(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110222189 A
(43)申请公布日 2019.09.10

(21)申请号 201910533276.9

(22)申请日 2019.06.19

(71)申请人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 蒋帅 陈思姣 梁海金 罗雨 卞东海

(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 王达佐 马晓亚

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/55(2019.01)
G06F 17/27(2006.01)
G06K 9/62(2006.01)

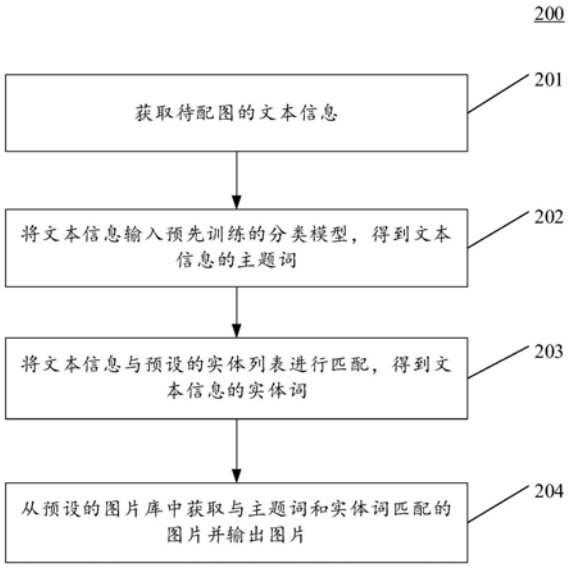
权利要求书2页 说明书9页 附图5页

(54)发明名称

用于输出信息的方法和装置

(57)摘要

本公开的实施例公开了用于输出信息的方法和装置。该方法的一具体实施方式包括：获取待配图的文本信息；将文本信息输入预先训练的分类模型，得到文本信息的主题词；将文本信息与预设的实体列表进行匹配，得到文本信息的实体词；从预设的图片库中获取与主题词和实体词匹配的图片并输出图片。该实施方式能够快速、高效的根据文本内容检索相关图片。



CN 110222189 A

1.一种用于输出信息的方法,包括:

获取待配图的文本信息;

将所述文本信息输入预先训练的分类模型,得到所述文本信息的主题词;

将所述文本信息与预设的实体列表进行匹配,得到所述文本信息的实体词;

从预设的图片库中获取与所述主题词和所述实体词匹配的图片并输出所述图片。

2.根据权利要求1所述的方法,其中,所述分类模型是通过以下步骤训练得到的:

获取训练样本集合,其中,训练样本包括文档和文档的类别标签;

将所述训练样本集合中的训练样本的文档和文档的类别标签分别作为fasttext模型的输入和输出,训练得到分类模型。

3.根据权利要求1所述的方法,其中,所述将所述文本信息与预设的实体列表进行匹配,得到所述文本信息的实体词,包括:

将所述文本信息与预设的实体列表进行匹配,得到至少一个候选实体词;

确定各候选实体词在所述文本信息中出现的频次;

将频次最高的预定数目个候选实体词确定为所述文本信息的实体词。

4.根据权利要求1所述的方法,其中,所述将所述文本信息与预设的实体列表进行匹配,得到所述文本信息的实体词,包括:

将所述文本信息的标题与预设的实体列表进行匹配,得到至少一个标题候选实体词;

将所述文本信息的正文与预设的实体列表进行匹配,得到至少一个正文候选实体词;

确定各标题候选实体词和正文候选实体词在所述文本信息中出现的频次;

分别根据标题候选实体词的权重和正文候选实体词的权重对频次进行加权,得到各标题候选实体词的权重值和正文候选实体词的权重值,其中,标题候选实体词的权重大于正文候选实体词的权重;

将权重值最高的预定数目个标题候选实体词或正文候选实体词确定为所述文本信息的实体词。

5.根据权利要求1所述的方法,其中,所述图片库的构造过程包括:

通过搜索引擎利用所述主题词和所述实体词作为关键词搜索图片集合;

从所述图片集合过滤掉低质图片后存储到图片库中。

6.根据权利要求1所述的方法,其中,所述图片库的构造过程包括:

从预设的图片站点抓取图片集合;

从所述图片集合过滤掉低质图片;

根据图片的关键字将过滤后的图片集合中的图片分类存储到图片库中。

7.根据权利要求6所述的方法,其中,所述图片库的构造过程包括:

从网络获取图片;

将所述图片输入预先训练的VGG模型,得到分类标签;

将所述图片与所述分类标签关联存储到图片库中。

8.一种用于输出信息的装置,包括:

获取单元,被配置成获取待配图的文本信息;

主题确定单元,被配置成将所述文本信息输入预先训练的分类模型,得到所述文本信息的主题词;

实体确定单元，被配置成将所述文本信息与预设的实体列表进行匹配，得到所述文本信息的实体词；

匹配单元，被配置成从预设的图片库中获取与所述主题词和所述实体词匹配的图片并输出所述图片。

9.根据权利要求8所述的装置，其中，所述分类模型是通过以下步骤训练得到的：

获取训练样本集合，其中，训练样本包括文档和文档的类别标签；

将所述训练样本集合中的训练样本的文档和文档的类别标签分别作为fasttext模型的输入和输出，训练得到分类模型。

10.根据权利要求8所述的装置，其中，所述实体确定单元进一步被配置成：

将所述文本信息与预设的实体列表进行匹配，得到至少一个候选实体词；

确定各候选实体词在所述文本信息中出现的频次；

将频次最高的预定数目个候选实体词确定为所述文本信息的实体词。

11.根据权利要求8所述的装置，其中，所述实体确定单元进一步被配置成：

将所述文本信息的标题与预设的实体列表进行匹配，得到至少一个标题候选实体词；

将所述文本信息的正文与预设的实体列表进行匹配，得到至少一个正文候选实体词；

确定各标题候选实体词和正文候选实体词在所述文本信息中出现的频次；

分别根据标题候选实体词的权重和正文候选实体词的权重对频次进行加权，得到各标题候选实体词的权重值和正文候选实体词的权重值，其中，标题候选实体词的权重大于正文候选实体词的权重；

