

(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



(10) 国际公布号

WO 2020/151318 A1

(43) 国际公布日

2020 年 7 月 30 日 (30.07.2020)

(51) 国际专利分类号:

G06F 16/332 (2019.01) G06Q 30/02 (2012.01)

G06F 16/35 (2019.01)

(21) 国际申请号:

PCT/CN2019/117698

(22) 国际申请日:

2019 年 11 月 12 日 (12.11.2019)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

201910065779.8 2019 年 1 月 24 日 (24.01.2019) 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: 深圳市世联合知识产权代理有限公司(SL
INTELLECTUAL PROPERTY CO., LTD.); 中国广
东省深圳市福田区泰然六路泰然科技园 205 栋
二层 203-206, 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: CORPUS CONSTRUCTION METHOD AND APPARATUS BASED ON CRAWLER MODEL, AND COMPUTER DEVICE

(54) 发明名称: 基于爬虫模型的语料构建方法、装置及计算机设备

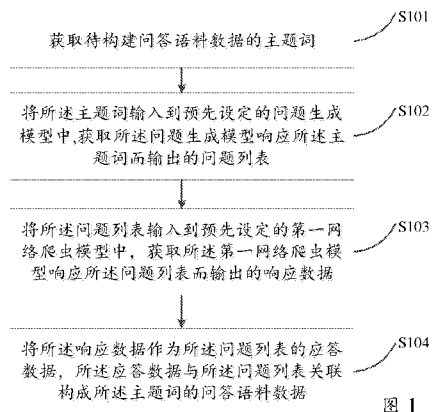


图 1

S101 Acquire subject words of question and answer corpus data to be constructed
S102 Input the subject words into a predetermined question generation model and acquire a question list outputted by the question generation model in response to the subject words
S103 Input the question list into a predetermined first web crawling model and acquire response data outputted by the first web crawling model in response to the question list
S104 Set the response data as answer data of the question list, the answer data and the question list being associated to form the question and answer corpus data of the subject words

(57) Abstract: A corpus construction method and apparatus based on a crawler model, a computer device, and a storage medium, the method comprising the following steps: acquiring subject words of question and answer corpus data to be constructed (S101); inputting the subject words into a predetermined question generation model and acquiring a question list outputted by the question generation model in response to the subject words (S102); inputting the question list into a predetermined first web crawling model and acquiring response data outputted by the first web crawling model in response to the question list (S103); and setting the response data as answer data of the question list, the answer data and the question list being associated to form the question and answer corpus data of the subject words (S104). The answer data acquired by means of the web crawler is real customer service answers. The present method increases the efficiency and quality of corpus construction and also increases the question hit rate of smart customer service, making customer service more artificially intelligent.

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: 一种基于爬虫模型的语料构建方法、装置、计算机设备及存储介质, 其中方法包括下述步骤: 获取待构建问答语料数据的主题词(S101); 将所述主题词输入到预先设定的问题生成模型中, 获取所述问题生成模型响应所述主题词而输出的问题列表(S102); 将所述问题列表输入到预先设定的第一网络爬虫模型中, 获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据(S103); 将所述响应数据作为所述问题列表的应答数据, 所述应答数据与所述问题列表关联构成所述主题词的问答语料数据(S104)。其中, 应答数据通过网络爬虫获取真实的客服应答。该方法提高了语料构建的效率和质量, 也提高了智能客服的问题命中率, 使客服人工智能化。

基于爬虫模型的语料构建方法、装置及计算机设备

本申请以2019年1月24日提交的申请号为201910065779.8，名称为“智能客服的语料构建方法、装置、计算机设备及存储介质”的中国发明专利申请为基础，并要求其优先权。

技术领域

本申请智能客服领域，尤其涉及一种基于爬虫模型的语料构建方法、装置、计算机设备及存储介质。

背景技术

随着人工智能技术的发展，智能客服系统也逐渐兴起。智能客服不仅为企业与海量用户之间建立起便捷的基于自然语言的沟通平台，有效地提高了客服工作的效率，而且能够为企业进行精细化管理提供直接来源于客户信息。

智能客服通常基于既有问答数据库才能提供客服功能，现有的智能客服，在建立问答数据库时需要人工整理知识点，人工扩展用户的问题点，最终生成问答数据库中一问一答的数据。

但是，发明人意识到人工整理知识点和人工扩展用户问题点的方法，费时费力，而且往往不能体现用户真实的热点问题，导致在使用过程中，用户提问命中问答数据库中的问题点的命中率低，使智能客服不能有效地解答用户提问，影响用户体验。

发明内容

本申请提供一种基于爬虫模型的语料构建方法、装置、计算机设备及存储介质，以解决智能客服构建问答语料库费时费力的问题。

为解决上述技术问题，本申请提出一种基于爬虫模型的语料构建方法，包括如下步骤：

获取待构建问答语料数据的主题词；

将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；

将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网

络爬虫模型响应所述问题列表而输出的响应数据；

将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

为解决上述问题，本申请还提供一种基于爬虫模型的问答语料数据构建装置，包括：

获取模块，用于获取待构建问答语料数据的主题词；

生成模块，用于将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；

处理模块，用于将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据，其中所述第一网络爬虫模型以所述问题列表为约束条件抓取目标数据；

执行模块，用于将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

为解决上述技术问题，本申请实施例还提供一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述处理器执行上述所述基于爬虫模型的语料构建方法的步骤。

为解决上述技术问题，本申请实施例还提供一个或多个非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时，使得所述处理器执行上述所述基于爬虫模型的语料构建方法的步骤。

本申请实施例的有益效果为：通过获取待构建问答语料数据的主题词；将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据；将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。其中，关于主题词的问题生成通过网络爬虫自动获取用户真实问题，或通过人工智能学习用户的真实意图生成问题列表，相应的应答数据也通过网络爬虫获取真实的客服应答。本发明提高了问答数据构建的效率和质量，也提高了智能客服的问题命中率。

