息，根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息；对新增数据信息中的文件是否为文字文件进行判断；若新增数据信息中的文件为非文字文件，通过预设信息识别模型将非文字文件转换为文字文件；将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

第二方面，本申请实施例提供了一种互联网信息链式存储装置，其包括：

网页监控单元，用于获取待监控网页的网址信息，根据待监控网页的网址

信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息；判断单元，用于对新增数据信息中的文件是否为文字文件进行判断；信息转换单元，用于若新增数据信息中的文件为非文字文件，通过预设信息识别模型将非文字文件转换为文字文件；信息存储单元，用于将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

第三方面，本申请实施例又提供了一种计算机设备，其包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述第一方面所述的互联网信息链式存储方法。

第四方面，本申请实施例还提供了一种计算机可读存储介质，其中所述计算机可读存储介质存储有计算机程序，所述计算机程序当被处理器执行时使所述处理器执行上述第一方面所述的互联网信息链式存储方法。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例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。

图1为本申请实施例提供的互联网信息链式存储方法的流程示意图；

图2为本申请实施例提供的互联网信息链式存储方法的子流程示意图；

图3为本申请实施例提供的互联网信息链式存储方法的另一子流程示意图；

图4为本申请实施例提供的互联网信息链式存储方法的另一子流程示意图；

图5为本申请实施例提供的互联网信息链式存储方法的另一子流程示意图；

图6为本申请实施例提供的互联网信息链式存储装置的示意性框图；

图7为本申请实施例提供的互联网信息链式存储装置的子单元示意性框图；

图8为本申请实施例提供的互联网信息链式存储装置的另一子单元示意性框图；

图9为本申请实施例提供的互联网信息链式存储装置的另一子单元示意性框图；

图10为本申请实施例提供的互联网信息链式存储装置的另一子单元示意性框图；

图 11 为本申请实施例提供的计算机设备的示意性框图。

具体实施方式

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。

应当理解，当在本说明书和所附权利要求书中使用时，术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

还应当理解，在此本申请说明书中所使用的术语仅仅是出于描述特定实施例的目的而并不意在限制本申请。如在本申请说明书和所附权利要求书中所使用的那样，除非上下文清楚地指明其它情况，否则单数形式的“一”、“一个”及“该”意在包括复数形式。

还应当进一步理解，在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。

请参阅图 1，图 1 是本申请实施例提供的互联网信息链式存储方法的流程示意图。该互联网信息链式存储方法应用于具有信息存储功能的终端设备中，例如台式电脑、笔记本电脑、平板电脑或手机等。

如图 1 所示，该方法包括步骤 S110~S140。

S110、获取待监控网页的网址信息，根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息。

获取待监控网页的网址信息，根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息。其中，待监控网页信息为用户所输入的待监控网页的网址信息，待监控网页可以是微博、微信、企业网址、政府网站等所有在互联网上所发布的数据信息，发布人可以是个人、企业、组织或政府部门，例如对某一名人在其微博中所发布的信息进行监控，则待监控网页信息即是该名人微博网页的网址信息。

待监控网页中所发布的数据信息中可包含多种格式的文件，例如文字格式的信息、视频格式的信息、音频格式、图片格式的信息等。由于待监控网页中所发布的数据信息为实时发布，因此需对待监控网页进行监控以实时获取该网页中所发布的最新的数据信息。

在一实施例中，如图2所示，步骤S110包括子步骤S111、S112和S113。

S111、若监控到待监控网页中发布数据信息，根据待监控网页的网址信息及所述数据信息的发布人生成发布源信息。

若监控到待监控网页中发布数据信息，根据待监控网页的网址信息及所述数据信息的发布人生成发布源信息。为获取新增数据信息的发布人，需根据待监控网页的网址信息及所述数据信息的发布人生成相应的发布源信息。发布源信息中包括待监控网页的网址信息以及该数据信息的发布人，待监控网页的网址信息也即是用户所输入的待监控网页信息；发布人也即是发布该新增数据信息的发布主体，发布人可以是个人、企业、组织或政府部门。

S112、根据所述数据信息的发布时间生成发布时间戳。

根据所述数据信息的发布时间生成新增数据信息的发布时间戳，为对新增数据信息的发布时间进行记录，需根据数据信息的发布时间生成相应的发布时间戳，发布时间戳生成后无法被更改，也即是确保新增数据信息的发布时间被及时记录且无法更改。

例如，待监控网页为某一名人的微博网页，每一条微博信息的发布均包含一个发布时间，获取该微博信息的发布时间即为相应新增数据信息的发布时间戳。

S113、获取所述发布数据信息中的所有文件及发布源信息、发布时间戳以得到新增数据信息。

获取所述发布数据信息中的所有文件及发布源信息、发布时间戳以得到新增数据信息。获取所发布数据信息中的所有文件作为新增数据信息，并获取所得到的发布源信息及发布时间戳即可得到新增数据信息，新增数据信息中可包含一个或多个文件。

例如，待监控网页为某一名人的微博网页，该名人微博网页中发布了一段文字信息及一个视频信息，则获取所发布的文字信息及视频信息、发布源信息、发布时间戳得到包含一个文字文件及一个视频文件的新增数据信息。

S120、对新增数据信息中的文件是否为文字文件进行判断。

对新增数据信息中的文件是否为文字文件进行判断，为对新增数据信息中各种格式的文件进行保存，需先对新增数据信息中的文件是否为文字文件进行判断。具体的，通过获取新增数据信息中各文件的格式信息以判断该文件是否为文字文件。

获取新增数据信息中各文件的格式信息。每一个文件都拥有各自的格式信息，不同文件的格式信息与相应的类型相匹配，通过格式信息即可对文件的具体类型进行判断。根据各文件的格式信息对各文件是否为文字文件进行判断，通过各文件的格式信息即可对文件的具体类型进行判断。