将权重值最高的预定数目个标题候选实体词或正文候选实体词确定为所述文本信息的实体词。

12.根据权利要求8所述的装置，其中，所述图片库的构造过程包括：

通过搜索引擎利用所述主题词和所述实体词作为关键词搜索图片集合；

从所述图片集合过滤掉低质图片后存储到图片库中。

13.根据权利要求8所述的装置，其中，所述图片库的构造过程包括：

从预设的图片站点抓取图片集合；

从所述图片集合过滤掉低质图片；

根据图片的关键字将过滤后的图片集合中的图片分类存储到图片库中。

14.根据权利要求13所述的装置，其中，所述图片库的构造过程包括：

从网络获取图片；

将所述图片输入预先训练的VGG模型，得到分类标签；

将所述图片与所述分类标签关联存储到图片库中。

15.一种电子设备，包括：

一个或多个处理器；

存储装置，其上存储有一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-7中任一所述的方法。

16.一种计算机可读介质，其上存储有计算机程序，其中，所述程序被处理器执行时实现如权利要求1-7中任一所述的方法。

用于输出信息的方法和装置

技术领域

[0001] 本公开的实施例涉及计算机技术领域，具体涉及用于输出信息的方法和装置。

背景技术

[0002] 信息时代，数据变得越来越重要。随着数据爆炸式增长，想要快速获取到自己想要的数据成本也越来越高，特别是针对无法形象描述的图片，获取的难度更大。

[0003] 目前市场上的文章配图，大多是直接通过人工搜索并筛选来获取，但这会造成大量人力财力的消耗，也不一定拿到优质的图片。现有方法的主要问题在于：(1)偏主观性：对于人工筛选的方式，不同的主观偏见往往会影响图片配置情况，影响文本和图片的匹配度。(2)局限性：由于人力有限，往往不能对所有的文本进行复述。(3)复杂低效，人工检索图片进行配图，成本高效能低。

发明内容

[0004] 本公开的实施例提出了用于输出信息的方法和装置。

[0005] 第一方面，本公开的实施例提供了一种用于输出信息的方法，包括：获取待配图的文本信息；将文本信息输入预先训练的分类模型，得到文本信息的主题词；将文本信息与预设的实体列表进行匹配，得到文本信息的实体词；从预设的图片库中获取与主题词和实体词匹配的图片并输出图片。

[0006] 在一些实施例中，分类模型是通过以下步骤训练得到的：获取训练样本集合，其中，训练样本包括文档和文档的类别标签；将训练样本集合中的训练样本的文档和文档的类别标签分别作为fasttext模型的输入和输出，训练得到分类模型。

[0007] 在一些实施例中，将文本信息与预设的实体列表进行匹配，得到文本信息的实体词，包括：将文本信息与预设的实体列表进行匹配，得到至少一个候选实体词；确定各候选实体词在文本信息中出现的频次；将频次最高的预定数目个候选实体词确定为文本信息的实体词。

[0008] 在一些实施例中，将文本信息与预设的实体列表进行匹配，得到文本信息的实体词，包括：将文本信息的标题与预设的实体列表进行匹配，得到至少一个标题候选实体词；将文本信息的正文与预设的实体列表进行匹配，得到至少一个正文候选实体词；确定各标题候选实体词和正文候选实体词在文本信息中出现的频次；分别根据标题候选实体词的权重和正文候选实体词的权重对频次进行加权，得到各标题候选实体词的权重值和正文候选实体词的权重值，其中，标题候选实体词的权重大于正文候选实体词的权重；将权重值最高的预定数目个标题候选实体词或正文候选实体词确定为文本信息的实体词。

[0009] 在一些实施例中，图片库的构造过程包括：通过搜索引擎利用主题词和实体词作为关键词搜索图片集合；从图片集合过滤掉低质图片后存储到图片库中。

[0010] 在一些实施例中，图片库的构造过程包括：从预设的图片站点抓取图片集合；从图片集合过滤掉低质图片；根据图片的关键字将过滤后的图片集合中的图片分类存储到图片

库中。

[0011] 在一些实施例中，图片库的构造过程包括：从网络获取图片；将图片输入预先训练的VGG模型，得到分类标签；将图片与分类标签关联存储到图片库中。

[0012] 第二方面，本公开的实施例提供了一种用于输出信息的装置，包括：获取单元，被配置成获取待配图的文本信息；主题确定单元，被配置成将文本信息输入预先训练的分类模型，得到文本信息的主题词；实体确定单元，被配置成将文本信息与预设的实体列表进行匹配，得到文本信息的实体词；匹配单元，被配置成从预设的图片库中获取与主题词和实体词匹配的图片并输出图片。

[0013] 在一些实施例中，分类模型是通过以下步骤训练得到的：获取训练样本集合，其中，训练样本包括文档和文档的类别标签；将训练样本集合中的训练样本的文档和文档的类别标签分别作为fasttext模型的输入和输出，训练得到分类模型。

[0014] 在一些实施例中，实体确定单元进一步被配置成：将文本信息与预设的实体列表进行匹配，得到至少一个候选实体词；确定各候选实体词在文本信息中出现的频次；将频次最高的预定数目个候选实体词确定为文本信息的实体词。

[0015] 在一些实施例中，实体确定单元进一步被配置成：将文本信息的标题与预设的实体列表进行匹配，得到至少一个标题候选实体词；将文本信息的正文与预设的实体列表进行匹配，得到至少一个正文候选实体词；确定各标题候选实体词和正文候选实体词在文本信息中出现的频次；分别根据标题候选实体词的权重和正文候选实体词的权重对频次进行加权，得到各标题候选实体词的权重值和正文候选实体词的权重值，其中，标题候选实体词的权重大于正文候选实体词的权重；将权重值最高的预定数目个标题候选实体词或正文候选实体词确定为文本信息的实体词。

[0016] 在一些实施例中，图片库的构造过程包括：通过搜索引擎利用主题词和实体词作为关键词搜索图片集合；从图片集合过滤掉低质图片后存储到图片库中。

[0017] 在一些实施例中，图片库的构造过程包括：从预设的图片站点抓取图片集合；从图片集合过滤掉低质图片；根据图片的关键字将过滤后的图片集合中的图片分类存储到图片库中。