附图说明

为了更清楚地说明本申请实施例中的技术方案，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，下面描述中的附图仅仅是本申请的一些实施例，对于本领域技术人员来讲，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为本申请实施例一种基于爬虫模型的语料构建方法基本流程示意图；

图2为本申请实施例基于第二网络爬虫模型生成问题列表的流程示意图；

图3为本申请实施例基于Seq2Seq 模型生成问题列表的流程示意图；

图4为本申请实施例基于过滤规则获取应答数据的流程示意图；

图5为本申请实施例基于深度神经网络模型获取应答数据的流程示意图；

图6为本申请实施例深度神经网络模型训练流程示意图；

图7为本申请实施例一种基于爬虫模型的问答语料数据构建装置基本结构框图；

图8为本申请实施计算机设备基本结构框图。

具体实施方式

为了使本技术领域的人员更好地理解本申请方案，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。

在本申请的说明书和权利要求书及上述附图中的描述的一些流程中，包含了按照特定顺序出现的多个操作，但是应该清楚了解，这些操作可以不按照其在本文中出现的顺序来执行或并行执行，操作的序号如 101、102 等，仅仅是用于区分开各个不同的操作，序号本身不代表任何的执行顺序。另外，这些流程可以包括更多或更少的操作，并且这些操作可以按顺序执行或并行执行。需要说明的是，本文中的“第一”、“第二”等描述，是用于区分不同的消息、设备、模块等，不代表先后顺序，也不限定“第一”和“第二”是不同的类型。

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例，都属于本申请保护的范围。

实施例

本技术领域技术人员可以理解，这里所使用的“终端”、“终端设备”既包括无

线信号接收器的设备，其仅具备无发射能力的无线信号接收器的设备，又包括接收和发射硬件的设备，其具有能够在双向通信链路上，执行双向通信的接收和发射硬件的设备。这种设备可以包括：蜂窝或其他通信设备，其具有单线路显示器或多线路显示器或没有多线路显示器的蜂窝或其他通信设备；PCS（Personal Communications Service，个人通信系统），其可以组合语音、数据处理、传真和/或数据通信能力；PDA（Personal Digital Assistant，个人数字助理），其可以包括射频接收器、寻呼机、互联网/内联网访问、网络浏览器、记事本、日历和/或GPS（Global Positioning System，全球定位系统）接收器；常规膝上型和/或掌上型计算机或其他设备，其具有和/或包括射频接收器的常规膝上型和/或掌上型计算机或其他设备。这里所使用的“终端”、“终端设备”可以是便携式、可运输、安装在交通工具（航空、海运和/或陆地）中的，或者适合于和/或配置为在本地运行，和/或以分布形式，运行在地球和/或空间的任何其他位置运行。这里所使用的“终端”、“终端设备”还可以是通信终端、上网终端、音乐/视频播放终端，例如可以是PDA、MID（Mobile Internet Device，移动互联网设备）和/或具有音乐/视频播放功能的移动电话，也可以是智能电视、机顶盒等设备。

本实施方式中的终端即为上述的终端。

具体地，请参阅图1，图1为本实施例一种基于爬虫模型的语料构建方法的基本流程示意图。

如图1所示，一种基于爬虫模型的语料构建方法，包括下述步骤：

S101、获取待构建问答语料数据的主题词；

主题词限定了待构建问答语料数据的主题，通过终端上可交互的页面获取用户输入的主题词。为了构建的问答语料数据更聚焦，建议输入的主题词描述的范围适当小。例如“手机”，涵盖范围较广，构建的问答语料可能较发散，为了问答语料数据更聚焦，可以将主题词限定为“XX型号手机”。

S102、将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；

将获取的主题词输入到预先设定的问题生成模型中，生成基于主题词的问题列表。问题生成模型可以是预先设置的固定的一系列问题，以主题词为参数变化。例如，预先设置的一系列问题为：

主题词	发布时间是什么时候
主题词	销售价格是多少

主题词	销售渠道有哪些
主题词	支持指纹识别吗
主题词	支持多用户登录吗

可以预先设定不同类型的主题词对应不同的问题列表。

本申请实施例中采用网络爬虫模型，获取线上用户的真实提问；或者通过预先训练的 Seq2Seq 模型生成问题列表。具体地，请参阅以下图 2 和图 3 描述。

S103、将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据；

获取到有关主题词的问题列表，将其输入到预先设定的网络爬虫模型中，这里称为第一网络爬虫模型。网络爬虫是一个自动提取网页的程序，具体地，利用 python 程序，模拟浏览器，向目标站点发送请求，目标站点服务器响应请求，返回 html、图片、视频等资源。第一网络爬虫模型以问题列表为检索条件，检索目标站点中与问题列表相关的数据，即第一网络爬虫模型影响问题列表而输出的响应数据。

S104、将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

将响应数据作为问题列表的应答数据，并将问题列表和应答数据按照一问一答形式关联。表现在数据库中，为一条数据包含两部分，一部分为问题，另一部分则为该问题的应答。

当智能客服接收到用户的问题时，可以通过关键词检索的方式，检索问答数据库与用户问题关键词一致的问题，并返回与该问题具有映射关系的应答。在一些实施方式中通过计算用户问题与问答数据库中问题相似度的方式，获取该问题对应的应答。相似度的计算可以采用计算编辑距离的算法计算相似度，例如：问答数据库中保存的问题为“手机卖多少钱”，接收到的用户问题为“手机多少钱”，两者的编辑距离为 1，即“手机多少钱”变“手机卖多少钱”只需插入“卖”。检索数据库中与用户提出的问题相似度最大的问题，返回该问题对应的应答。