例如，若某一文件的格式信息为 txt、string，则该文件为文字文件；若某一文件的格式信息为 wav、mp3、wma，则该文件为音频文件；若某一文件的格式信息为 avi、flv、rmvb，则该文件为视频文件。

S130、若新增数据信息中的文件为非文字文件，通过信息识别模型将非文字文件转换为文字文件。

若新增数据信息中的文件为非文字文件，通过预设信息识别模型将非文字文件转换为文字文件。具体的，非文字文件也是文件中的一种，非文字文件包括音频文件、视频文件、图片。信息识别模型即是对非文字文件进行识别及转换的模型，其中，信息识别模型中包括音频识别模型及图片识别模型。

在一实施例中，如图 3 所示，步骤 S130 包括子步骤 S131、S132 和 S133。

S131、获取所述非文字文件的格式信息并判断该文件是否为音频文件，若该文件为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件。

获取文件的格式信息并判断该文件是否为音频文件，若该文件为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件。通过音频识别模型即可对音频文件中的语音信息进行识别并转换，以得到相应的包含文字信息的文字文件，进行转换后每一个音频文件均对应得到一个文字文件。其中，音频识别模型包括声学模型、语音特征词典及语义解析模型。

在一实施例中，如图 4 所示，步骤 S131 包括子步骤 S1311 和 S1312。

S1311、根据音频识别模型中的声学模型对音频文件中的语音信息进行切分以得到语音信息中所包含的多个音素。

根据音频识别模型中的声学模型对音频文件中的语音信息进行切分以得到语音信息中所包含的多个音素。具体的，语音信息由多个字符发音的音素而组成，一个字符的音素包括该字符发音的频率和音色。声学模型中包含所有字符发音的音素，通过将语音信息与声学模型中所有的音素进行匹配，即可对语音信息中单个字符的音素进行切分，通过切分最终得到该语音信息中所包含的多个音素。

S1312、根据音频识别模型中的语音特征词典对所得到的音素进行匹配以将所有音素转换为拼音信息。

根据音频识别模型中的语音特征词典对所得到的音素进行匹配，即可将所有音素转换为拼音信息。语音特征词典中包含所有字符拼音对应的音素信息，通过将所得到的音素与字符拼音对应的音素信息进行匹配，即可将单个字符的音素转换为语音特征词典中与该音素相匹配的字符拼音，以实现将语音信息中所包含的所有音素转换为拼音信息。

S1313、根据音频识别模型中的语义解析模型对所得到的拼音信息进行语义解析以得到包含文字信息的文字文件。

根据音频识别模型中的语义解析模型对所得到的拼音信息进行语义解析，以实现将拼音信息转换为对应的文字文件。语义解析模型中包含拼音信息与文字信息之间所对应的映射关系，通过语义解析模型中所包含的映射关系即可对所得到的拼音信息进行语义解析以将拼音信息转换为包含文字信息的文字文件。

S132、获取所述非文字文件的格式信息并判断该文件是否为图片，若该文件为图片则通过信息识别模型中图片识别模型的对该文件进行识别以得到相应的文字文件。

获取所述非文字文件的格式信息并判断该文件是否为图片，若该文件为图片则通过信息识别模型中图片识别模型的对该文件中所包含的文字进行识别以得到相应的文字文件。具体的，文字模板即是用对图片中文字进行识别的模板信息，一个文字模板与图片中的一个文字相对应，一个文字模板包含相应文字所对应的多种字体，图片中的文字均能与对应文字模板中的某一种字体相匹配，通过图片识别模型中的文字模板与图片进行匹配，即可对该图片中所包含的文字进行识别以得到相应的文字文件。

S133、获取所述非文字文件的格式信息并判断该文件是否为视频文件，若该文件为视频文件则通过信息识别模型中的音频识别模型及图片识别模型对该文件进行识别以得到相应的文字文件。

获取所述非文字文件的格式信息并判断该文件是否为视频文件，若该文件为视频文件则通过信息识别模型中的音频识别模型及图片识别模型对该文件进行识别以得到相应的文字文件。若该文件为视频文件，则先获取该视频文件中的语音信息，并通过音频识别模型即可对该视频文件中的语音信息进行识别并转换以得到该语音信息对应的文字信息，具体的识别及转换方法与所述步骤 S131 相同；获取该视频文件中所包含的每一帧图片，并通过图片识别模型对该视频文件中所包含的每一帧图片进行识别，以得到每一帧图片中所包含的文字信息，具体的识别方法与所述步骤 S132 相同。获取视频文件的语音信息所对应的文字信息，及该视频文件中每一帧图片所包含的文字信息，即可最终得到该视频文件所对应的文字文件，也即是进行转换后每一个视频文件均对应得到一个文字文件。

此外，若非文字文件不是视频文件、音频文件及图片中的任意一种，信息识别模型无法对该文件进行处理，则生成报警提示信息以提示用户无法对该文件进行处理。

S140、将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中以对新增数据信息进行保存。数据链表为终端设备中所预设的用于存储信息的数据库，具体的，数据链表为根据时间轴对新增数据信息中所包含的文字文件进行存储的数据库，数据链表中所存储的数据信息的逻辑顺序是通过数据链表中的指针链接次序实现的，在本实施例中以新增数据信息的发布时间戳作为数据链表的逻辑顺序，也即是通过时间信息为指针链接次序将新增数据信息中的文字文件存储至数据链表中。通过以时间顺序作为链表的逻辑顺序对新增数据信息进行存储，用户可通过数据链表获取到以时间信息为顺序的文字文件列表，数据链表所存储的信息具有无法删除的特性。

此外，由于将其他非文字文件转换为包含文字信息的文字文件进行存储，因此可极大地压缩对相应数据信息进行存储所需的存储空间，便于用户进行使用。

在一实施例中，如图 5 所示，步骤 S140 包括子步骤 S141、S142 和 S143。

S141、获取待监控网页中新增数据信息的发布源信息及发布时间戳。

获取待监控网页中新增数据信息的发布源信息及发布时间戳。为方便对新增数据信息进行存储以后期对所存储的信息数据信息进行检索，需获取新增数据信息的发布源信息及发布时间戳，具体的，发布源信息中包括待监控网页的网址信息以及该数据信息的发布人。