[0018] 在一些实施例中，图片库的构造过程包括：从网络获取图片；将图片输入预先训练的VGG模型，得到分类标签；将图片与分类标签关联存储到图片库中。

[0019] 第三方面，本公开的实施例提供了一种电子设备，包括：一个或多个处理器；存储装置，其上存储有一个或多个程序，当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现如第一方面中任一的方法。

[0020] 第四方面，本公开的实施例提供了一种计算机可读介质，其上存储有计算机程序，其中，程序被处理器执行时实现如第一方面中任一的方法。

[0021] 本公开的实施例提供的用于输出信息的方法和装置，利用神经网络模型进行图片分类，并结合文本主题挖掘&实体识别，快速、高效的根据文本内容检索相关图片。

附图说明

[0022] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本公开的其它特征、目的和优点将会变得更明显：

[0023] 图1是本公开的一个实施例可以应用于其中的示例性系统架构图；

[0024] 图2是根据本公开的用于输出信息的方法的一个实施例的流程图；

[0025] 图3是根据本公开的用于输出信息的方法的一个应用场景的示意图；

[0026] 图4是根据本公开的用于输出信息的方法的又一个实施例的流程图；

[0027] 图5是根据本公开的用于输出信息的装置的一个实施例的结构示意图；

[0028] 图6是适于用来实现本公开的实施例的电子设备的计算机系统的结构示意图。

具体实施方式

[0029] 下面结合附图和实施例对本公开作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

[0030] 需要说明的是，在不冲突的情况下，本公开中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本公开。

[0031] 图1示出了可以应用本公开的用于输出信息的方法或用于输出信息的装置的实施例的示例性系统架构100。

[0032] 如图1所示，系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0033] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用，例如音频转文字应用、网页浏览器应用、购物类应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等。

[0034] 终端设备101、102、103可以是硬件，也可以是软件。当终端设备101、102、103为硬件时，可以是具有麦克风、显示屏并且支持音频转文字的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving Picture ExpertsGroup Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、膝上型便携计算机和台式计算机等等。当终端设备101、102、103为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块(例如用来提供分布式服务)，也可以实现成单个软件或软件模块。在此不做具体限定。

[0035] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103上显示的文字提供支持的后台编辑服务器。后台编辑服务器可以对接收到的文本信息等数据进行分析等处理，并将处理结果(例如文本信息的配图)反馈给终端设备。

[0036] 需要说明的是，服务器可以是硬件，也可以是软件。当服务器为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器为软件时，可以实现成多个软件或软件模块(例如用来提供分布式服务的多个软件或软件模块)，也可以实现成单个软件或软件模块。在此不做具体限定。

[0037] 需要说明的是，本公开的实施例所提供的用于输出信息的方法可以由终端设备101、102、103执行，也可以由服务器105执行。相应地，用于输出信息的装置可以设置于终端

设备101、102、103中，也可以设置于服务器105中。在此不做具体限定。

[0038] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0039] 继续参考图2，示出了根据本公开的用于输出信息的方法的一个实施例的流程200。该用于输出信息的方法，包括以下步骤：

[0040] 步骤201，获取待配图的文本信息。

[0041] 在本实施例中，用于输出信息的方法的执行主体(例如图1所示的服务器)可以通过有线连接方式或者无线连接方式从用户利用其进行配图的终端接收待配图的文本信息。或者接收音频文件，再由服务器将音频文件通过语音识别技术转换成文本信息。

[0042] 步骤202，将文本信息输入预先训练的分类模型，得到文本信息的主题词。

[0043] 在本实施例中，用于文本分类的分类模型可包括基于规则的模型、基于概率的模型、基于几何的模型、基于统计的模型等。分类的结果得到文本信息的类别即主题词。主题词可包括社会、体育、娱乐等。分类模型可包括SVM，Logistic Regression和neural network等模型。

[0044] 在本实施例的一些可选的实现方式中，分类模型是通过以下步骤训练得到的：获取训练样本集合，其中，训练样本包括文档和文档的类别标签；将训练样本集合中的训练样本的文档和文档的类别标签分别作为fasttext模型的输入和输出，训练得到分类模型。fastText是一种Facebook AI Research在16年开源的一个文本分类器。

[0045] 文本主题挖掘使用fasttext模型，具体操作如下。

[0046] 1、首先标注一定数量的文档及其分类数据，文档可以是新闻；

[0047] 2、针对文档切词并去掉停用词(例如，的、地、得等无实际意义的词)，每个词用空格分割，作为模型的特征；

[0048] 3、将文档的类别名前追加”__label__”，作为模型训练的标签；

[0049] 4、将样本数据按照2、3步构造成标准格式，对模型进行训练，即可得到训练好的模型。

[0050] 步骤203，将文本信息与预设的实体列表进行匹配，得到文本信息的实体词。

[0051] 在本实施例中，实体词包括待配图文本中出现的人名、地名、机构名、媒体、作者及文本的主题关键词。为了提升实体词抽取的效率，可以预先离线挖掘全网的地名、实物、图片集中的标签等，经过网页抓取、数据清洗后，形成实体列表。而后可以根据离线挖掘到的实体列表，对待配图文本信息进行实体词抽取，得到第一实体词。可以根据计算待配图文本正文中各实体词的词频-逆向文件频率(Term Frequency-Inverse Document Frequency，简称TF-IDF)，根据TF-IDF，选取权重在待配图文本中较高的第二实体词。，在确定第一实体词和第二实体词后，可以根据第一实体词和第二实体词，生成最终的实体词。由此，生成的最终的实体词更能表达待配图文本的配图意图。

[0052] 可选地，根据离线挖掘到的词表，对待配图文本进行实体词抽取，得到候选实体词，可将在文本信息中出现的频次最高的预定数目个候选实体词作为最终的实体词。例如，只选择出现频率最高的地名、实物、图片集中的标签等作为实体词。

