(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110309355 B
(45)授权公告日 2023.05.16

(21)申请号 201810622125.6

(22)申请日 2018.06.15

(65)同一申请的已公布的文献号
申请公布号 CN 110309355 A

(43)申请公布日 2019.10.08

(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 孙子荀

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
专利代理师 祝亚男

(51)Int.Cl.
G06F 16/73(2019.01)

(56)对比文件
CN 103365904 A,2013.10.23
US 2016371283 A1,2016.12.22
杨肖.基于主题的互联网信息抓取研究.中国博士学位论文全文数据库信息科技辑.2016,第4章.
Michal Barla等.On Deriving Tagsonomies: Keyword Relations Coming from Crowd.International Conference on Computational Collective Intelligence.2009,第309-320页.

审查员 张甜

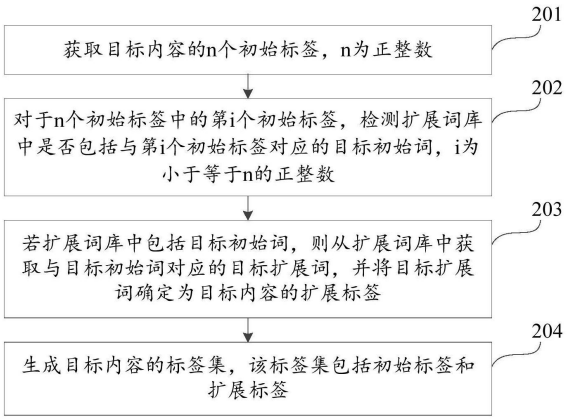
权利要求书2页 说明书9页 附图4页

(54)发明名称

内容标签的生成方法、装置、设备及存储介质

(57)摘要

本申请实施例公开了一种内容标签的生成方法、装置、设备及存储介质。所述方法包括:获取目标内容的n个初始标签;对于n个初始标签中的第i个初始标签,检测扩展词库中是否包括与第i个初始标签对应的目标初始词;其中,扩展词库中包括至少一组初始词和扩展词之间的对应关系;若扩展词库中包括目标初始词,则从扩展词库中获取与目标初始词对应的目标扩展词,并将目标扩展词确定为目标内容的扩展标签;生成目标内容的标签集,该标签集包括初始标签和扩展标签。本申请实施例通过构建扩展词库,在生成内容的初始标签之后,结合初始标签和扩展词库生成内容的扩展标签,从而实现了对标签数量的扩展,使得内容的标签更加丰富。



CN 110309355 B

1.一种内容标签的生成方法,其特征在于,所述方法应用于搜索场景中,所述方法包括:

获取目标内容的n个初始标签,n为正整数;

对于所述n个初始标签中的第i个初始标签,检测扩展词库中是否包括与所述第i个初始标签对应的目标初始词;其中,所述扩展词库中包括至少一组初始词和扩展词之间的对应关系,i为小于等于n的正整数;所述第i个初始标签对应的所述目标初始词为所述第i个初始标签,或者为所述第i个初始标签的同义词;

获取实体词库,所述实体词库中包括至少一个实体词;从所述实体词库中筛选出符合预设条件的实体词作为所述初始词,得到初始词库;分别为所述初始词库中的每个初始词生成对应的扩展词,得到所述扩展词库;符合预设条件的所述实体词是指表征意义显著且不存在一词多义的实体词;所述扩展词库包括上位词库和代表词库;所述上位词库中包括至少一组所述初始词和上位词之间的对应关系,所述上位词是指相较于所述初始词,在概念上外延更广的主题词;所述代表词库中包括至少一组所述初始词和代表词之间的对应关系,所述代表词是指代表所述初始词的词语;

若所述上位词库中包括所述目标初始词,则从所述上位词库中获取与所述目标初始词对应的目标上位词;若所述代表词库中包括所述目标初始词,则从所述代表词库中获取与所述目标初始词对应的目标代表词;将所述目标上位词和所述目标代表词确定为所述目标内容的扩展标签;

生成所述目标内容的标签集,所述标签集包括所述初始标签和所述扩展标签;所述标签集中的标签用于在所述搜索场景中与搜索关键词进行匹配,将与所述搜索关键词相符的所述标签对应的目标内容确定为搜索结果。

2.根据权利要求1所述的方法,其特征在于,所述获取目标内容的n个初始标签,包括:

获取所述目标内容的描述信息,所述描述信息包括用于对所述目标内容进行介绍说明的信息;

对所述描述信息执行分词处理,生成候选词;

对所述候选词执行聚类处理,得到至少一个类,每个类中包含至少一个候选词;

获取每个类的主题词,作为所述目标内容的初始标签。

3.根据权利要求2所述的方法,其特征在于,所述对所述候选词执行聚类处理,得到至少一个类,包括:

提取每个候选词的词向量;

根据每两个候选词的词向量,计算每两个候选词之间的相似度;

根据每两个候选词之间的相似度,对所述候选词执行聚类处理,得到所述至少一个类。

4.一种内容标签的生成装置,其特征在于,所述装置包括:

标签获取模块,用于获取目标内容的n个初始标签,n为正整数;

检测模块,用于对于所述n个初始标签中的第i个初始标签,检测扩展词库中是否包括与所述第i个初始标签对应的目标初始词;其中,所述扩展词库中包括至少一组初始词和扩展词之间的对应关系,i为小于等于n的正整数;所述第i个初始标签对应的所述目标初始词为所述第i个初始标签,或者为所述第i个初始标签的同义词;

词库获取模块,用于获取实体词库,所述实体词库中包括至少一个实体词;筛选模块,

用于从所述实体词库中筛选出符合预设条件的实体词作为所述初始词，得到初始词库；词库创建模块，用于分别为所述初始词库中的每个初始词生成对应的扩展词，得到所述扩展词库；符合预设条件的实体词是指表征意义显著且不存在一词多义的实体词；所述扩展词库包括上位词库和代表词库；所述上位词库中包括至少一组所述初始词和上位词之间的对应关系，所述上位词是指相较于所述初始词，在概念上外延更广的主题词；所述代表词库中包括至少一组所述初始词和代表词之间的对应关系，所述代表词是指代表所述初始词的词语；