如图 2 所示，步骤 S102 具体还包括下述步骤：

S111、将所述主题词输入到第二网络爬虫模型中，获取所述第二网络爬虫模型响应所述主题词而输出的疑问候选数据；

本申请实施例中采用网络爬虫模型获取与输入主题词相关的问题，为区别

与前述网络爬虫模型的区别，这里称为第二网络爬虫模型，该模型以主题词为检索条件，获取目标站点上，与主题词相关的内容，这里称为疑问候选数据。

S112、按照预设的匹配规则对所述疑问候选数据进行匹配，获取疑问匹配数据，其中所述匹配规则至少包含疑问语料匹配规则；

获取的疑问候选数据中包含非疑问语料数据和疑问语料数据，为了得到疑问语料，预设了匹配规则，通过预设的匹配规则对疑问候选数据进行处理，得到疑问匹配数据。匹配规则为包含“？”“什么”“多少”“哪里”“怎么”等表示疑问的语气词。本申请实施例中采用正则匹配的算法，正则表达式是对字符串操作的一种逻辑公式，用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。正则表达式是一种文本模式，描述在搜索文本时要匹配的一个或多个字符串。例如，可以用“*主题词*什么*”这一正则表达式查找包含“主题词”和“什么”的任意字符串。

由于通过第二网络爬虫模型获取的是目标站点记录的用户真实问题，所以通过这种方式获取的问题列表与现实更接近，基于此构建的问答语料数据命中用户实际提问的命中率更高。

S113、将所述疑问匹配数据作为所述主题词的问题列表。

疑问候选数据经过预设的匹配规则后，得到的疑问匹配数据就是与主题词相关的问题列表。

如图3所示，在一些实施方式中，步骤S102具体还包括下述步骤：

S121、将所述主题词输入到预先训练的 Seq2Seq 模型中；

在一些实施方式中，通过将主题词输入到预先训练的 Seq2Seq 模型中获取问题列表。Seq2Seq 模型是一个 Encoder（编码器）- Decoder（解码器）结构的网络，它的输入是一个序列，输出也是一个序列，Encoder 中将一个可变长度的信号序列变为固定长度的向量表达，Decoder 将这个固定长度的向量变成可变长度的目标的信号序列。具体地，先将接收的主题词进行 one-hot 词汇编码，然后，通过 word2vec 词向量模型，转化为对应的词向量，将词向量输入到 Encoder 层，其中，Encoder 层是以双向 LSTM 层或 RNN（卷积神经网络）作为基本的神经元单位的多层神经元层，生成 final_state 状态层和 final_output 状态向量；

然后将以上步骤输出的 encoder 的 final output 状态向量，输入到全局信息层，生成全局状态层 context；

最后将以上步骤得到的 final_state 和 final_output 状态向量和全局信息层生成的 context 向量,输入到 Decoder 层中,输出 decoder 层的 final_state 向量和 output 向量,其中 Decoder 层也是以双向 LSTM 层或 RNN 作为基本的神经元单位的多层神经元层。输出结果即为以输入主题词为主题的基本问题列表。

S122、获取所述 Seq2Seq 模型响应所述主题词而输出的问题列表。

将 Seq2Seq 模型输出的关于主题词的响应作为问题列表。Seq2Seq 模型需要经过训练才具备输出问题列表的功能。具体的训练过程为准备训练语料,即准备输入序列和对应的输出序列,将输入序列输入到 Seq2Seq 模型,计算得到输出序列的概率,调整 Seq2Seq 模型的参数,使整个样本,即所有输入序列经过 Seq2Seq 输出对应输出序列的概率最高。

如图 4 所示,在步骤 S103 之后,还包括下述步骤:

S131、按照预设的过滤规则对所述响应数据进行过滤,获取过滤数据,其中,所述过滤规则至少包含疑问语料数据过滤规则;

在一些实施方式中,对于获取的响应数据进行进一步的处理,由于这里需要获取的应答数据,所以首先需要过滤掉表示疑问的语料。同样可以通过正则匹配算法,将所有包含“什么”“怎么”“多少”等表示疑问语义的语料过滤掉。另外,过滤规则还可以包括敏感词的过滤,根据设定的敏感词表,过滤掉包含敏感词的语料。

S132、将所述过滤数据作为所述问题列表的应答数据。

响应数据经过过滤后得到过滤数据即为问题列表的应答数据。

如图 5 所示, 在一些实施方式中在步骤 S103 之后,还包括下述步骤:

S141、将所述响应数据输入到预先训练的神经网络模型中,获取所述神经网络模型输出的对所述响应数据的分类信息,其中,所述分类信息至少将所述响应数据区分为疑问语料数据和非疑问语料数据;

在一些实施方式中,通过预先训练的神经网络模型对响应数据进行分类,其中预先训练的神经网络模型至少可以识别出疑问语料和非疑问语料。神经网络的具体训练过程请参阅图 6。

S142、将所述非疑问语料数据作为所述问题列表的应答数据。

经神经网络识别后的非语料数据即为问题列表对应的应答数据。

如图 6 所示,在步骤 S141 中使用的神经网络模型按下述步骤训练:

S151、获取标记有语料类别的训练样本,其中所述语料类别至少包含疑问

语料和非疑问语料；

本申请实施例中，深度神经网络模型的训练目标是可以识别出疑问语料和非疑问语料。所以训练样本中包含两类语料，每一个样本均标记有语料类别。

S152、将所述训练样本输入深度卷积神经网络模型获取所述训练样本的参照语料类别；

将样本输入到深度卷积神经网络模型，输出训练样本的参照语料类别，即输出样本语料是疑问语料还是非疑问语料。

S153、比对所述训练样本内不同样本的参照语料类别与所述语料类别是否一致；

通过损失函数判断深度卷积神经网络输出的参照语料类别与样本标注的语料类别是否一致。本申请实施例中，损失函数采用 Softmax 交叉熵损失函数。在训练过程中，调整深度卷积神经网络模型中的权重，使 Softmax 交叉熵损失函数尽可能收敛，也就是说继续调整权重，得到的损失函数的值不再缩小，反而增大时，认为深度卷积神经网络训练可以结束。