[0053] 步骤204，从预设的图片库中获取与主题词和实体词匹配的图片并输出图片。

[0054] 在本实施例中，这里的图片可以基于关键字通过百度检索、谷歌检索，也可以在相

关的图片站点抓取，根据其指定的图片标签，进行归类。还可获取到无类别标签的图片，然后通过图片分类模型归类，得到图片标签。将主题词和实体词分别与图片的关键字、图片标签进行匹配。将匹配度最高的图片作为配图进行输出。如果有多个实体词，则可得到多个配图。

[0055] 在本实施例的一些可选的实现方式中，图片库的构造过程包括：通过搜索引擎利用主题词和实体词作为关键词搜索图片集合；从图片集合过滤掉低质图片后存储到图片库中。这里的图片可以基于关键字通过百度检索、谷歌检索，将检索到的图片标记上关键字，用于后续的主题词和实体词的匹配。然后基于KG(Knowledge Graph，知识图谱)等图片判别相关工具进行图片低质判别，主要包括图片尺寸、清晰度、是否有水印等判断。根据图片的关键字将过滤后的图片集合中的图片分类存储到图片库中。

[0056] 在本实施例的一些可选的实现方式中，图片库的构造过程包括：从预设的图片站点抓取图片集合；从图片集合过滤掉低质图片；根据图片的关键字将过滤后的图片集合中的图片分类存储到图片库中。可以在相关的图片站点抓取，根据其指定的图片标签，进行归类。然后基于KG(Knowledge Graph，知识图谱)等图片判别相关工具进行图片低质判别，主要包括图片尺寸、清晰度、是否有水印等判断。根据图片的关键字将过滤后的图片集合中的图片分类存储到图片库中。

[0057] 在本实施例的一些可选的实现方式中，图片库的构造过程包括：从网络获取图片；将图片输入预先训练的VGG模型，得到分类标签；将图片与分类标签关联存储到图片库中。VGG模型是计算机视觉组(Visual Geometry Group)和Google DeepMind公司的研究员一起研究的深度卷积神经网络。具体过程如下：

[0058] 1、图片预处理：将图片resize到256*256像素的图片，等比例缩放，多出的区域用空白填充；

[0059] 2、使用图片作为输入，类目作为标签；

[0060] 3、使用VGG模型进行训练，最终得到训练好的图片分类模型。可选地，可采用其它神经网络模型来进行图片分类，例如GoogleNet等基于图像语义分析的分类模型。

[0061] 训练好后的模型可以直接用作图片挖掘，及从网络挖掘高质图片，通过模型进行分类，将分类准确的图片存入相关图片库。

[0062] 继续参见图3，图3是根据本实施例的用于输出信息的方法的应用场景的一个示意图。在图3的应用场景中，用户通过以下至少一种方式构建图片库：1、将确认出的主题词和实体词作为关键词通过搜索引擎进行图片搜索，然后从搜索结果中过滤掉低质图片后保存到图片库中。2、在相关的图片站点抓取，根据图片关键字分类，然后过滤掉低质图片保存到图片库中。3、从网络挖掘高质图片，然后通过图片分类模型进行分类，将分类准确的图片存入图片库。当需要给文本信息配图时，根据文本信息进行主题挖掘和实体挖掘。根据挖掘出的主题词和实体词在图片库中查找匹配的图片并输出。

[0063] 本公开的上述实施例提供的方法，通过图片分类模型进行图片分类等方式构建图片库，并结合文本主题挖掘&实体识别，快速、高效的根据文本内容从图片库中检索相关图片。

[0064] 进一步参考图4，其示出了用于输出信息的方法的又一个实施例的流程400。该用于输出信息的方法的流程400，包括以下步骤：

[0065] 步骤401，获取待配图的文本信息。

[0066] 步骤402，将文本信息输入预先训练的分类模型，得到文本信息的主题词。

[0067] 步骤401-402与步骤201-202基本相同，因此不再赘述。

[0068] 步骤403，将文本信息的标题与预设的实体列表进行匹配，得到至少一个标题候选实体词。

[0069] 在本实施例中，文本信息通常是按一定格式排版的，第一行位于中间位置的内容通常是标题。或者存储文本信息的文件名通常是文本信息的标题。实体词包括待配图文本中出现的人名、地名、机构名、媒体、作者及文本的主题关键词。为了提升实体词抽取的效率，可以预先离线挖掘全网的地名、实物、图片集中的标签等，经过网页抓取、数据清洗后，形成实体列表。而后可以根据离线挖掘到的实体列表，对待配图文本信息的标题进行实体词抽取，得到至少一个标题候选实体词。

[0070] 步骤404，将文本信息的正文与预设的实体列表进行匹配，得到至少一个正文候选实体词。

[0071] 在本实施例中，文本信息中除标题之外的主体内容就是正文了。将正文与步骤403中提到的实体列表进行匹配，得到至少一个正文候选实体词。正文候选实体词与标题候选实体词之间有可能有交集。

[0072] 步骤405，确定各标题候选实体词和正文候选实体词在文本信息中出现的频次。

[0073] 在本实施例中，分别统计两种类型的候选实体词在文本信息中出现的频次。正文候选实体词与标题候选实体词之间重复的词先分别统计，后续步骤中再加权合并。

[0074] 步骤406，分别根据标题候选实体词的权重和正文候选实体词的权重对频次进行加权，得到各标题候选实体词的权重值和正文候选实体词的权重值，其中，标题候选实体词的权重大于正文候选实体词的权重。

[0075] 在本实施例中，可针对标题候选实体词和正文候选实体词，我们设置不同的权重，例如，标题为2，正文为1，对整个文档的实体权重进行线性加权，然后得到整个文档的不同实体的权重值。对于一个实体词既是标题候选实体词和正文候选实体词的情况，将该实体词的两个权重值之和作为该实体词的权重值。