S142、根据发布源信息将新增数据信息中所包含的文字文件和/或转换得到的文字文件进行分类。

根据发布源信息将新增数据信息中所包含的文字文件和/或转换得到的文字文件进行分类。具体的，为对新增数据信息进行分类存储，需根据发布源信息中的发布人对新增数据信息进行分类，其中，一个发布人对应一个类别，每一个类别与数据链表中的一个子链表相对应，相同发布人所发布的新增数据信息则分至该发布源信息对应的子链表中进行保存，通过发布源信息中的发布人，即可将新增数据信息中所包含的文字文件和/或转换得到的文字文件进行分类以对新增数据信息进行分类保存。

S143、根据发布时间戳将新增数据信息中所包含的文字文件和/或转换得到的文字文件保存至预设数据链表中与发布源信息对应的子链表。

根据发布时间戳将新增数据信息中所包含的文字文件和/或转换得到的文字文件保存至预设数据链表中相应的子链表中进行保存。一个发布人对应一个类别，每一个类别与数据链表中的一个子链表相对应，且由于数据链表中的数据信息均即是根据时间轴进行存储，因此需根据新增数据信息的发布时间戳将新增数据信息中所对应的文字文件存储至数据链表中与发布人类别相对应的子链表中，即可实现对新增数据信息进行保存。

由于数据链表所存储的文字文件无法删除和修改，因此可实现对待监控网页中所发布的历史数据信息进行保存，以方便后期进行对相应发布人所发布的历史数据信息进行取证。

通过对网页中所发布的数据信息进行监控并判断其中的文件是否为文字文

件，将非文字文件的文件转换为文字文件，并对所有文字文件存储至数据链表中以实现互联网信息进行链式存储，能够确保所存储的文字文件无法删除和修改，能够方便用户获取互联网上已删除的数据信息以协助用户对相关数据信息进行取证，具有极大的实用价值。

本申请实施例还提供一种互联网信息链式存储装置，该互联网信息链式存储装置用于执行前述互联网信息链式存储方法的任一实施例。具体地，请参阅图 6，图 6 是本申请实施例提供的互联网信息链式存储装置的示意性框图。该互联网信息链式存储装置可以配置于台式电脑、笔记本电脑、平板电脑或手机等终端设备中。

如图 6 所示，互联网信息链式存储装置 100 包括网页监控单元 110、判断单元 120、信息转换单元 130、信息存储单元 140。

网页监控单元 110，用于获取待监控网页的网址信息，根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息。

其他申请实施例中，如图 7 所示，所述网页监控单元 110 包括子单元：发布源信息生成单元 111、发布时间戳生成单元 112 和新增数据信息获取单元 113。

发布源信息生成单元 111，用于若监控到待监控网页中发布数据信息，根据待监控网页的网址信息及所述数据信息的发布人生成发布源信息。

发布时间戳生成单元 112，用于根据所述数据信息的发布时间生成发布时间戳。

新增数据信息获取单元 113，用于获取所述发布数据信息中的所有文件及发布源信息、发布时间戳以得到新增数据信息。

判断单元 120，用于对新增数据信息中的文件是否为文字文件进行判断。

信息转换单元 130，用于若新增数据信息中的文件为非文字文件，通过预设信息识别模型将非文字文件转换为文字文件。

其他申请实施例中，如图 8 所示，所述信息转换单元 130 包括子单元：第一文字文件获取单元 131、第二文字文件获取单元 132 和第三文字文件获取单元 133。

第一文字文件获取单元 131，用于获取所述非文字文件的格式信息并判断该文件是否为音频文件，若该文件为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件。

其他申请实施例中，如图 9 所示，所述第一文字文件获取单元 131 包括子单元：音素切分单元 1311、音素转换单元 1312 和语音解析单元 1313。

音素切分单元 1311，用于根据音频识别模型中的声学模型对音频文件中的语音信息进行切分以得到语音信息中所包含的多个音素。

音素转换单元 1312，用于根据音频识别模型中的语音特征词典对所得到的音素进行匹配以将所有音素转换为拼音信息。

语音解析单元 1313，用于根据音频识别模型中的语义解析模型对所得到的拼音信息进行语义解析以得到包含文字信息的文字文件。

第二文字文件获取单元 132，用于获取所述非文字文件的格式信息并判断该文件是否为图片，若该文件为图片则通过信息识别模型中图片识别模型的对该文件进行识别以得到相应的文字文件。

第三文字文件获取单元 133，用于获取所述非文字文件的格式信息并判断该文件是否为视频文件，若该文件为视频文件则通过信息识别模型中的音频识别模型及图片识别模型对该文件进行识别以得到相应的文字文件。

信息存储单元 140，用于将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

其他申请实施例中，如图 10 所示，所述信息存储单元 140 包括子单元：信息获取单元 141、文件分类单元 142 和文件存储单元 143。

信息获取单元 141，用于获取待监控网页中新增数据信息的发布源信息及发布时间戳。

文件分类单元 142，用于根据发布源信息将新增数据信息中所包含的文字文件和/或转换得到的文字文件进行分类。

文件存储单元 143，用于根据发布时间戳将新增数据信息中所包含的文字文件和/或转换得到的文字文件保存至预设数据链表中与发布源信息对应的子链表。

上述互联网信息链式存储装置可以实现为计算机程序的形式，该计算机程序可以在如图 11 所示的计算机设备上运行。

请参阅图 11，图 11 是本申请实施例提供的计算机设备的示意性框图。

参阅图 11，该计算机设备 500 包括通过系统总线 501 连接的处理器 502、存储器和网络接口 505，其中，存储器可以包括非易失性存储介质 503 和内存存储器 504。该非易失性存储介质 503 可存储操作系统 5031 和计算机程序 5032。该