标签扩展模块，包括：上位词扩展单元和代表词扩展单元；所述上位词扩展单元，用于若所述上位词库中包括所述目标初始词，则从所述上位词库中获取与所述目标初始词对应的目标上位词；所述代表词扩展单元，用于若所述代表词库中包括所述目标初始词时，则从所述代表词库中获取与所述目标初始词对应的目标代表词；将所述目标上位词和所述目标代表词确定为所述目标内容的扩展标签；

标签生成模块，用于生成所述目标内容的标签集，所述标签集包括所述初始标签和所述扩展标签；所述标签集中的标签用于在搜索场景中与搜索关键词进行匹配，将与所述搜索关键词相符的所述标签对应的目标内容确定为搜索结果。

5.根据权利要求4所述的装置，其特征在于，所述标签获取模块，包括：

信息获取单元，用于获取所述目标内容的描述信息，所述描述信息包括用于对所述目标内容进行介绍说明的信息；

分词单元，用于对所述描述信息执行分词处理，生成候选词；

聚类单元，用于对所述候选词执行聚类处理，得到至少一个类，每个类中包含至少一个候选词；

标签获取单元，用于获取每个类的主题词，作为所述目标内容的初始标签。

6.根据权利要求5所述的装置，其特征在于，所述聚类单元，用于：

提取每个候选词的词向量；

根据每两个候选词的词向量，计算每两个候选词之间的相似度；

根据每两个候选词之间的相似度，对所述候选词执行聚类处理，得到所述至少一个类。

7.一种计算机设备，其特征在于，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一段程序，所述至少一段程序由所述处理器加载并执行以实现如权利要求1至3任一所述的内容标签的生成方法。

8.一种计算机可读存储介质，其特征在于，所述存储介质中存储有至少一段程序，所述至少一段程序由处理器加载并执行以实现如权利要求1至3任一所述的内容标签的生成方法。

内容标签的生成方法、装置、设备及存储介质

技术领域

[0001] 本申请实施例涉及互联网技术领域，特别涉及一种内容标签的生成方法、装置、设备及存储介质。

背景技术

[0002] 目前，互联网能够提供纷繁多样的内容资源以供用户观看阅览，如电影、电视剧、综艺、动漫、音乐、小说等。提供上述内容资源的网站或者应用程序，通常会为内容设置标签，以便用户基于标签来了解内容的核心点。例如，一部电影的标签可以包括：动作、2018、超级英雄等。

[0003] 在相关技术中，主要从内容的标题或者内容本身中提取内容的标签。这种方法生成的标签数量较少，存在一定局限性。

发明内容

[0004] 本申请实施例提供了一种内容标签的生成方法、装置、设备及存储介质，可以解决相关技术生成的标签数量较少的问题，减小其局限性。技术方案如下：

[0005] 一方面，本申请实施例提供一种内容标签的生成方法，所述方法包括：

[0006] 获取目标内容的n个初始标签，所述n为正整数；

[0007] 对于所述n个初始标签中的第i个初始标签，检测扩展词库中是否包括与所述第i个初始标签对应的目标初始词；其中，所述扩展词库中包括至少一组初始词和扩展词之间的对应关系，所述i为小于等于n的正整数；

[0008] 若所述扩展词库中包括所述目标初始词，则从所述扩展词库中获取与所述目标初始词对应的目标扩展词，并将所述目标扩展词确定为所述目标内容的扩展标签；

[0009] 生成所述目标内容的标签集，所述标签集包括所述初始标签和所述扩展标签。

[0010] 另一方面，本申请实施例提供一种内容标签的生成装置，所述装置包括：

[0011] 标签获取模块，用于获取目标内容的n个初始标签，所述n为正整数；

[0012] 检测模块，用于对于所述n个初始标签中的第i个初始标签，检测扩展词库中是否包括与所述第i个初始标签对应的目标初始词；其中，所述扩展词库中包括至少一组初始词和扩展词之间的对应关系，所述i为小于等于n的正整数；

[0013] 标签扩展模块，用于当所述扩展词库中包括所述目标初始词时，从所述扩展词库中获取与所述目标初始词对应的目标扩展词，并将所述目标扩展词确定为所述目标内容的扩展标签；

[0014] 标签生成模块，用于生成所述目标内容的标签集，所述标签集包括所述初始标签和所述扩展标签。

[0015] 再一方面，本申请实施例提供一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上述方

面所述的内容标签的生成方法。

[0016] 又一方面，本申请实施例提供一种计算机可读存储介质，所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如上述方面所述的内容标签的生成方法。

[0017] 又一方面，本申请实施例提供一种计算机程序产品，当该计算机程序产品被执行时，其用于执行上述方面所述的内容标签的生成方法。

[0018] 本申请实施例提供的技术方案中，通过构建扩展词库，在生成内容的初始标签之后，结合初始标签和扩展词库生成内容的扩展标签，从而实现了对标签数量的扩展，使得内容的标签更加丰富。

附图说明

[0019] 图1是本申请一个实施例提供的实施环境的示意图；

[0020] 图2是本申请一个实施例提供的内容标签的生成方法的流程图；

[0021] 图3示例性示出了一种初始词和上位词之间关系的示意图；

[0022] 图4示例性示出了一种扩展词库的构建过程的流程图；

[0023] 图5是本申请一个实施例提供的内容标签的生成装置的框图；

[0024] 图6是本申请另一个实施例提供的内容标签的生成装置的框图；

[0025] 图7是本申请一个实施例提供的计算机设备的结构框图。

具体实施方式

[0026] 下面将结合附图对本申请实施方式作进一步地详细描述。

[0027] 本申请实施例提供的技术方案中，通过构建扩展词库，在生成内容的初始标签之后，结合初始标签和扩展词库生成内容的扩展标签，从而实现了对标签数量的扩展，使得内容的标签更加丰富。