[0076] 步骤407，将权重值最高的预定数目个标题候选实体词或正文候选实体词确定为文本信息的实体词。

[0077] 在本实施例中，按照权重值由大到小顺序排序，排在前面的预定数目个实体词为文本信息的实体词。通常情况下预定数目个设置为1。

[0078] 步骤408，从预设的图片库中获取与主题词和实体词匹配的图片并输出图片。

[0079] 在本实施例中，这里的图片可以基于关键字通过百度检索、谷歌检索，也可以在相关的图片站点抓取，根据其指定的图片标签，进行归类。还可获取到无类别标签的图片，然后通过图片分类模型归类，得到图片标签。将主题词和实体词分别与图片的关键字、图片标签进行匹配。将匹配度最高的图片作为配图进行输出。如果有多个实体词，则可得到多个配图。

[0080] 从图4中可以看出，与图2对应的实施例相比，本实施例中的用于输出信息的方法的流程400体现了对实体词进行挖掘的步骤。由此，本实施例描述的方案可以引入更多的实体词相关数据，从而实现更全面的候选图片的选取和更有效的配图。

[0081] 进一步参考图5，作为对上述各图所示方法的实现，本公开提供了一种用于输出信息的装置的一个实施例，该装置实施例与图2所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0082] 如图5所示，本实施例的用于输出信息的装置500包括：获取单元501、主题确定单元502、实体确定单元503、匹配单元504。其中，获取单元501，被配置成获取待配图的文本信息；主题确定单元502，被配置成将文本信息输入预先训练的分类模型，得到文本信息的主题词；实体确定单元503，被配置成将文本信息与预设的实体列表进行匹配，得到文本信息的实体词；匹配单元504，被配置成从预设的图片库中获取与主题词和实体词匹配的图片并输出图片。

[0083] 在本实施例的一些可选的实现方式中，分类模型是通过以下步骤训练得到的：获取训练样本集合，其中，训练样本包括文档和文档的类别标签；将训练样本集合中的训练样本的文档和文档的类别标签分别作为fasttext模型的输入和输出，训练得到分类模型。

[0084] 在本实施例的一些可选的实现方式中，实体确定单元进一步被配置成：将文本信息与预设的实体列表进行匹配，得到至少一个候选实体词；确定各候选实体词在文本信息中出现的频次；将频次最高的预定数目个候选实体词确定为文本信息的实体词。

[0085] 在本实施例的一些可选的实现方式中，实体确定单元进一步被配置成：将文本信息的标题与预设的实体列表进行匹配，得到至少一个标题候选实体词；将文本信息的正文与预设的实体列表进行匹配，得到至少一个正文候选实体词；确定各标题候选实体词和正文候选实体词在文本信息中出现的频次；分别根据标题候选实体词的权重和正文候选实体词的权重对频次进行加权，得到各标题候选实体词的权重值和正文候选实体词的权重值，其中，标题候选实体词的权重大于正文候选实体词的权重；将权重值最高的预定数目个标题候选实体词或正文候选实体词确定为文本信息的实体词。

[0086] 在本实施例的一些可选的实现方式中，图片库的构造过程包括：通过搜索引擎利用主题词和实体词作为关键词搜索图片集合；从图片集合过滤掉低质图片后存储到图片库中。

[0087] 在本实施例的一些可选的实现方式中，图片库的构造过程包括：从预设的图片站点抓取图片集合；从图片集合过滤掉低质图片；根据图片的关键字将过滤后的图片集合中的图片分类存储到图片库中。

[0088] 在本实施例的一些可选的实现方式中，图片库的构造过程包括：从网络获取图片；将图片输入预先训练的VGG模型，得到分类标签；将图片与分类标签关联存储到图片库中。

[0089] 下面参考图6，其示出了适于用来实现本公开的实施例的电子设备(例如图1中的服务器)600的结构示意图。图6示出的服务器仅仅是一个示例，不应对本公开的实施例的功能和使用范围带来任何限制。

[0090] 如图6所示，电子设备600可以包括处理装置(例如中央处理器、图形处理器等)601，其可以根据存储在只读存储器(ROM)602中的程序或者从存储装置608加载到随机访问存储器(RAM)603中的程序而执行各种适当的动作和处理。在RAM 603中，还存储有电子设备600操作所需的各种程序和数据。处理装置601、ROM 602以及RAM 603通过总线604彼此相连。输入/输出(I/O)接口605也连接至总线604。

[0091] 通常，以下装置可以连接至I/O接口605：包括例如触摸屏、触摸板、键盘、鼠标、摄

像头、麦克风、加速度计、陀螺仪等的输入装置606;包括例如液晶显示器(LCD)、扬声器、振动器等的输出装置607;包括例如磁带、硬盘等的存储装置608;以及通信装置609。通信装置609可以允许电子设备600与其他设备进行无线或有线通信以交换数据。虽然图6示出了具有各种装置的电子设备600,但是应理解的是,并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图6中示出的每个方框可以代表一个装置,也可以根据需要代表多个装置。

[0092] 特别地,根据本公开的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本公开的实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信装置609从网络上被下载和安装,或者从存储装置608被安装,或者从ROM 602被安装。在该计算机程序被处理装置601执行时,执行本公开的实施例的方法中限定的上述功能。需要说明的是,本公开的实施例所述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开的实施例中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开的实施例中,计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:电线、光缆、RF(射频)等等,或者上述的任意合适的组合。

[0093] 上述计算机可读介质可以是上述电子设备中所包含的;也可以是单独存在,而未装配入该电子设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被该电子设备执行时,使得该电子设备:获取待配图的文本信息;将文本信息输入预先训练的分类模型,得到文本信息的主题词;将文本信息与预设的实体列表进行匹配,得到文本信息的实体词;从预设的图片库中获取与主题词和实体词匹配的图片并输出图片。

[0094] 可以以一种或多种程序设计语言或其组合来编写用于执行本公开的实施例的操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++,还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0095] 附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0096] 描述于本公开的实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器包括获取单元、主题确定单元、实体确定单元和匹配单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，例如，获取单元还可以被描述为“获取待配图的文本信息的单元”。

[0097] 以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离所述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