假设共有 N 个训练样本，针对网络最后分层第 i 个样本的输入特征为 X_i ，其对应的标记为 Y_i 是最终的分类结果（即样本 i 是疑问句还是非疑问句）， $h = (h_1, h_2, \dots, h_c)$ 为网络的最终输出，即样本 i 的预测结果。其中 C 是最后所有分类的数量。

Softmax 交叉熵损失函数为

$$L_{crossEntropyLoss} = -\frac{1}{N} \log \left(\frac{e^{h_{yi}}}{\sum_{j=1}^c e^{h_j}} \right)$$

S154、当参照语料类别与所述语料类别不一致时，反复循环迭代的更新所述深度神经网络模型中的权重，至所述参照语料类别与所述语料类别一致时结束。

如前所述当损失函数没有收敛时，更新深度卷积神经网络模型中的各节点的权重，本申请实施例中采用梯度下降法，梯度下降法是一个最优化算法，用于机器学习和人工智能当中用来递归性地逼近最小偏差模型。

为解决上述技术问题本申请实施例还提供一种基于爬虫模型的问答语料数据构建装置。具体请参阅图 7，图 7 为本实施例基于爬虫模型的问答语料数据构建装置的基本结构框图。

如图 7 所示，一种基于爬虫模型的问答语料数据构建装置，包括：获取模

块 210、生成模块 220、处理模块 230 和执行模块 240。其中，获取模块 210，用于获取待构建问答语料数据的主题词；生成模块 220，用于将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；处理模块 230，用于将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据，其中所述第一网络爬虫模型以所述问题列表为约束条件抓取目标数据；执行模块 240，用于将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

本申请实施例通过获取待构建问答语料数据的主题词；将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据；将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。其中，关于主题词的问题生成通过网络爬虫自动获取用户真实问题，或通过人工智能学习用户的真实意图生成问题列表，相应的应答数据也通过网络爬虫获取真实的客服应答。本申请提高了问答数据构建的效率和质量，也提高了智能客服的问题命中率。

在一些实施方式中，所述生成模块220还包括：第一处理子模块、第一匹配子模块、第一执行子模块。其中，第一处理子模块，用于将所述主题词输入到第二网络爬虫模型中，获取所述第二网络爬虫模型响应所述主题词而输出的疑问候选数据；第一匹配子模块，用于按照预设的匹配规则对所述疑问候选数据进行匹配，获取疑问匹配数据，其中所述匹配规则至少包含疑问语料匹配规则；第一执行子模块，用于将所述疑问匹配数据作为所述主题词的问题列表。

在一些实施方式中，所述第一匹配子模块中，采用正则匹配算法获取疑问匹配数据。

在一些实施方式中，所述生成模块220还包括：第二处理子模块和第一获取子模块。其中，第二处理子模块，用于将所述主题词输入到预先训练的Seq2Seq模型中；第一获取子模块，用于获取所述Seq2Seq模型响应所述主题词而输出的问题列表。

在一些实施方式中，所述基于爬虫模型的问答语料数据构建装置还包括：，第一过滤子模块和第二执行子模块。其中，第一过滤子模块，用于按照预设的

过滤规则对所述响应数据进行过滤，获取过滤数据，其中，所述过滤规则至少包含疑问语料数据过滤规则；第二执行子模块，用于将所述过滤数据作为所述问题列表的应答数据。

在一些实施方式中，所述基于爬虫模型的问答语料数据构建装置还包括：第一分类子模块和第三执行子模块。其中，第一分类子模块，用于将所述响应数据输入到预先训练的深度神经网络模型中，获取所述深度神经网络模型输出的对所述响应数据的分类信息，其中，所述分类信息至少将所述响应数据区分为疑问语料数据和非疑问语料数据；第三执行子模块，用于将所述非疑问语料数据作为所述问题列表的应答数据。

在一些实施方式中，所述基于爬虫模型的问答语料数据构建装置还包括：第二获取子模块、第三处理子模块、第一比对子模块和第一更新子模块。其中，第二获取子模块，用于获取标记有语料类别的训练样本，其中所述语料类别至少包含疑问语料和非疑问语料；第三处理子模块，用于将所述训练样本输入深度卷积神经网络模型获取所述训练样本的参照语料类别；第一比对子模块，用于比对所述训练样本内不同样本的参照语料类别与所述语料类别是否一致；第一更新子模块，用于当参照语料类别与所述语料类别不一致时，反复循环迭代的更新所述深度神经网络模型中的权重，至所述参照语料类别与所述语料类别一致时结束。

为解决上述技术问题，本申请实施例还提供计算机设备。具体请参阅图 8，图 8 为本实施例计算机设备基本结构框图。

如图 8 所示，计算机设备的内部结构示意图。如图 8 所示，该计算机设备包括通过系统总线连接的处理器、非易失性存储介质、存储器和网络接口。其中，该计算机设备的非易失性存储介质存储有操作系统、数据库和计算机可读指令，数据库中可存储有控件信息序列，该计算机可读指令被处理器执行时，可使得处理器实现一种基于爬虫模型的问答语料数据构建的方法。该计算机设备的处理器用于提供计算和控制能力，支撑整个计算机设备的运行。该计算机设备的存储器中可存储有计算机可读指令，该计算机可读指令被处理器执行时，可使得处理器执行一种基于爬虫模型的问答语料数据构建的方法。该计算机设备的网络接口用于与终端连接通信。本领域技术人员可以理解，图 8 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的限定，具体的计算机设备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。

本实施方式中处理器用于执行图 7 中获取模块 210、生成模块 220、处理模块 230 和执行模块 240 的具体内容，存储器存储有执行上述模块所需的可读指令代码和各类数据。网络接口用于向用户终端或服务器之间的数据传输。本实施方式中的存储器存储有基于爬虫模型的语料构建方法中执行所有子模块所需的可读指令代码及数据，服务器能够调用服务器的可读指令代码及数据执行所有子模块的功能。