[0028] 在本申请实施例中，“标签”是指能够体现内容的特征的词语。另外，本申请实施例提及的“内容”，其可以是视频、音乐、小说等媒体资源。以视频为例，可以包括电影、电视剧、综艺、体育节目、动漫等。在本申请实施例中，主要以内容是视频为例，对本申请技术方案进行介绍和说明。对于其它类型的内容，本申请技术方案对于解决上述问题以实现对标签数量的扩展同样适用。

[0029] 请参考图1，其示出了本申请一个实施例提供的实施环境的示意图。该实施环境可以包括：终端10和服务器20。

[0030] 终端10可以是诸如手机、平板电脑、电子书阅读器、多媒体播放设备、可穿戴设备、PC(Personal Computer，个人计算机)等电子设备。终端10中可以安装浏览器或者应用程序客户端，通过浏览器或者应用程序客户端从服务器20中获取内容并进行展示。

[0031] 服务器20用于为终端10提供内容。例如，服务器20可以是用于提供内容的网站或者应用程序的后台服务器。服务器20可以是一台服务器，也可以是由多台服务器组成的服务器集群，或者是一个云计算服务中心。

[0032] 终端10和服务器20之间可通过网络30进行互相通信。该网络30可以是有线网络，也可以是无线网络。

[0033] 在一种可能的应用场景中，终端10在展示内容时，会将内容的标签同步展示，以便用户基于标签来了解内容的核心点。在另一种可能的应用场景中，终端10支持基于标签的内容搜索功能。用户在终端10中输入搜索关键词之后，终端10将具有与该搜索关键词相符的标签的内容作为搜索结果提供给用户。当然，上述仅是列举了两种关于内容标签的典型应用场景，对于其它可能的应用场景，本申请实施例不作赘述。

[0034] 请参考图2，其示出了本申请一个实施例提供的内容标签的生成方法的流程图。该方法可应用于图1所示的实施环境的服务器20中。该方法可以包括如下几个步骤：

[0035] 步骤201，获取目标内容的n个初始标签，n为正整数。

[0036] 初始标签是指根据目标内容的相关信息提取到的标签，上述相关信息可以是任何与目标内容相关的信息，如标题、描述信息、目标内容本身、评论信息等。

[0037] 在本申请实施例中，对获取目标内容的初始标签的方式不作限定，下文实施例中会对一种可能的实现方式进行介绍说明。

[0038] 步骤202，对于n个初始标签中的第i个初始标签，检测扩展词库中是否包括与第i个初始标签对应的目标初始词，i为小于等于n的正整数。

[0039] 第i个初始标签可以是上述n个初始标签中的任意一个标签。另外，第i个初始标签对应的目标初始词可以是该第i个初始标签本身，也可以是该第i个初始标签的同义词。

[0040] 在本申请实施例中，预先构建扩展词库。扩展词库中包括至少一组初始词和扩展词之间的对应关系。同一个初始词可以对应一个或者多个扩展词，同一个扩展词也可以对应一个或者多个初始词。对于任意一组对应的初始词和扩展词，该扩展词是指与该初始词具有强相关性的词语。

[0041] 可选地，扩展词库包括上位词库和/或代表词库。

[0042] 上位词库中包括至少一组初始词和上位词之间的对应关系。同一个初始词可以对应一个或者多个上位词，同一个上位词也可以对应一个或者多个初始词。对于任意一组对应的初始词和上位词，该上位词是指相较于该初始词，在概念上外延更广的主题词。例如，“花”是“鲜花”的上位词，“植物”是“花”的上位词，“音乐”是“mp3”的上位词。一个初始词所表达概念的任何一种属性、任何一种归类方式，都可以是该初始词的上位词。例如，“鲜花快递”的上位词可以是“鲜花”、“快递”、“网上购物”、“鲜花礼仪”、“鲜花店”、“礼品公司”等。

[0043] 结合参考图3，其示例性示出了一种初始词和上位词之间关系的示意图。“菊花”和“牡丹花”的上位词是“花”，“苹果树”和“桃树”的上位词是“树”，“花”和“树”的上位词是“植物”。

[0044] 在上位词库中，每一组对应的初始词和上位词可以采用如下格式存储：{key：“初始词”；relation：“hypernym”；value：“上位词”}。

[0045] 代表词库中包括至少一组初始词和代表词之间的对应关系。同一个初始词可以对应一个或者多个代表词，同一个代表词也可以对应一个或者多个初始词。对于任意一组对应的初始词和代表词，该代表词是指能够代表该初始词的词语。

[0046] 在代表词库中，每一组对应的初始词和代表词可以采用如下格式存储：{key：“初始词”；relation：“expand”；value：“代表词”}。

[0047] 步骤203，若扩展词库中包括目标初始词，则从扩展词库中获取与目标初始词对应的目标扩展词，并将目标扩展词确定为目标内容的扩展标签。

[0048] 可选地，若上位词库中包括目标初始词，则从上位词库中获取与目标初始词对应的目标上位词，并将目标上位词确定为目标内容的扩展标签；若代表词库中包括目标初始词，则从代表词库中获取与目标初始词对应的目标代表词，并将目标代表词确定为目标内容的扩展标签。