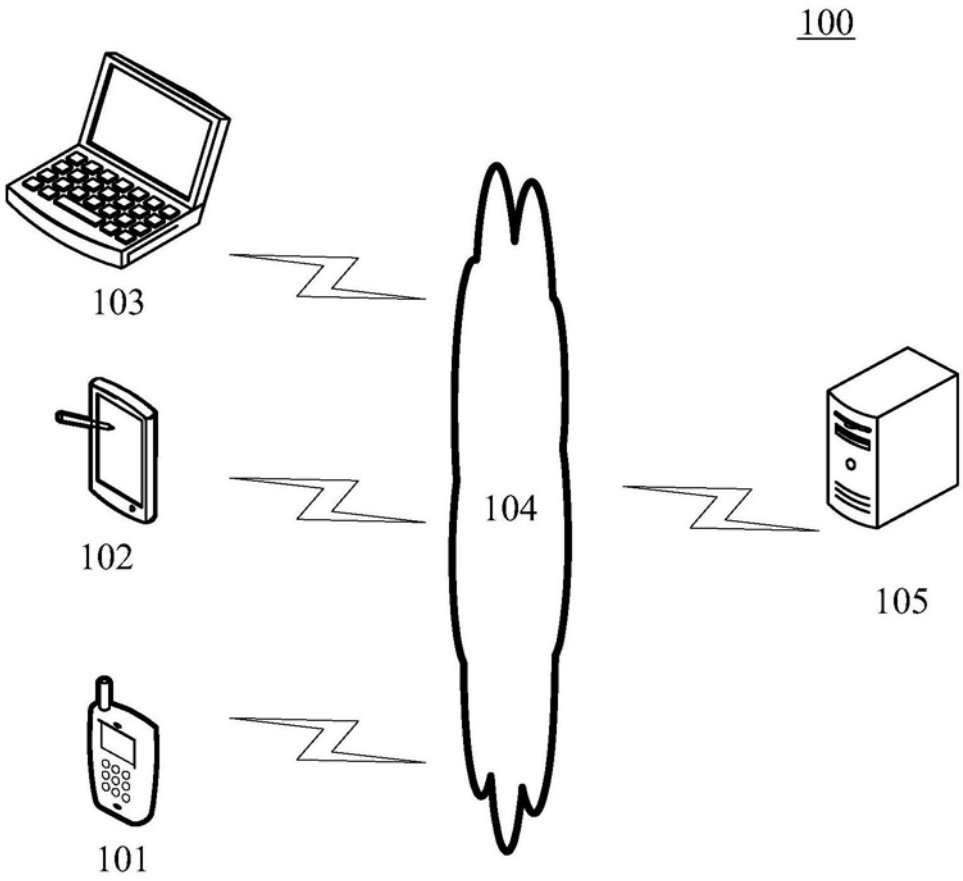


图1

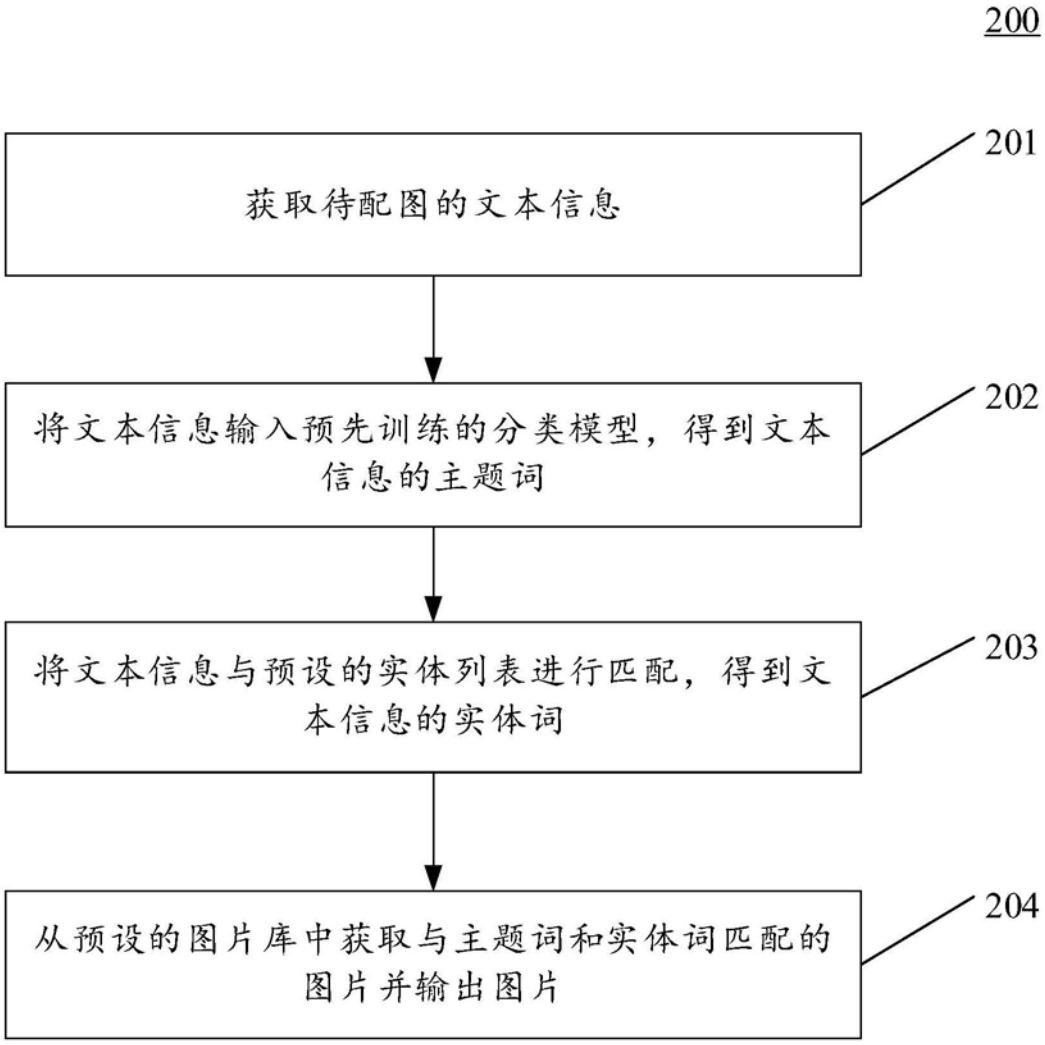

200
201
获取待配图的文本信息
202
将文本信息输入预先训练的分类模型，得到文本信息的主题词
203
将文本信息与预设的实体列表进行匹配，得到文本信息的实体词
204
从预设的图片库中获取与主题词和实体词匹配的图片并输出图片