计算机程序 5032 被执行时,可使得处理器 502 执行互联网信息链式存储方法。该处理器 502 用于提供计算和控制能力,支撑整个计算机设备 500 的运行。该内存储器 504 为非易失性存储介质 503 中的计算机程序 5032 的运行提供环境,该计算机程序 5032 被处理器 502 执行时,可使得处理器 502 执行互联网信息链式存储方法。该网络接口 505 用于进行网络通信,如提供数据信息的传输等。本领域技术人员可以理解,图 11 中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备 500 的限定,具体的计算机设备 500 可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

其中,所述处理器 502 用于运行存储在存储器中的计算机程序 5032,以实现本实施例的互联网信息链式存储方法。

本领域技术人员可以理解,图 11 中示出的计算机设备的实施例并不构成对计算机设备具体构成的限定,在其他实施例中,计算机设备可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。例如,在一些实施例中,计算机设备可以仅包括存储器及处理器,在这样的实施例中,存储器及处理器的结构及功能与图 11 所示实施例一致,在此不再赘述。

应当理解,在本申请实施例中,处理器 502 可以是中央处理单元 (Central Processing Unit, CPU),该处理器 502 还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中,通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

在本申请的另一实施例中提供计算机可读存储介质。该计算机可读存储介质可以为非易失性的计算机可读存储介质。该计算机可读存储介质存储有计算机程序,其中计算机程序被处理器执行时实现本申请实施例的互联网信息链式存储方法。

所述存储介质可以是前述设备的内部存储单元,例如设备的硬盘或内存。所述存储介质也可以是所述设备的外部存储设备,例如所述设备上配备的插接式硬盘,智能存储卡 (Smart Media Card, SMC),安全数字 (Secure Digital, SD) 卡,闪存卡 (Flash Card) 等。进一步地,所述存储介质还可以既包括所述设备

的内部存储单元也包括外部存储设备。

所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，上述描述的设备、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

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，可轻易想到各种等效的修改或替换，这些修改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以权利要求的保护范围为准。

权利要求书

1. 一种互联网信息链式存储方法，包括：

获取待监控网页的网址信息，根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息；

对新增数据信息中的文件是否为文字文件进行判断；

若新增数据信息中的文件为非文字文件，通过预设信息识别模型将非文字文件转换为文字文件；

将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

2. 根据权利要求 1 所述的互联网信息链式存储方法，其中，所述根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息，包括：

若监控到待监控网页中发布数据信息，根据待监控网页的网址信息及所述数据信息的发布人生成发布源信息；

根据所述数据信息的发布时间生成发布时间戳；

获取所述发布数据信息中的所有文件及发布源信息、发布时间戳以得到新增数据信息。

3. 根据权利要求 1 所述的互联网信息链式存储方法，其中，所述通过预设信息识别模型将非文字文件转换为文字文件，包括：

获取所述非文字文件的格式信息并判断该文件是否为音频文件，若该文件为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件；

获取所述非文字文件的格式信息并判断该文件是否为图片，若该文件为图片则通过信息识别模型中图片识别模型的对该文件进行识别以得到相应的文字文件；

获取所述非文字文件的格式信息并判断该文件是否为视频文件，若该文件为视频文件则通过信息识别模型中的音频识别模型及图片识别模型对该文件进行识别以得到相应的文字文件。

4. 根据权利要求 3 所述的互联网信息链式存储方法，其中，所述若该文件

为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件，包括：

根据音频识别模型中的声学模型对音频文件中的语音信息进行切分以得到语音信息中所包含的多个音素；

根据音频识别模型中的语音特征词典对所得到的音素进行匹配以将所有音素转换为拼音信息；

根据音频识别模型中的语义解析模型对所得到的拼音信息进行语义解析以得到包含文字信息的文字文件。

5.根据权利要求2所述的互联网信息链式存储方法，其中，所述将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中，包括：

获取待监控网页中新增数据信息的发布源信息及发布时间戳；

根据发布源信息将新增数据信息中所包含的文字文件和/或转换得到的文字文件进行分类；

根据发布时间戳将新增数据信息中所包含的文字文件和/或转换得到的文字文件保存至预设数据链表中与发布源信息对应的子链表。

6.根据权利要求1所述的互联网信息链式存储方法，其中，所述对新增数据信息中的文件是否为文字文件进行判断，包括：

根据新增数据信息中各文件的格式信息对各文件是否为文字文件进行判断。

7.根据权利要求3所述的互联网信息链式存储方法，其中，所述通过预设信息识别模型将非文字文件转换为文字文件，还包括：

若所述非文字文件不是视频文件、音频文件及图片中的任意一种，生成报警提示信息以提示用户无法对所述文件进行处理。

8.一种互联网信息链式存储装置，包括：

网页监控单元，用于获取待监控网页的网址信息，根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息；

判断单元，用于对新增数据信息中的文件是否为文字文件进行判断；

信息转换单元，用于若新增数据信息中的文件为非文字文件，通过预设信息识别模型将非文字文件转换为文字文件；

信息存储单元，用于将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

9.根据权利要求 8 所述的互联网信息链式存储装置,其中,所述网页监控单元,包括:

发布源信息生成单元,用于若监控到待监控网页中发布数据信息,根据待监控网页的网址信息及所述数据信息的发布人生成发布源信息;

发布时间戳生成单元,用于根据所述数据信息的发布时间生成发布时间戳;

新增数据信息获取单元,用于获取所述发布数据信息中的所有文件及发布源信息、发布时间戳以得到新增数据信息。

10.根据权利要求 8 所述的互联网信息链式存储装置,其中,所述信息转换单元,包括:

第一文字文件获取单元,用于获取所述非文字文件的格式信息并判断该文件是否为音频文件,若该文件为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件;

第二文字文件获取单元,用于获取所述非文字文件的格式信息并判断该文件是否为图片,若该文件为图片则通过信息识别模型中图片识别模型的对该文件进行识别以得到相应的文字文件;