[0049] 可选地，对于上述n个初始标签中的每一个初始标签，服务器均执行上述步骤202和203，以分别获得与每一个初始标签对应的扩展标签。

[0050] 另外，如果扩展词库中不包括与第i个初始标签对应的目标初始词，则服务器无法获得与该第i个初始标签对应的扩展标签。

[0051] 步骤204，生成目标内容的标签集，该标签集包括初始标签和扩展标签。

[0052] 服务器获得目标内容的扩展标签之后，将目标内容的初始标签和扩展标签整合，得到目标内容的标签集。

[0053] 可选地，服务器将目标内容的标签集提供给审核人员进行审核，由审核人员从中筛选出合适的标签，最终作为目标内容的标签。

[0054] 综上所述，本申请实施例提供的技术方案中，通过构建扩展词库，在生成内容的初始标签之后，结合初始标签和扩展词库生成内容的扩展标签，从而实现了对标签数量的扩展，使得内容的标签更加丰富。

[0055] 在基于图2实施例提供的一个可选实施例中，通过下述方法生成扩展词库：

[0056] 1、获取实体词库；

[0057] 实体词库中包括至少一个实体词。实体词是指用于表征人或事物的词语，实体词通常为名词。可选地，通过网络爬虫技术从百科网站中爬取实体词，构建实体词库。百科网站是指提供各种不同领域的知识介绍的网站，如艺术、科学、自然、文化、地理、生活、社会、人物、经济、体育、历史等领域。百科网站中对各个不同领域的人或事物有较为权威的分类和定义，因此从百科网站中爬取实体词具有可行性，且较为准确可靠。

[0058] 2、从实体词库中筛选出符合预设条件的实体词作为初始词，得到初始词库；

[0059] 可选地，符合预设条件的实体词是指表征意义显著且不存在一词多义的实体词。从实体词库中筛选出初始词库的过程，可以由人工筛选实现。

[0060] 3、分别为初始词库中的每个初始词生成对应的扩展词，得到扩展词库。

[0061] 在上文已经介绍，扩展词包括上位词和/或代表词。当扩展词包括上位词和代表词时，可以通过一个词库(如称为扩展词库)记录初始词和扩展词之间的对应关系，也可以通过两个词库(如包括上位词库和代表词库)分别记录初始词和上位词之间的对应关系，以及初始词和代表词之间的对应关系。

[0062] 可选地，对于生成上位词的方法，包括但不限于以下几种：

[0063] 1、词前后缀法

[0064] 通过获取初始词的前缀或者后缀，作为该初始词的上位词。例如，“牡丹花”、“菊花”的后缀是“花”，可以将“花”作为“牡丹花”、“菊花”的上位词。

[0065] 2、共现词法

[0066] 通过获取初始词的共现词，作为该初始词的上位词。所谓初始词的共现词，是指与该初始词共同出现的频率高于预设门限值的词语。可选地，通过获取包含有初始词的相关语料，对该相关语料进行分析，从中提取初始词的共现词。

[0067] 3、规则模板法

[0068] 通过规则模板从包含有初始词、且符合特定句式的句子中抽取该初始词的上位词。

[0069] 上文介绍的几种生成上位词的方法仅是示例性和解释性的，本申请实施例并不限定还可采用其它方法来生成上位词。另外，在生成上位词时，可以采用其中一种方法，也可以采用多种方法的组合，例如对于某一初始词，分别采用多种不同的方法来生成该初始词的上位词，然后将生成的上位词整合，最终将出现次数大于阈值的上位词确定为该初始词的上位词。

[0070] 可选地，对于生成代表词的方法，可以采用基于规则的方法。例如，基于球队和队长之间的代表关系、基于电影和主演之间的代表关系、基于综艺和主持人之间的代表关系，构建代表词的生成规则，然后基于上述规则，通过网络爬虫技术从相关网站爬取球队的队长、电影的主演信息、综艺的主持人等信息，得到初始词的代表词。

[0071] 结合参考图4，其示例性示出了扩展词库的构建过程的流程图。首先，通过网络爬虫技术从百科网站中爬取实体词，构建实体词库51；然后从实体词库中筛选出符合预设条件的实体词作为初始词，得到初始词库52；之后，分别通过上位词生成和代表词生成，得到上位词库53和代表词库54。

[0072] 综上所述，本申请实施例提供的技术方案中，通过获取实体词库，从实体词库中筛选出符合预设条件的实体词构建初始词库，然后基于上位词生成规则和/或代表词生成规则，生成扩展词库，从而构建出一个用于标签扩展的知识图谱，为标签扩展提供数据支持。

[0073] 在基于图2实施例或者上述可选实施例提供的另一个可选实施例中，通过下述方法获取目标内容的n个初始标签：

[0074] 1、获取目标内容的描述信息；

[0075] 描述信息包括用于对目标内容进行介绍说明的信息。可选地，通过网络爬虫技术从相关网站爬取目标内容的描述信息。以电影为例，可以通过网络爬虫技术从百科类网站或影视类网站中爬取电影的描述信息，如电影的剧情简介。

[0076] 2、对描述信息执行分词处理，生成候选词；

[0077] 在本申请实施例中，对分词处理所采用的算法不作限定。例如，对于中文来说，可以采用开源的jieba分词软件进行分词处理。

[0078] 可选地，本步骤包括如下几个子步骤：

[0079] (1)、对描述信息执行分词处理，得到至少两个词语；

[0080] (2)、从至少两个词语中选取目标词性的词语，作为候选词。

[0081] 由于需要提取的候选词是能够体现内容特征的描述性词语，因此在分词处理之后，可以根据词语的词性筛选出一些词语作为候选词。例如，上述目标词性包括以下至少一项：名词、形容词、动词。而将非目标词性的词语筛除，不作为候选词。

[0082] (3)、对候选词执行聚类处理，得到至少一个类，每个类中包含至少一个候选词；

[0083] 分词处理后得到的各个候选词之间是没有关联的，在本申请实施例中，根据各个候选词之间的语义相似度，对候选词执行聚类处理，得到至少一个类。属于同一类的候选词具有相同或者相似的语义。