计算机设备通过获取待构建问答语料数据的主题词；将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据；将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。其中，关于主题词的问题生成通过网络爬虫自动获取用户真实问题，或通过人工智能学习用户的真实意图生成问题列表，相应的应答数据也通过网络爬虫获取真实的客服应答。本申请提高了问答数据构建的效率和质量，也提高了智能客服的问题命中率。

本申请还提供一个或多个存储有计算机可读指令的非易失性存储介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行上述任一实施例所述基于爬虫模型的语料构建方法的步骤。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，该计算机可读指令可存储于非易失性可读存储介质中，该可读指令在执行时，可包括如上述各方法的实施例的流程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory, ROM）等非易失性存储介质，或随机存储记忆体（Random Access Memory, RAM）等。

应该理解的是，虽然附图的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，其可以以其他的顺序执行。而且，附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，其执行顺序也不必然是依次进行，而是可以与其他步骤或者其他步骤的子

步骤或者阶段的至少一部分轮流或者交替地执行。

以上所述仅是本申请的部分实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

权利要求书

1、一种基于爬虫模型的语料构建方法,其特征在于,包括下述步骤:

获取待构建问答语料数据的主题词;

将所述主题词输入到预先设定的问题生成模型中,获取所述问题生成模型响应所述主题词而输出的问题列表;

将所述问题列表输入到预先设定的第一网络爬虫模型中,获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据;

将所述响应数据作为所述问题列表的应答数据,所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

2、根据权利要求1所述的基于爬虫模型的语料构建方法,其特征在于,在所述将所述主题词输入到预先设定的问题生成模型中,获取所述问题生成模型响应所述主题词而输出的问题列表的步骤中,具体包括下述步骤:

将所述主题词输入到第二网络爬虫模型中,获取所述第二网络爬虫模型响应所述主题词而输出的疑问候选数据;

按照预设的匹配规则对所述疑问候选数据进行匹配,获取疑问匹配数据,其中所述匹配规则至少包含疑问语料匹配规则;

将所述疑问匹配数据作为所述主题词的问题列表。

3、根据权利要求2所述的基于爬虫模型的语料构建方法,其特征在于,在所述按照预设的匹配规则对所述疑问候选数据进行匹配的步骤中,所述匹配的步骤采用正则匹配算法获取疑问匹配数据。

4、根据权利要求1所述的基于爬虫模型的语料构建方法,其特征在于,在所述将所述主题词输入到预先设定的问题生成模型中,获取所述问题生成模型响应所述主题词而输出的问题列表的步骤中,具体包括下述步骤:

将所述主题词输入到预先训练的 Seq2Seq 模型中;

获取所述 Seq2Seq 模型响应所述主题词而输出的问题列表。

5、根据权利要求1所述的基于爬虫模型的语料构建方法,其特征在于,在所述将所述问题列表输入到预先设定的第一网络爬虫模型中,获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据的步骤之后,还包括下述步骤:

按照预设的过滤规则对所述响应数据进行过滤,获取过滤数据,其中,所述过滤规则至少包含疑问语料数据过滤规则;

将所述过滤数据作为所述问题列表的应答数据。

6、根据权利要求1所述的基于爬虫模型的语料构建方法，其特征在于，在所述将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据的步骤之后，还包括下述步骤：

将所述响应数据输入到预先训练的深度神经网络模型中，获取所述深度神经网络模型输出的对所述响应数据的分类信息，其中，所述分类信息至少将所述响应数据区分为疑问语料数据和非疑问语料数据；

将所述非疑问语料数据作为所述问题列表的应答数据。

7、根据权利要求6所述的基于爬虫模型的语料构建方法，其特征在于，所述预先训练的深度神经网络模型是通过以下步骤进行训练：

获取标记有语料类别的训练样本，其中所述语料类别至少包含疑问语料和非疑问语料；

将所述训练样本输入深度卷积神经网络模型获取所述训练样本的参照语料类别；

比对所述训练样本内不同样本的参照语料类别与所述语料类别是否一致；

当参照语料类别与所述语料类别不一致时，反复循环迭代的更新所述深度神经网络模型中的权重，至所述参照语料类别与所述语料类别一致时结束。

8、一种基于爬虫模型的问答语料数据构建装置，其特征在于，包括：

获取模块，用于获取待构建问答语料数据的主题词；

生成模块，用于将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；

处理模块，用于将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据，其中所述第一网络爬虫模型以所述问题列表为约束条件抓取目标数据；

执行模块，用于将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

9、根据权利要求8所述的基于爬虫模型的问答语料数据构建装置，其特征在于，所述生成模块还包括：

第一处理子模块，用于将所述主题词输入到第二网络爬虫模型中，获取所述第二网络爬虫模型响应所述主题词而输出的疑问候选数据；

第一匹配子模块，用于按照预设的匹配规则对所述疑问候选数据进行匹配，

获取疑问匹配数据，其中所述匹配规则至少包含疑问语料匹配规则；

第一执行子模块，用于将所述疑问匹配数据作为所述主题词的问题列表。

10、根据权利要求8所述的基于爬虫模型的问答语料数据构建装置，其特征在于，所述生成模块还包括：

第二处理子模块，用于将所述主题词输入到预先训练的 Seq2Seq 模型中；

第一获取子模块，用于获取所述 Seq2Seq 模型响应所述主题词而输出的问题列表。

11、一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述处理器执行如下步骤：

获取待构建问答语料数据的主题词；

将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；

将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据；

将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

12、根据权利要求11所述的计算机设备，其特征在于，在所述将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表的步骤中，具体包括下述步骤：