第三文字文件获取单元,用于获取所述非文字文件的格式信息并判断该文件是否为视频文件,若该文件为视频文件则通过信息识别模型中的音频识别模型及图片识别模型对该文件进行识别以得到相应的文字文件。

11.一种计算机设备,包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,其中,所述处理器执行所述计算机程序时实现以下步骤:

获取待监控网页的网址信息,根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息;

对新增数据信息中的文件是否为文字文件进行判断;

若新增数据信息中的文件为非文字文件,通过预设信息识别模型将非文字文件转换为文字文件;

将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

12. 根据权利要求 11 所述的计算机设备,其中,所述根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息,

包括：

若监控到待监控网页中发布数据信息，根据待监控网页的网址信息及所述数据信息的发布人生成发布源信息；

根据所述数据信息的发布时间生成发布时间戳；

获取所述发布数据信息中的所有文件及发布源信息、发布时间戳以得到新增数据信息。

13. 根据权利要求 11 所述的计算机设备，其中，所述通过预设信息识别模型将非文字文件转换为文字文件，包括：

获取所述非文字文件的格式信息并判断该文件是否为音频文件，若该文件为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件；

获取所述非文字文件的格式信息并判断该文件是否为图片，若该文件为图片则通过信息识别模型中图片识别模型的对该文件进行识别以得到相应的文字文件；

获取所述非文字文件的格式信息并判断该文件是否为视频文件，若该文件为视频文件则通过信息识别模型中的音频识别模型及图片识别模型对该文件进行识别以得到相应的文字文件。

14. 根据权利要求 13 所述的计算机设备，其中，所述若该文件为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件，包括：

根据音频识别模型中的声学模型对音频文件中的语音信息进行切分以得到语音信息中所包含的多个音素；

根据音频识别模型中的语音特征词典对所得到的音素进行匹配以将所有音素转换为拼音信息；

根据音频识别模型中的语义解析模型对所得到的拼音信息进行语义解析以得到包含文字信息的文字文件。

15. 根据权利要求 12 所述的计算机设备，其中，所述将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中，包括：