图2

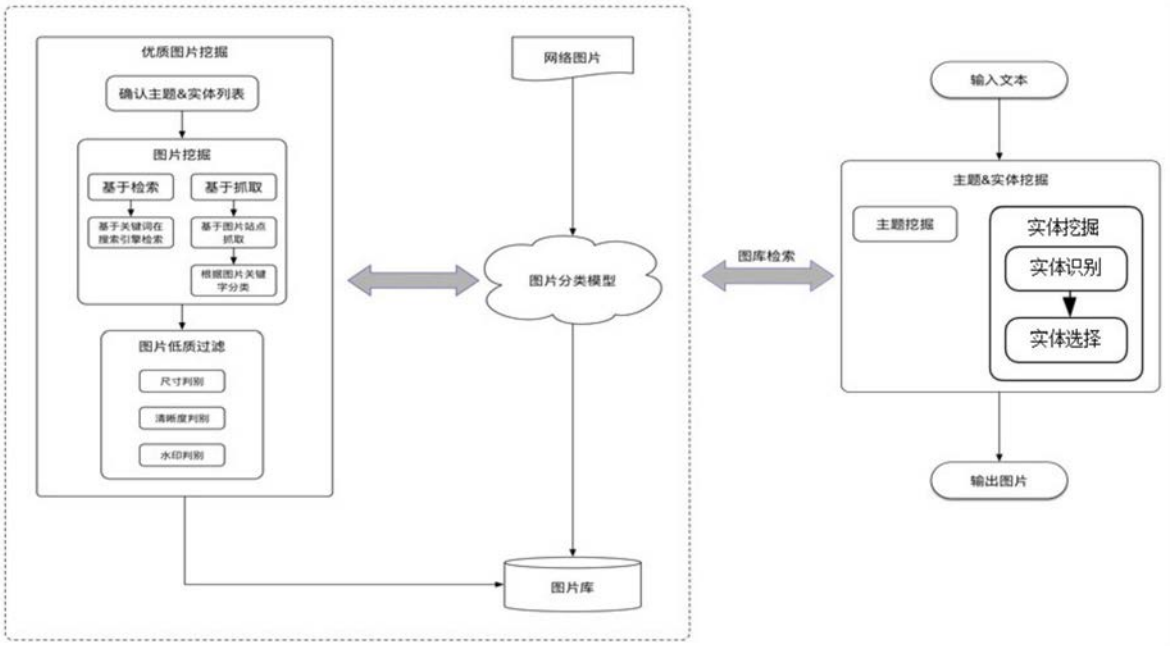

优质图片挖掘
确认主题&实体列表
图片挖掘
基于检索
基于抓取
基于关键词在搜索引擎检索
基于图片站点抓取
根据图片关键字分类
图片低质过滤
尺寸判别
清晰度判别
水印判别
网络图片
图片分类模型
图片库
图库检索
输入文本
主题&实体挖掘
主题挖掘
实体挖掘
实体识别
实体选择
输出图片