将所述主题词输入到第二网络爬虫模型中，获取所述第二网络爬虫模型响应所述主题词而输出的疑问候选数据；

按照预设的匹配规则对所述疑问候选数据进行匹配，获取疑问匹配数据，其中所述匹配规则至少包含疑问语料匹配规则；

将所述疑问匹配数据作为所述主题词的问题列表。

13、根据权利要求12所述的计算机设备，其特征在于，在所述按照预设的匹配规则对所述疑问候选数据进行匹配的步骤中，所述匹配的步骤采用正则匹配算法获取疑问匹配数据。

14、根据权利要求11所述的计算机设备，其特征在于，在所述将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表的步骤中，具体包括下述步骤：

将所述主题词输入到预先训练的 Seq2Seq 模型中；
获取所述 Seq2Seq 模型响应所述主题词而输出的问题列表。

15、根据权利要求 11 所述的计算机设备，其特征在于，在所述将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据的步骤之后，还包括下述步骤：

按照预设的过滤规则对所述响应数据进行过滤，获取过滤数据，其中，所述过滤规则至少包含疑问语料数据过滤规则；

将所述过滤数据作为所述问题列表的应答数据。

16、一个或多个非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如下步骤：

获取待构建问答语料数据的主题词；

将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表；

将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据；

将所述响应数据作为所述问题列表的应答数据，所述应答数据与所述问题列表关联构成所述主题词的问答语料数据。

17、根据权利要求 16 所述的非易失性可读存储介质，其特征在于，在所述将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表的步骤中，具体包括下述步骤：

将所述主题词输入到第二网络爬虫模型中，获取所述第二网络爬虫模型响应所述主题词而输出的疑问候选数据；

按照预设的匹配规则对所述疑问候选数据进行匹配，获取疑问匹配数据，其中所述匹配规则至少包含疑问语料匹配规则；

将所述疑问匹配数据作为所述主题词的问题列表。

18、根据权利要求 17 所述的非易失性可读存储介质，其特征在于，在所述按照预设的匹配规则对所述疑问候选数据进行匹配的步骤中，所述匹配的步骤采用正则匹配算法获取疑问匹配数据。

19、根据权利要求 16 所述的非易失性可读存储介质，其特征在于，在所述将所述主题词输入到预先设定的问题生成模型中，获取所述问题生成模型响应所述主题词而输出的问题列表的步骤中，具体包括下述步骤：

将所述主题词输入到预先训练的 Seq2Seq 模型中；
获取所述 Seq2Seq 模型响应所述主题词而输出的问题列表。

20、根据权利要求 16 所述的非易失性可读存储介质，其特征在于，在所述将所述问题列表输入到预先设定的第一网络爬虫模型中，获取所述第一网络爬虫模型响应所述问题列表而输出的响应数据的步骤之后，还包括下述步骤：