获取待监控网页中新增数据信息的发布源信息及发布时间戳；

根据发布源信息将新增数据信息中所包含的文字文件和/或转换得到的文字

文件进行分类；

根据发布时间戳将新增数据信息中所包含的文字文件和/或转换得到的文字文件保存至预设数据链表中与发布源信息对应的子链表。

16.根据权利要求 11 所述的计算机设备，其中，所述对新增数据信息中的文件是否为文字文件进行判断，包括：

根据各文件的格式信息对各文件是否为文字文件进行判断。

17.根据权利要求 13 所述的计算机设备，其中，所述通过预设信息识别模型将非文字文件转换为文字文件，还包括：

若所述非文字文件不是视频文件、音频文件及图片中的任意一种，生成报警提示信息以提示用户无法对所述文件进行处理。

18.一种计算机可读存储介质，其中，所述计算机可读存储介质存储有计算机程序，所述计算机程序当被处理器执行时使所述处理器执行以下操作：

获取待监控网页的网址信息，根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息；

对新增数据信息中的文件是否为文字文件进行判断；

若新增数据信息中的文件为非文字文件，通过预设信息识别模型将非文字文件转换为文字文件；

将新增数据信息中的文字文件和/或转换得到的文字文件保存至预设数据链表中。

19. 根据权利要求 18 所述的计算机可读存储介质，其中，所述根据待监控网页的网址信息对待监控网页中所发布的数据信息进行实时监控以获取新增数据信息，包括：

若监控到待监控网页中发布数据信息，根据待监控网页的网址信息及所述数据信息的发布人生成发布源信息；

根据所述数据信息的发布时间生成发布时间戳；

获取所述发布数据信息中的所有文件及发布源信息、发布时间戳以得到新增数据信息。

20. 根据权利要求 18 所述的计算机可读存储介质，其中，所述通过预设信息识别模型将非文字文件转换为文字文件，包括：

获取所述非文字文件的格式信息并判断该文件是否为音频文件，若该文件

为音频文件则通过信息识别模型中的音频识别模型对该文件进行识别以得到相应的文字文件；

获取所述非文字文件的格式信息并判断该文件是否为图片，若该文件为图片则通过信息识别模型中图片识别模型的对该文件进行识别以得到相应的文字文件；

获取所述非文字文件的格式信息并判断该文件是否为视频文件，若该文件为视频文件则通过信息识别模型中的音频识别模型及图片识别模型对该文件进行识别以得到相应的文字文件。

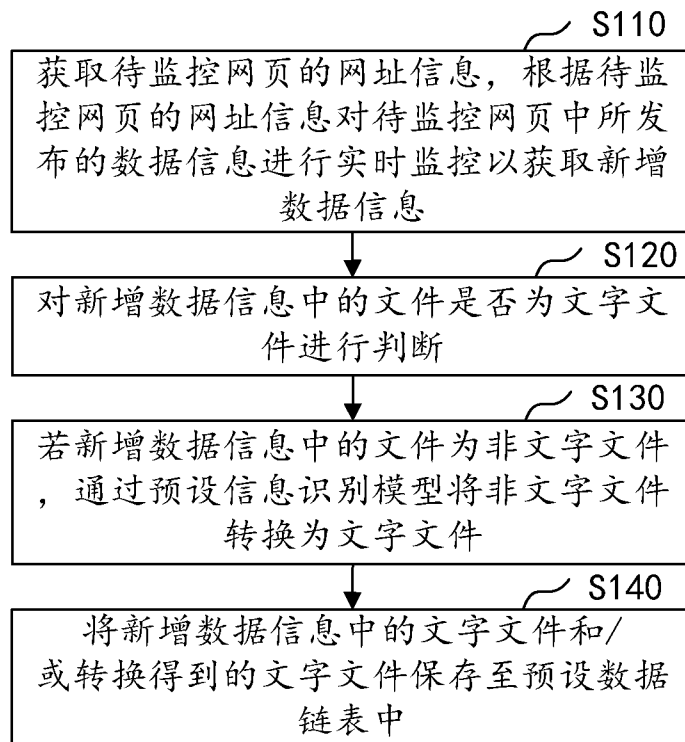


图 1

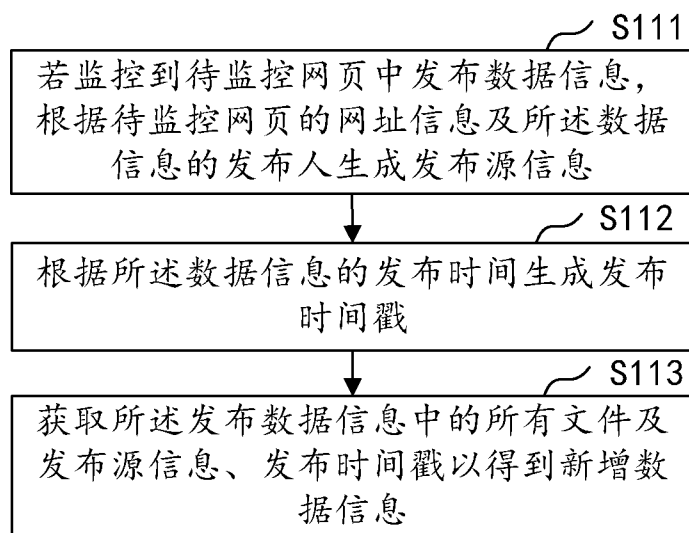


图 2

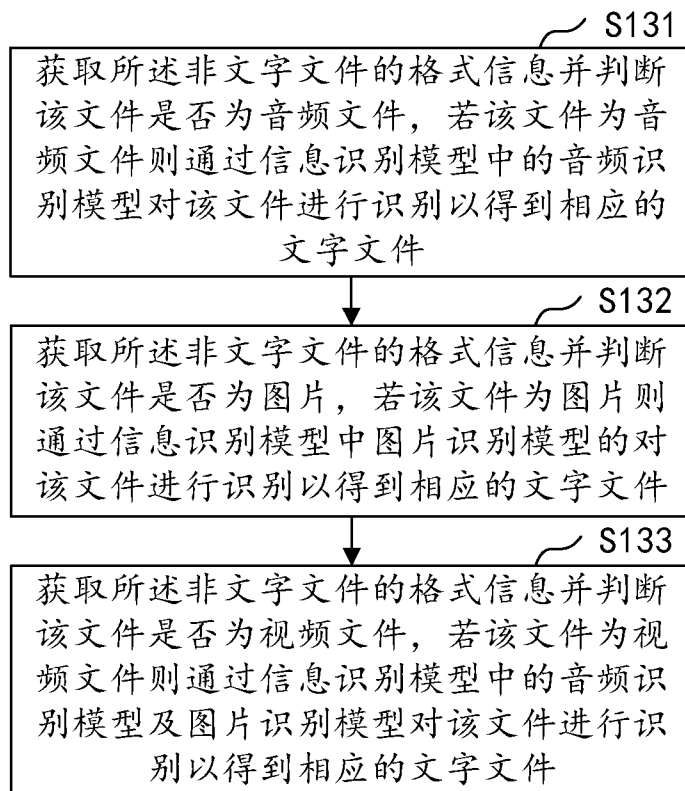


图 3

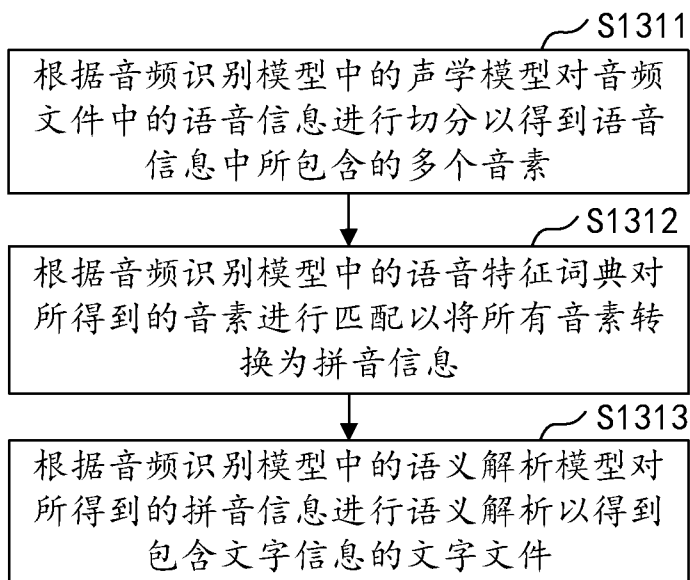


图 4

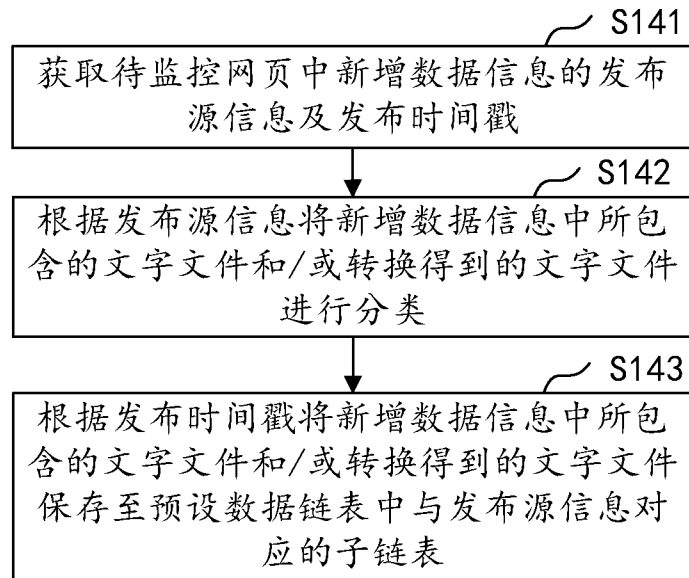


图 5

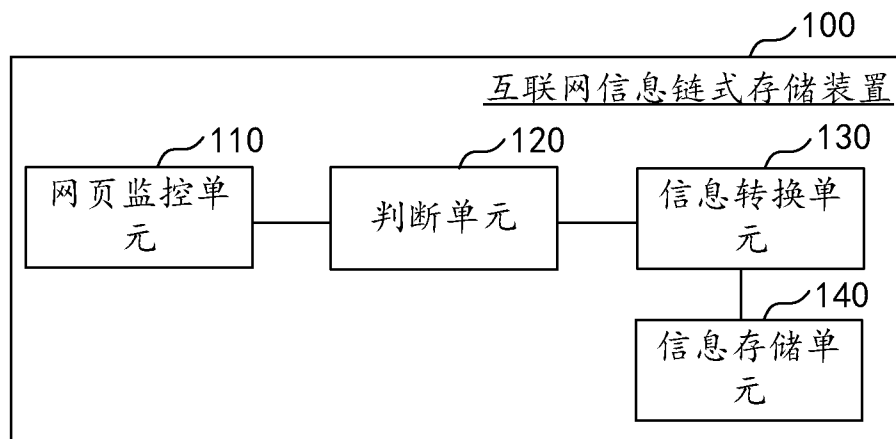


图 6

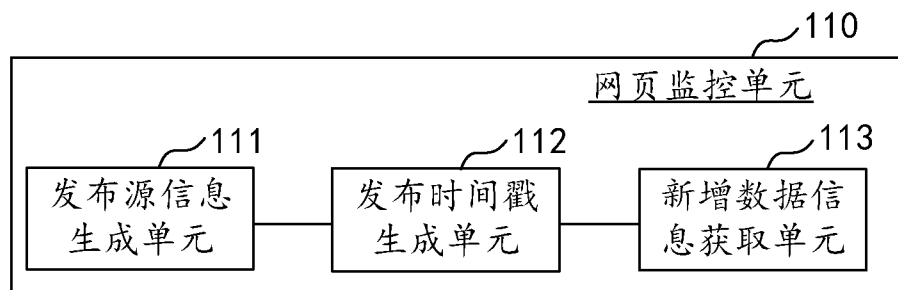


图 7

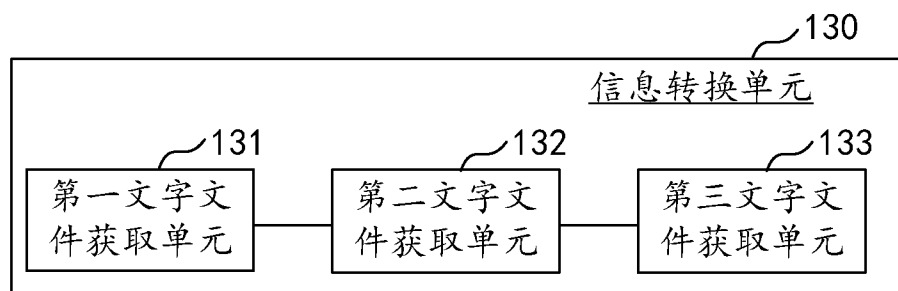


图 8

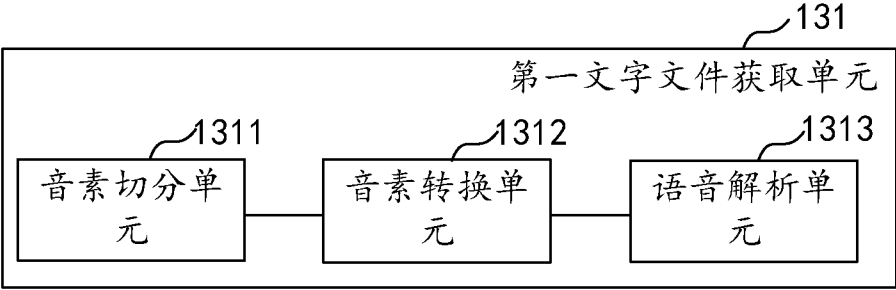


图 9

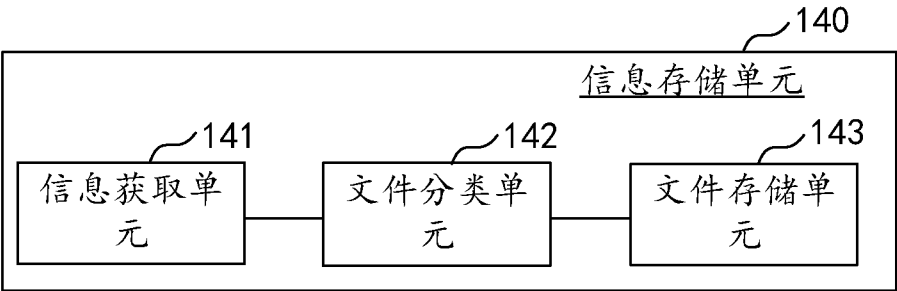


图 10

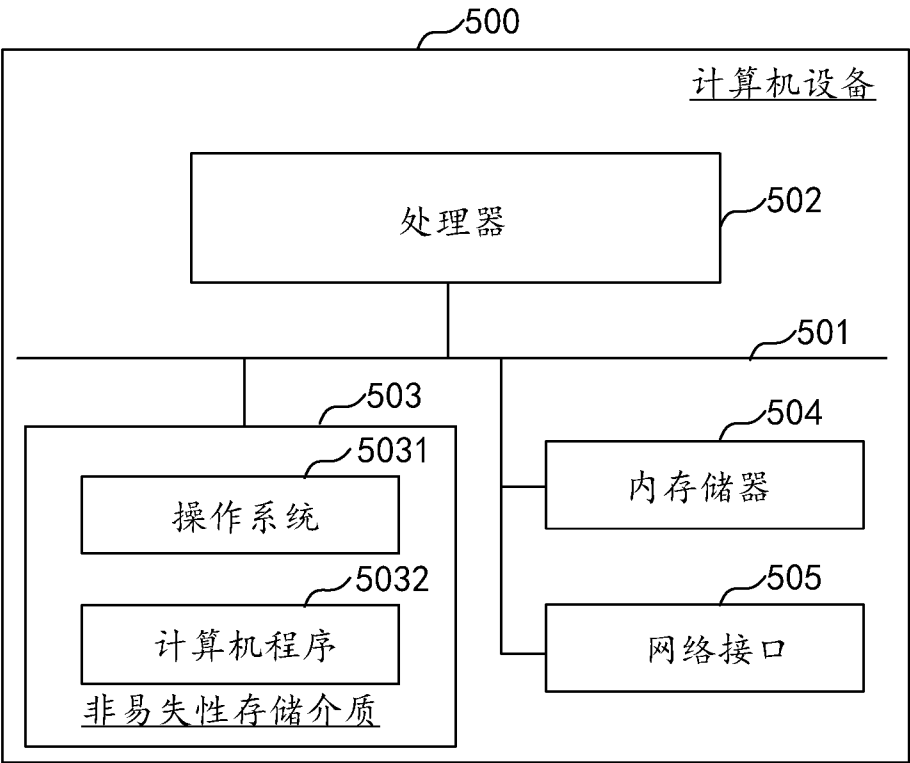


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/092551