[0084] 可选地，本步骤包括如下几个子步骤：

[0085] (1)、提取每个候选词的词向量；

[0086] (2)、根据每两个候选词的词向量，计算每两个候选词之间的相似度；

[0087] (3)、根据每两个候选词之间的相似度，对候选词执行聚类处理，得到至少一个类。

[0088] 在本申请实施例中，可以通过计算两个候选词的词向量之间的相似度，来得到两个候选词之间的相似度。也就是说，将判断两个候选词的语义是否相似的问题，转换成计算词向量的相似度。可选地，采用开源的word2vec工具对候选词进行词向量训练，训练结果是每个候选词都表示成一个k维的向量，k为正整数。在提取各个候选词的词向量之后，需要通过聚类的方法，将词向量相似的候选词聚集成一个类，这样做的原因是，不同词汇所表达的意思可能是相同或者相近的，因此需要将语义相同或者相近的不同候选词进行聚类。在本申请实施例中，对聚类所采用的算法不作限定，如K-Means算法。

[0089] 4、获取每个类的主题词，作为目标内容的初始标签。

[0090] 在对候选词进行聚类之后，获取每个类的主题词，主题词用于代表该类中包含的候选词。在一个示例中，采用人工标注的方式为每个类标注主题词。在另一个示例中，从每个类所包含的候选词中选择一个候选词作为该类的主题词，例如可以选择类中的第一个候选词或者随机选择一个候选词作为主题词。

[0091] 示例性地，某个类中包含如下候选词：救回、送回、救出、逃走、解救，可以将“救亡”作为该类的主题词。

[0092] 服务器获取到每个类的主题词之后，将获取到的各个主题词确定为目标内容的初始标签。

[0093] 综上所述，本申请实施例提供的技术方案中，提供了一种从内容的描述信息中提取内容的初始标签的方式。当然，还可采用其它方式提取内容的初始标签，以视频为例，可以从视频的标题中提取初始标签，还可以通过语音识别技术识别出视频中的语音信息对应的文本，从上述文本中提取初始标签，也可以基于深度学习技术从视频内容中提取初始标签，等等。在本申请实施例中，对提取内容的初始标签所采用的方式不作具体限定。

[0094] 下述为本申请装置实施例，可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节，请参照本申请方法实施例。

[0095] 请参考图5，其示出了本申请一个实施例提供的内容标签的生成装置的框图。该装置具有实现上述方法示例的功能，所述功能可以由硬件实现，也可以由硬件执行相应的软件实现。该装置600可以包括：标签获取模块610、检测模块620、标签扩展模块630和标签生成模块640。

[0096] 标签获取模块610，用于获取目标内容的n个初始标签，所述n为正整数。

[0097] 检测模块620，用于对于所述n个初始标签中的第i个初始标签，检测扩展词库中是否包括与所述第i个初始标签对应的目标初始词；其中，所述扩展词库中包括至少一组初始词和扩展词之间的对应关系，所述i为小于等于n的正整数。

[0098] 标签扩展模块630，用于当所述扩展词库中包括所述目标初始词时，从所述扩展词库中获取与所述目标初始词对应的目标扩展词，并将所述目标扩展词确定为所述目标内容的扩展标签。

[0099] 标签生成模块640，用于生成所述目标内容的标签集，所述标签集包括所述初始标签和所述扩展标签。

[0100] 综上所述，本申请实施例提供的技术方案中，通过构建扩展词库，在生成内容的初始标签之后，结合初始标签和扩展词库生成内容的扩展标签，从而实现了对标签数量的扩展，使得内容的标签更加丰富。

[0101] 在基于图5实施例提供的一个可选实施例中，所述扩展词库包括上位词库，所述上位词库中包括至少一组初始词和上位词之间的对应关系。

[0102] 相应地，如图6所示，所述标签扩展模块630，包括：上位词扩展单元630a。

[0103] 所述上位词扩展单元630a，用于当所述上位词库中包括所述目标初始词时，从所述上位词库中获取与所述目标初始词对应的目标上位词，并将所述目标上位词确定为所述目标内容的扩展标签。

[0104] 在基于图5实施例或者上述可选实施例提供的另一个可选实施例中，所述扩展词库包括代表词库，所述代表词库中包括至少一组初始词和代表词之间的对应关系。

[0105] 相应地，如图6所示，所述标签扩展模块630，包括：代表词扩展单元630b。

[0106] 所述代表词扩展单元630b，用于当所述代表词库中包括所述目标初始词时，从所述代表词库中获取与所述目标初始词对应的目标代表词，并将所述目标代表词确定为所述目标内容的扩展标签。

[0107] 在基于图5实施例或者上述可选实施例提供的另一个可选实施例中，如图6所示，所述装置600还包括：词库获取模块650、筛选模块660和词库创建模块670。

[0108] 词库获取模块650，用于获取实体词库，所述实体词库中包括至少一个实体词。

[0109] 筛选模块660，用于从所述实体词库中筛选出符合预设条件的实体词作为所述初始词，得到初始词库。

[0110] 词库创建模块670，用于分别为所述初始词库中的每个初始词生成对应的扩展词，得到所述扩展词库。

[0111] 在基于图5实施例或者上述可选实施例提供的另一个可选实施例中，所述标签获取模块610，包括：信息获取单元、分词单元、聚类单元和标签获取单元(图中未示出)。

[0112] 信息获取单元，用于获取所述目标内容的描述信息，所述描述信息包括用于对所述目标内容进行介绍说明的信息。

[0113] 分词单元，用于对所述描述信息执行分词处理，生成候选词。

[0114] 聚类单元，用于对所述候选词执行聚类处理，得到至少一个类，每个类中包含至少一个候选词。

[0115] 标签获取单元，用于获取每个类的主题词，作为所述目标内容的初始标签。

[0116] 可选地，所述聚类单元，用于：提取每个候选词的词向量；根据每两个候选词的词向量，计算每两个候选词之间的相似度；根据每两个候选词之间的相似度，对所述候选词执行聚类处理，得到所述至少一个类。