按照预设的过滤规则对所述响应数据进行过滤，获取过滤数据，其中，所述过滤规则至少包含疑问语料数据过滤规则；
将所述过滤数据作为所述问题列表的应答数据。

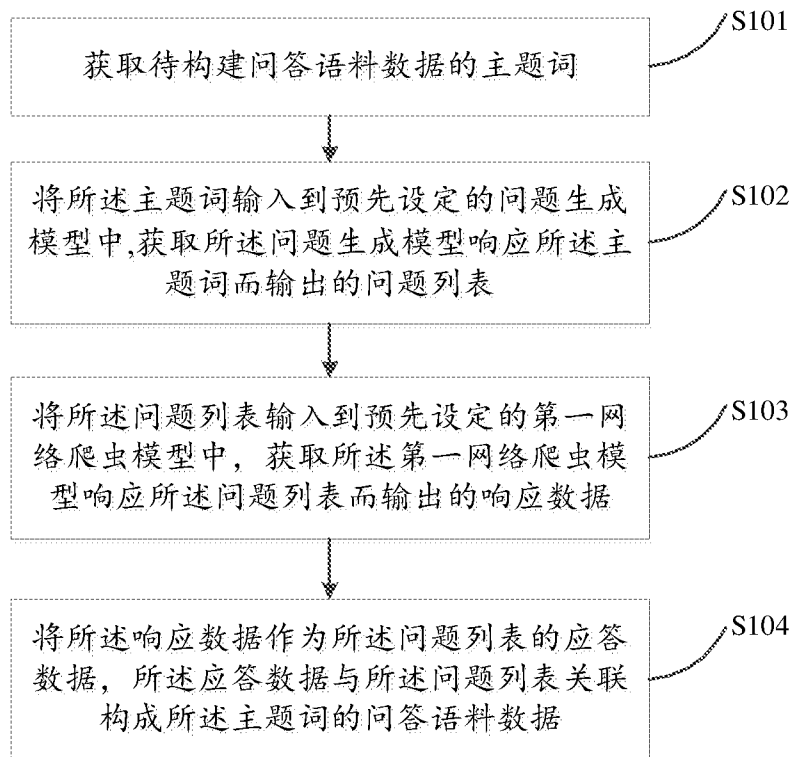


图 1

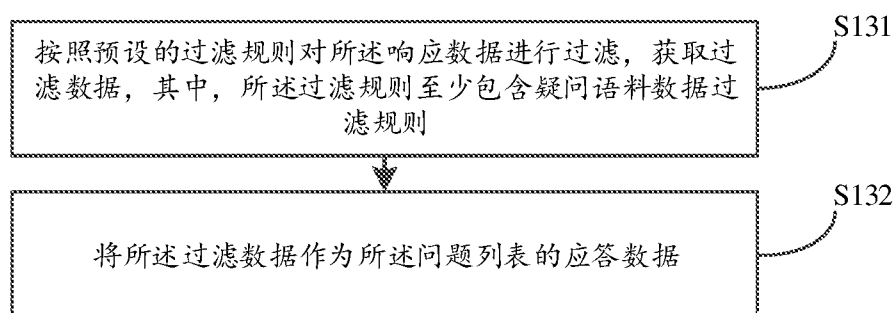
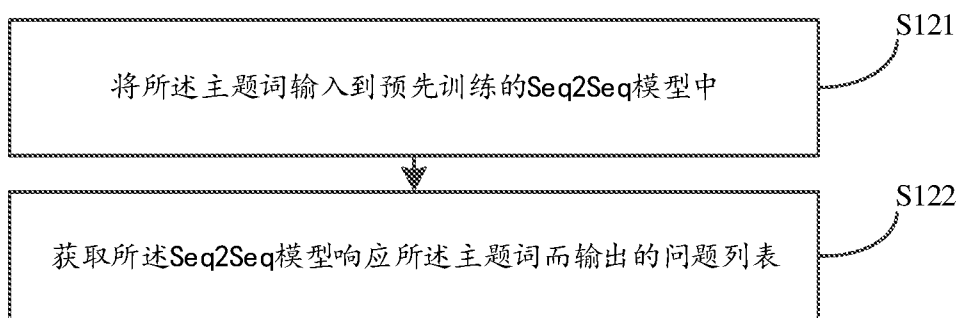
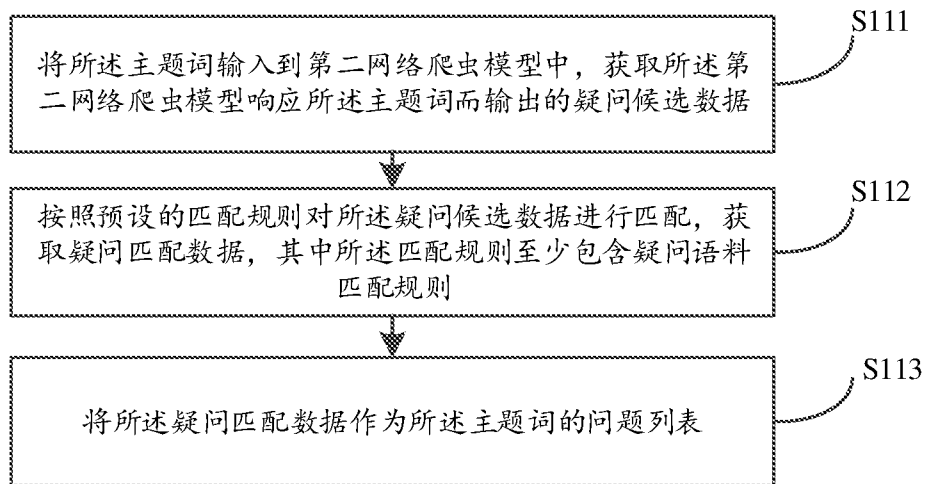