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/958(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC, GOOGLE: 网页, 微博, 新增, 更新, 监控, 检测, 监测, 监视, 检查, 实时, 随时, 文字, 文本, 非文字, 非文本, 图片, 音频, 视频, 链表, 保存, 存储, 储存, webpage, text, non-text, link, table, real-time, audio, video, sav+, picture, monitor+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109657181 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 19 April 2019 (2019-04-19) claims 1-10, and description, paragraphs 0093 and 0110	1-20
Y	CN 101882162 A (BEIJING SOGOU TECHNOLOGY DEVELOPMENT CO., LTD.) 10 November 2010 (2010-11-10) claims 1-17 and description, paragraphs 0096-0104	1-20
Y	CN 101364955 A (HANGZHOU DIANZI UNIVERSITY) 11 February 2009 (2009-02-11) claim 1 and description, pp. 4-8	1-20
A	WO 2017206739 A1 (GUANGZHOU DONG JING COMPUTER SCIENCE & TECHNOLOGY CO., LTD.) 07 December 2017 (2017-12-07) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

25 July 2019

Date of mailing of the international search report

27 August 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/092551

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109657181	A	19 April 2019	None			
CN	101882162	A	10 November 2010	WO	2012000423	A1	05 January 2012
CN	101364955	A	11 February 2009	None			
WO	2017206739	A1	07 December 2017	CN	107450900	A	08 December 2017

国际检索报告

国际申请号

PCT/CN2019/092551

A. 主题的分类 G06F 16/958(2019.01) i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																	
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G06F; H04L 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) CNPAT、CNKI、WPI、EP0DOC、GOOGLE: 网页、微博、新增、更新、监控、检测、监测、监视、检查、实时、随时、文字、文本、非文字、非文本、图片、音频、视频、链表、保存、存储、储存、webpage、text、non-text、link、table、real-time、audio、video、sav+、picture、monitor+																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 109657181 A (平安科技深圳有限公司) 2019年 4月 19日 (2019 - 04 - 19) 权利要求1-10, 说明书第0093、0110段</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 101882162 A (北京搜狗科技发展有限公司) 2010年 11月 10日 (2010 - 11 - 10) 权利要求1-17, 说明书第0096-0104段</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 101364955 A (杭州电子科技大学) 2009年 2月 11日 (2009 - 02 - 11) 权利要求1, 说明书第4-8页</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>WO 2017206739 A1 (广州市动景计算机科技有限公司) 2017年 12月 7日 (2017 - 12 - 07) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 109657181 A (平安科技深圳有限公司) 2019年 4月 19日 (2019 - 04 - 19) 权利要求1-10, 说明书第0093、0110段	1-20	Y	CN 101882162 A (北京搜狗科技发展有限公司) 2010年 11月 10日 (2010 - 11 - 10) 权利要求1-17, 说明书第0096-0104段	1-20	Y	CN 101364955 A (杭州电子科技大学) 2009年 2月 11日 (2009 - 02 - 11) 权利要求1, 说明书第4-8页	1-20	A	WO 2017206739 A1 (广州市动景计算机科技有限公司) 2017年 12月 7日 (2017 - 12 - 07) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 109657181 A (平安科技深圳有限公司) 2019年 4月 19日 (2019 - 04 - 19) 权利要求1-10, 说明书第0093、0110段	1-20															
Y	CN 101882162 A (北京搜狗科技发展有限公司) 2010年 11月 10日 (2010 - 11 - 10) 权利要求1-17, 说明书第0096-0104段	1-20															
Y	CN 101364955 A (杭州电子科技大学) 2009年 2月 11日 (2009 - 02 - 11) 权利要求1, 说明书第4-8页	1-20															
A	WO 2017206739 A1 (广州市动景计算机科技有限公司) 2017年 12月 7日 (2017 - 12 - 07) 全文	1-20															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																	
国际检索实际完成的日期 2019年 7月 25日		国际检索报告邮寄日期 2019年 8月 27日															
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 王怡轩 电话号码 86-10-53961621															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/092551

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	109657181	A	2019年 4月 19日	无			
CN	101882162	A	2010年 11月 10日	WO	2012000423	A1	2012年 1月 5日
CN	101364955	A	2009年 2月 11日	无			
WO	2017206739	A1	2017年 12月 7日	CN	107450900	A	2017年 12月 8日