[0117] 需要说明的是，上述实施例提供的装置，在实现其功能时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的装置与方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

[0118] 请参考图7，其示出了本申请一个实施例提供的计算机设备的结构框图。该计算机

设备可用于实施上述实施例中提供的内容标签的生成方法。该计算机设备可以是PC或者服务器，或者其它具备数据处理和存储能力的设备。具体来讲：

[0119] 所述计算机设备800包括中央处理单元(CPU)801、包括随机存取存储器(RAM)802和只读存储器(ROM)803的系统存储器804，以及连接系统存储器804和中央处理单元801的系统总线805。所述计算机设备800还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统(I/O系统)806，和用于存储操作系统813、应用程序814和其他程序模块815的大容量存储设备807。

[0120] 所述基本输入/输出系统806包括有用于显示信息的显示器808和用于用户输入信息的诸如鼠标、键盘之类的输入设备809。其中所述显示器808和输入设备809都通过连接到系统总线805的输入输出控制器810连接到中央处理单元801。所述基本输入/输出系统806还可以包括输入输出控制器810以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入输出控制器810还提供输出到显示屏、打印机或其他类型的输出设备。

[0121] 所述大容量存储设备807通过连接到系统总线805的大容量存储控制器(未示出)连接到中央处理单元801。所述大容量存储设备807及其相关联的计算机可读介质为计算机设备800提供非易失性存储。也就是说，所述大容量存储设备807可以包括诸如硬盘或者CD-ROM驱动器之类的计算机可读介质(未示出)。

[0122] 不失一般性，所述计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括RAM、ROM、EPROM、EEPROM、闪存或其他固态存储其技术，CD-ROM、DVD或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然，本领域技术人员可知所述计算机存储介质不局限于上述几种。上述的系统存储器804和大容量存储设备807可以统称为存储器。

[0123] 根据本申请的各种实施例，所述计算机设备800还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即计算机设备800可以通过连接在所述系统总线805上的网络接口单元811连接到网络812，或者说，也可以使用网络接口单元811来连接到其他类型的网络或远程计算机系统(未示出)。

[0124] 所述存储器还包括一个或者一个以上的程序，所述一个或者一个以上程序存储于存储器中，且经配置以由一个或者一个以上处理器执行。上述一个或者一个以上程序包含用于执行上述内容标签的生成方法的指令。

[0125] 在示例中实施例中，还提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集。所述至少一条指令、至少一段程序、代码集或指令集经配置以由一个或者一个以上处理器执行，以实现上述内容标签的生成方法。

[0126] 在示例性实施例中，还提供了一种计算机可读存储介质，所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或所述指令集在被计算机设备的处理器执行时实现上述内容标签的生成方法。