图 4

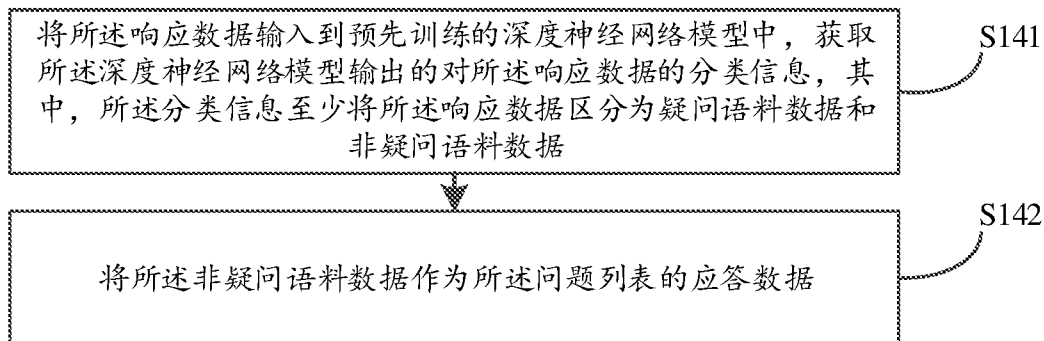


图 5

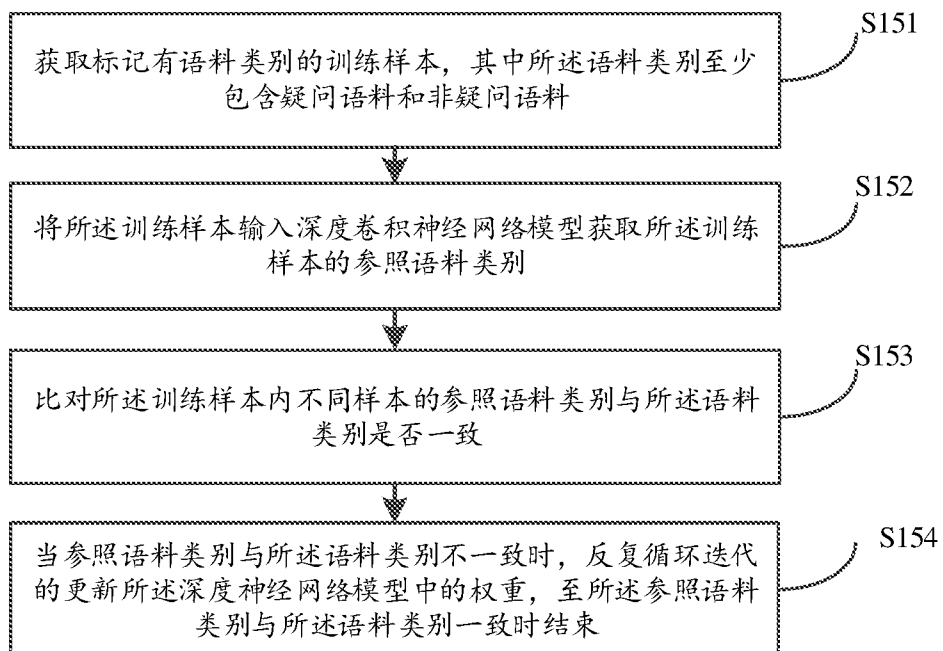


图 6

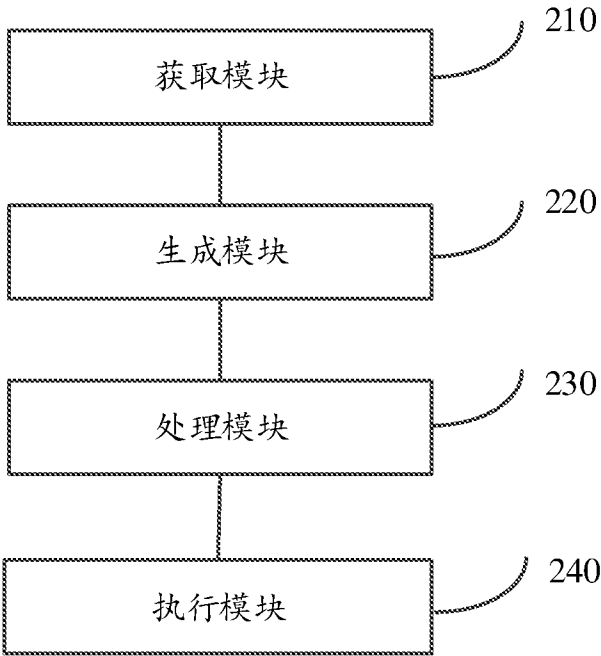


图 7

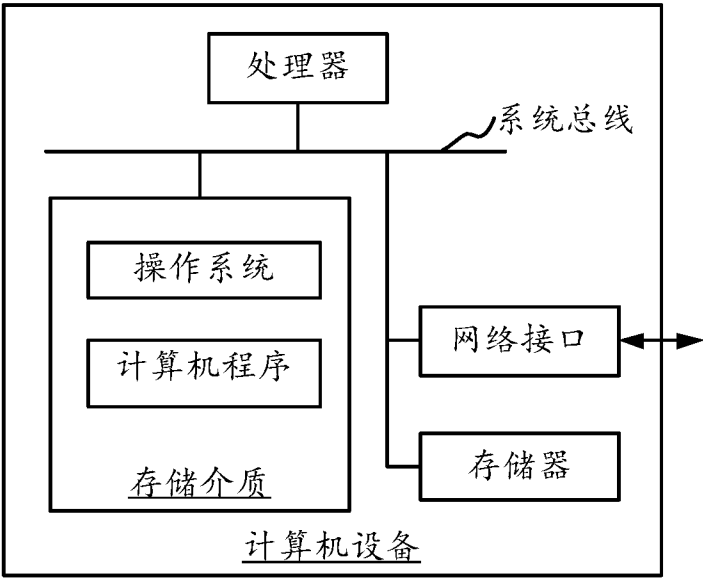


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/117698

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/332(2019.01)i; G06F 16/35(2019.01)i; G06Q 30/02(2012.01)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; G06Q

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, WPI, EPODOC, CNKI, IEEE: 爬虫, 语料, 构建, 主题, 关键词, 问题生成模型, 问答, 神经网络, 匹配, 训练, 过滤, crawler, corpus, build, topic, keyword, generation w model, question, answer, neural w network, match, train, filter

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109918486 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 21 June 2019 (2019-06-21) claims 1-10	1-20
Y	CN 108959559 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 07 December 2018 (2018-12-07) description, paragraphs [0078]-[0085] and [0164]	1-20
Y	CN 108345640 A (SHANGHAI UNIVERSITY) 31 July 2018 (2018-07-31) description, paragraphs [0006] and [0014]	1-20
A	US 2017308531 A1 (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 26 October 2017 (2017-10-26) entire document	1-20
A	CN 109190062 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 11 January 2019 (2019-01-11) 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

21 January 2020

Date of mailing of the international search report

02 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration
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/CN2019/117698

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109918486	A	21 June 2019	None			
CN	108959559	A	07 December 2018	None			
CN	108345640	A	31 July 2018	None			
US	2017308531	A1	26 October 2017	WO	2016112679	A1	21 July 2016
				CN	104573028	A	29 April 2015
				EP	3096246	A1	23 November 2016
				KR	20160125375	A	31 October 2016
				JP	2017511922	A	27 April 2017
CN	109190062	A	11 January 2019	None			

国际检索报告

国际申请号

PCT/CN2019/117698

A. 主题的分类 G06F 16/332(2019.01)i; G06F 16/35(2019.01)i; G06Q 30/02(2012.01)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; G06Q 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, WPI, EPDOC, CNKI, IEEE: 爬虫, 语料, 构建, 主题, 关键词, 问题生成模型, 问答, 神经网络, 匹配, 训练, 过滤, crawler, corpus, build, topic, keyword, generation w model, question, answer, neural network, match, train, filter		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 109918486 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 权利要求1-10	1-20
Y	CN 108959559 A (北京百度网讯科技有限公司) 2018年 12月 7日 (2018 - 12 - 07) 说明书第[0078]-[0085], [0164]段	1-20
Y	CN 108345640 A (上海大学) 2018年 7月 31日 (2018 - 07 - 31) 说明书第[0006], [0014]段	1-20
A	US 2017308531 A1 (BAIDU ONLINE NETWORK TECHNOLOGY BEIJING CO., LTD.) 2017年 10月 26日 (2017 - 10 - 26) 全文	1-20
A	CN 109190062 A (平安科技深圳有限公司) 2019年 1月 11日 (2019 - 01 - 11) 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 2020年 1月 21日		国际检索报告邮寄日期 2020年 3月 2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		授权官员 李玉坤 电话号码 86-(10)-53961358

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/117698

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	109918486	A	2019年 6月 21日	无			
CN	108959559	A	2018年 12月 7日	无			
CN	108345640	A	2018年 7月 31日	无			
US	2017308531	A1	2017年 10月 26日	WO	2016112679	A1	2016年 7月 21日
				CN	104573028	A	2015年 4月 29日
				EP	3096246	A1	2016年 11月 23日
				KR	20160125375	A	2016年 10月 31日
				JP	2017511922	A	2017年 4月 27日
CN	109190062	A	2019年 1月 11日	无			