图3

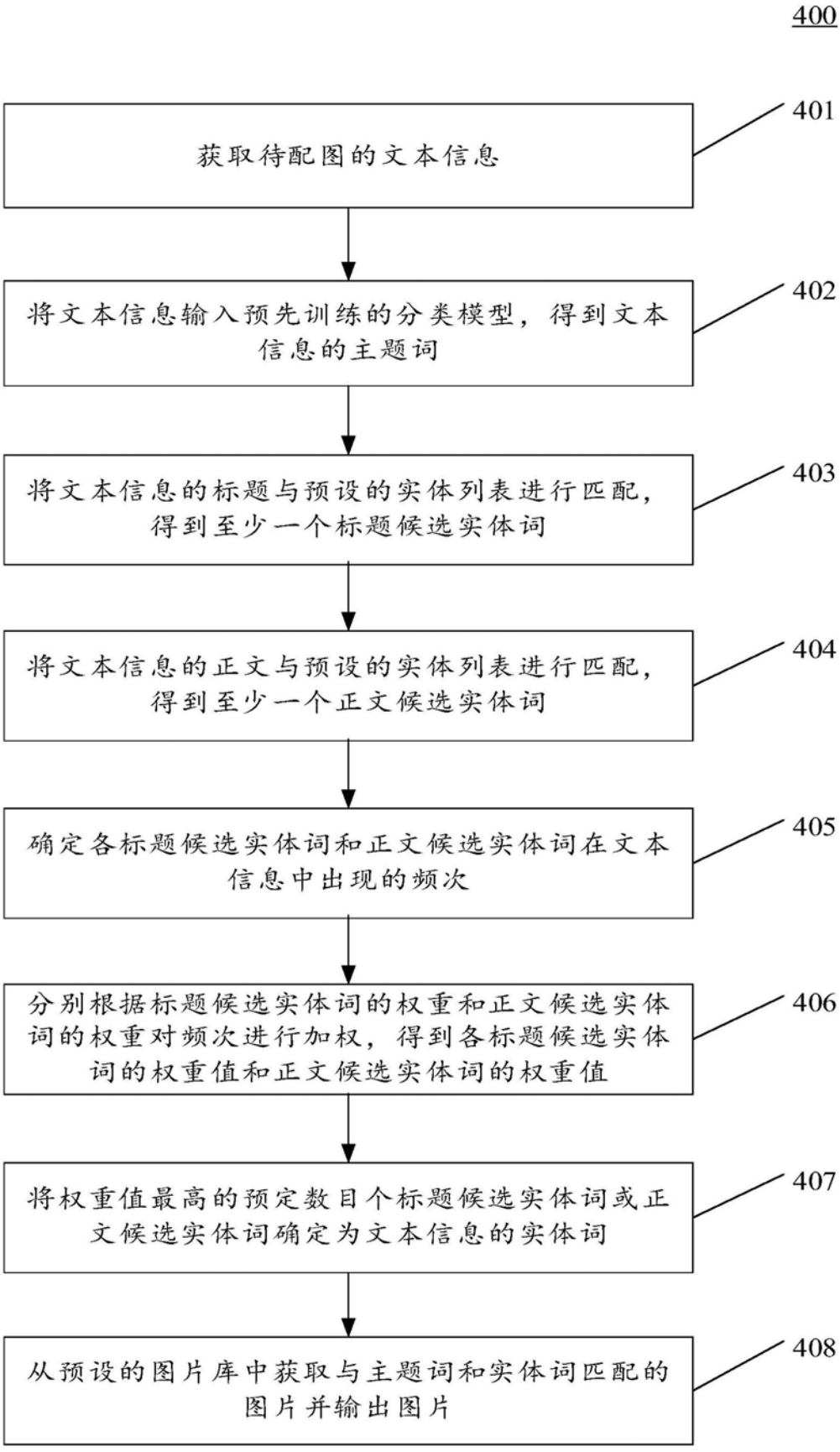

400
获取待配图的文本信息
401
将文本信息输入预先训练的分类模型，得到文本信息的主题词
402
将文本信息的标题与预设的实体列表进行匹配，得到至少一个标题候选实体词
403
将文本信息的正文与预设的实体列表进行匹配，得到至少一个正文候选实体词
404
确定各标题候选实体词和正文候选实体词在文本信息中出现的频次
405
分别根据标题候选实体词的权重和正文候选实体词的权重对频次进行加权，得到各标题候选实体词的权重值和正文候选实体词的权重值
406
将权重值最高的预定数目个标题候选实体词或正文候选实体词确定为文本信息的实体词
407
从预设的图片库中获取与主题词和实体词匹配的图片并输出图片
408

图4

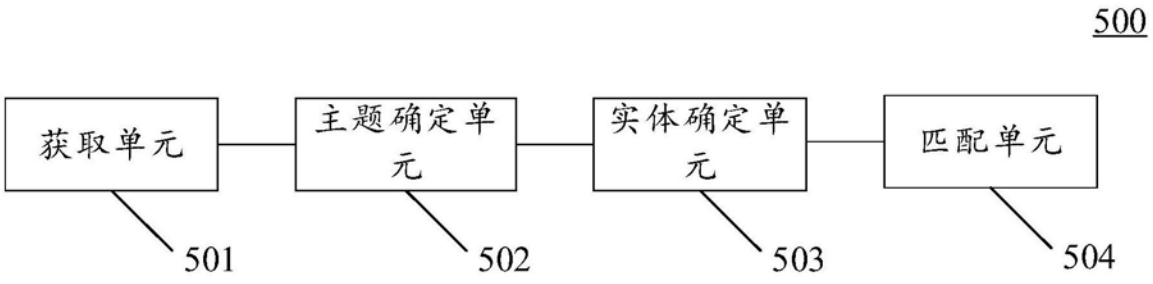

500
获取单元
主题确定单元
实体确定单元
匹配单元
501
502
503
504

图5

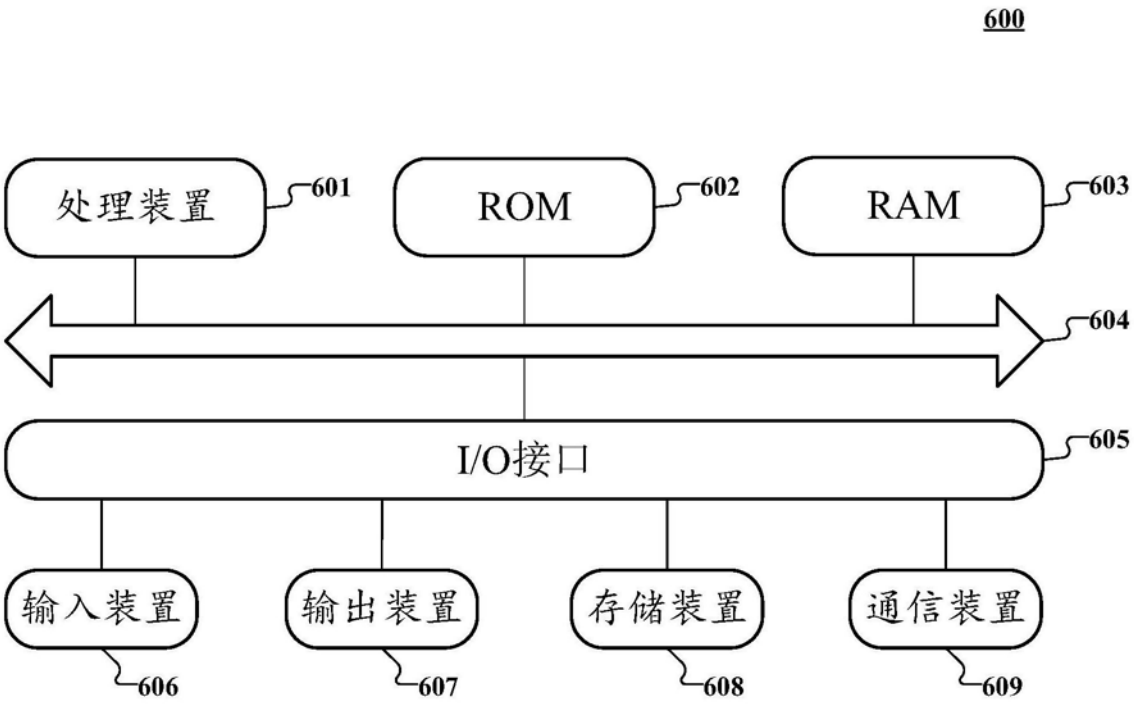

600
处理装置
601
ROM
602
RAM
603
604
I/O接口
605
输入装置
606
输出装置
607
存储装置
608
通信装置
609

图6