[0127] 可选地，上述计算机可读存储介质可以是ROM、RAM、CD-ROM、磁带、软盘和光数据存储设备等。

[0128] 在示例性实施例中，还提供了一种计算机程序产品，当该计算机程序产品被执行时，其用于实现上述内容标签的生成方法。

[0129] 应当理解的是，在本文中提及的“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

[0130] 以上所述仅为本申请的示例性实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

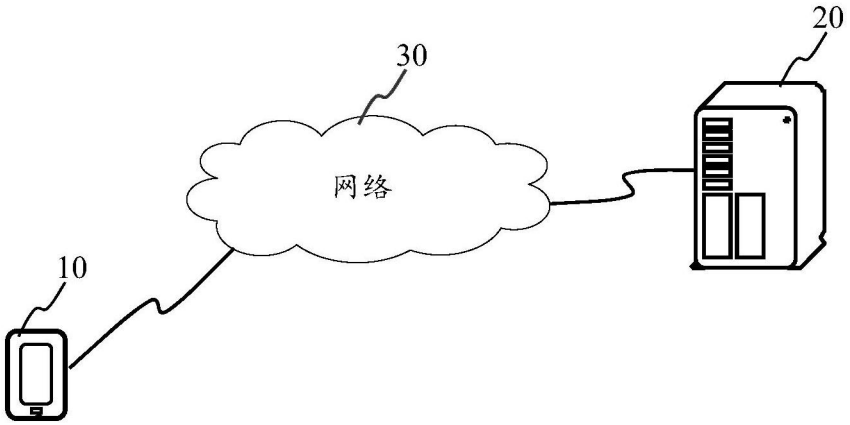

20
30
网络
10

图1

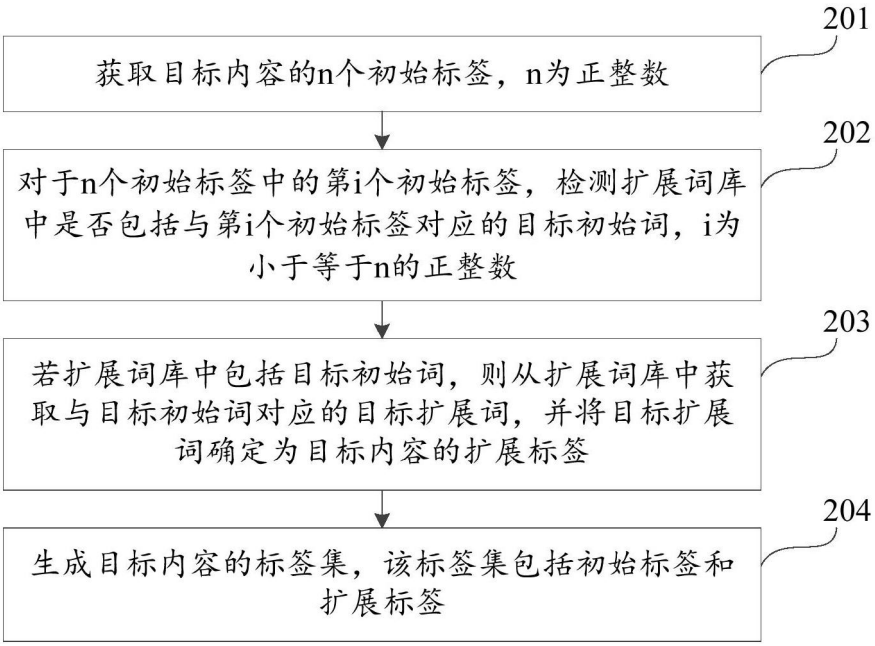

201
获取目标内容的n个初始标签，n为正整数
202
对于n个初始标签中的第i个初始标签，检测扩展词库中是否包括与第i个初始标签对应的目标初始词，i为小于等于n的正整数
203
若扩展词库中包括目标初始词，则从扩展词库中获取与目标初始词对应的目标扩展词，并将目标扩展词确定为目标内容的扩展标签
204
生成目标内容的标签集，该标签集包括初始标签和扩展标签

图2

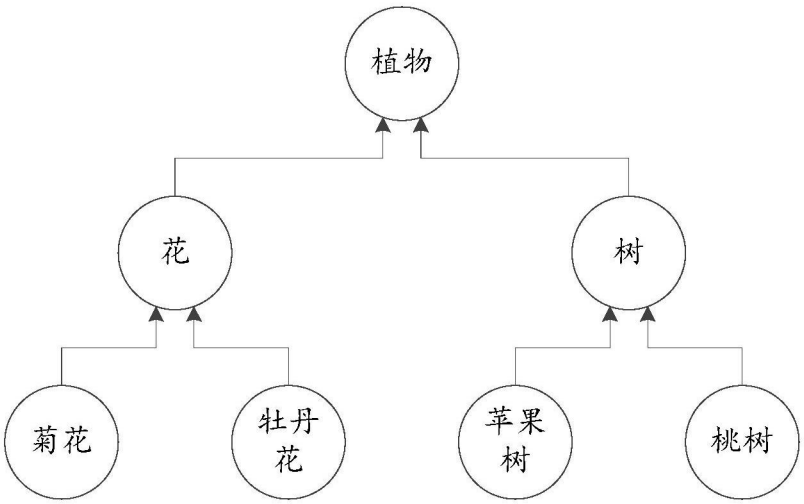


图3

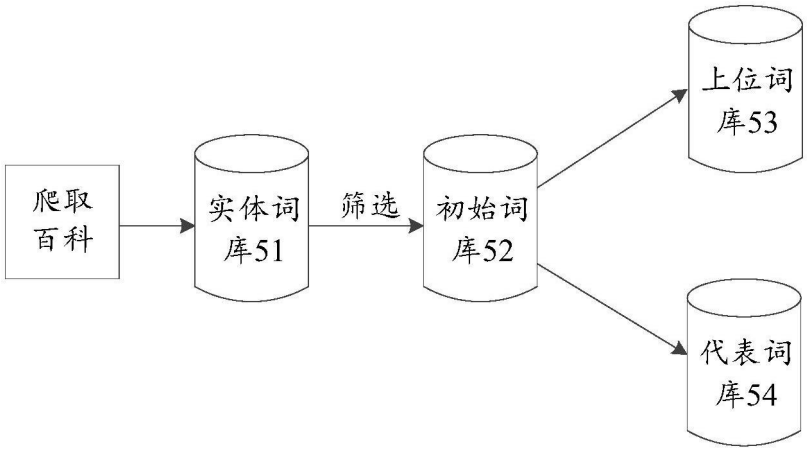


图4

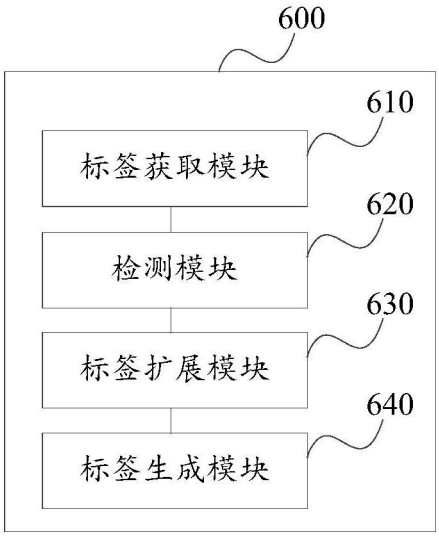

600
610
标签获取模块
620
检测模块
630
标签扩展模块
640
标签生成模块

图5

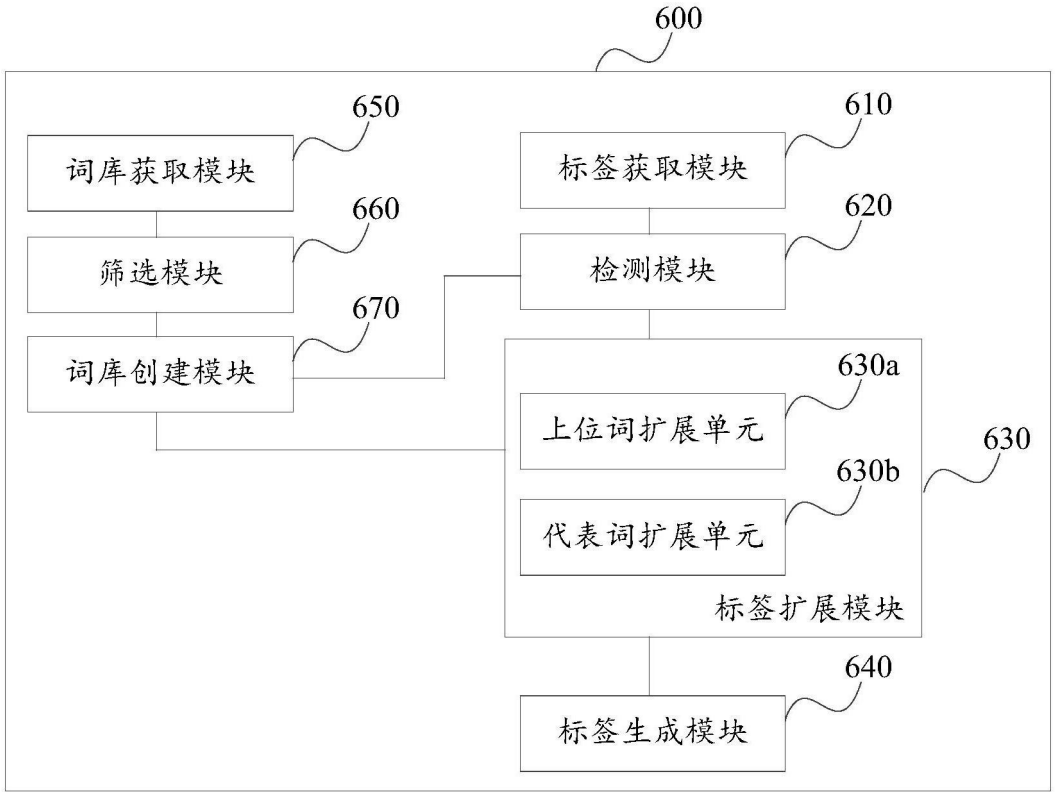

600
650
词库获取模块
660
筛选模块
670
词库创建模块
610
标签获取模块
620
检测模块
630a
上位词扩展单元
630b
代表词扩展单元
630
标签扩展模块
640
标签生成模块

图6

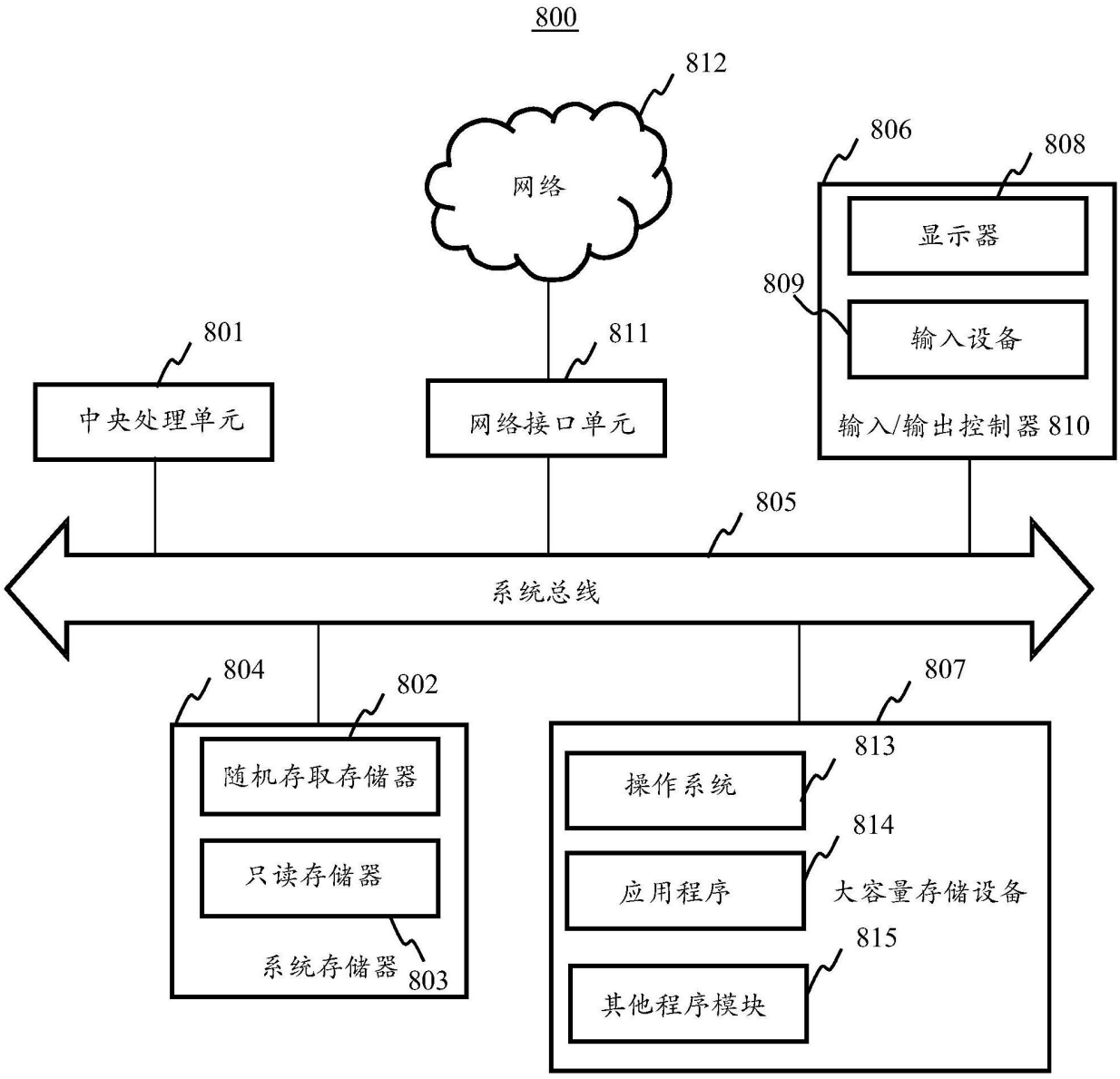

800
812
网络
806
808
显示器
809
输入设备
801
811
中央处理单元
网络接口单元
输入/输出控制器 810
805
系统总线
804
802
807
随机存取存储器
813
操作系统
只读存储器
814
应用程序
大容量存储设备
系统存储器
803
815
其他程序模块

图7