(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 112926297 A
(43)申请公布日 2021.06.08

(21)申请号 202110222722.1

(22)申请日 2021.02.26

(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 徐海东 刘继辉 邢卓然

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 吕朝蕙

(51)Int.Cl.
G06F 40/194(2020.01)
G06F 40/211(2020.01)
G06F 40/295(2020.01)
G06K 9/62(2006.01)

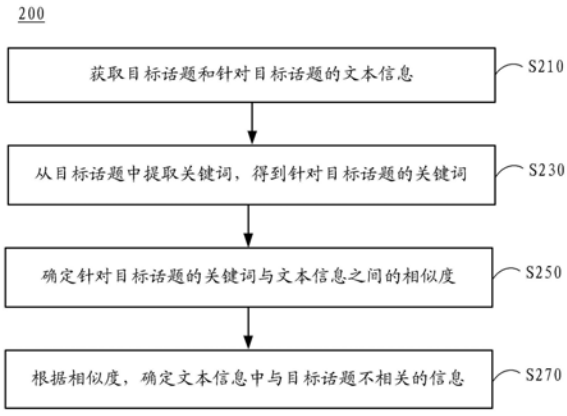
权利要求书2页 说明书10页 附图6页

(54)发明名称

处理信息的方法、装置、设备和存储介质

(57)摘要

本公开公开了一种处理信息的方法、装置、设备和存储介质，应用于计算机技术领域，具体应用于自然语言处理领域和深度学习领域。具体实现方案为：获取目标话题和针对目标话题的文本信息；从目标话题中提取关键词，得到针对目标话题的关键词；确定针对目标话题的关键词与文本信息之间的相似度；以及根据相似度，确定文本信息中与目标话题不相关的信息。



CN 112926297 A

1.一种处理信息的方法,包括:

获取目标话题和针对所述目标话题的文本信息;

从所述目标话题中提取关键词,得到针对所述目标话题的关键词;

确定针对所述目标话题的关键词与所述文本信息之间的相似度;以及

根据所述相似度,确定所述文本信息中与所述目标话题不相关的信息。

2.根据权利要求1所述的方法,其中,从所述目标话题中提取关键词,得到针对所述目标话题的关键词包括:

采用词法分析工具分析所述目标话题的话题名,得到话题名中的专有名词,作为针对所述目标话题的关键词;

在所述话题名中的专有名词个数小于第一预定值的情况下,采用词权重计算算法从所述目标话题的话题名中提取权重满足预定条件的词,作为针对所述目标话题的关键词。

3.根据权利要求1所述的方法,其中,从所述目标话题中提取关键词,得到针对所述目标话题的关键词还包括:

采用邻近算法确定针对所述权重满足预定条件的词的邻近词,作为针对所述目标话题的关键词。

4.根据权利要求1所述的方法,其中,确定针对所述目标话题的关键词与所述文本信息之间的相似度包括:

确定所述文本信息包括的语句个数;

确定与所述语句个数具有映射关系的相似度模型;以及

采用所述相似度模型确定针对所述目标话题的关键词与所述文本信息之间的相似度。

5.根据权利要求4所述的方法,其中,确定与所述语句个数具有映射关系的相似度模型包括:

在所述语句个数大于等于第二预定值的情况下,确定所述相似度模型为基于广义回归神经网络的相似度模型;

在所述语句个数小于所述第二预定值的情况下,确定所述相似度模型为基于潜在狄利克雷分布算法的相似度模型。

6.根据权利要求1所述的方法,其中,获取目标话题包括:

获取预定话题库中的多个话题;

对所述多个话题中每个话题的话题名进行词法分析,得到针对所述每个话题的目标词;以及

根据针对所述每个话题的目标词,确定所述多个话题中的目标话题。

7.根据权利要求6所述的方法,其中:

对所述多个话题中每个话题的话题名进行词法分析包括:采用词法分析工具分析所述每个话题的话题名,得到话题名包括的名词和动词,以作为针对所述每个话题的目标词;

确定所述多个话题中的目标话题包括:

在针对所述每个话题的目标词中包括专有名词的情况下,确定所述每个话题为目标话题;

在针对所述每个话题的目标词中不包括专有名词,且包括的名词个数或包括的动词个数大于等于第三预定值的情况下,确定所述每个话题为目标话题。

8.根据权利要求1所述的方法,其中,获取针对所述目标话题的文本信息包括:

获取针对所述目标话题的图文信息;

从所述图文信息中提取文本字段;以及

剔除所述文本字段中的特殊字段,得到针对所述目标话题的文本信息。

9.根据权利要求1所述的方法,其中,针对所述目标话题的关键词为多个,所述文本信息为多个,确定所述文本信息中与所述目标话题不相关的信息包括:

针对多个文本信息中的每个文本信息,在多个关键词中每个关键词与所述每个文本信息之间的相似度均小于相似度阈值的情况下,确定所述每个文本信息为与所述目标话题不相关的信息。

10.一种处理信息的装置,包括:

信息获取模块,用于获取目标话题和针对所述目标话题的文本信息;

关键词提取模块,用于从所述目标话题中提取关键词,得到针对所述目标话题的关键词;

相似度确定模块,用于确定针对所述目标话题的关键词与所述文本信息之间的相似度;以及

信息确定模块,用于根据所述相似度,确定所述文本信息中与所述目标话题不相关的信息。

11.一种电子设备,其特征在于,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1～9中任一项所述的方法。

12.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使所述计算机执行根据权利要求1～9中任一项所述的方法。

13.一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现根据权利要求1～9中任一项所述的方法。

处理信息的方法、装置、设备和存储介质

技术领域

[0001] 本公开涉及计算机技术领域，具体涉及自然语言处理领域和深度学习领域，更具体地涉及一种处理信息的方法、装置、设备和存储介质。

背景技术

[0002] 信息更新的实时性和热点覆盖度为影响用户体验的重要指标。话题作为一种新颖的热点展现形态，通过将带有话题标识的各种资源进行聚合排序，具备良好的实时性及资源多样性，能满足用户需求。

发明内容

[0003] 提供了一种处理信息的方法、装置、设备、介质和程序产品，以提高针对各话题的信息的准确性。

[0004] 根据第一方面，提供了一种处理信息的方法，包括：获取目标话题和针对目标话题的文本信息；从目标话题中提取关键词，得到针对目标话题的关键词；确定针对目标话题的关键词与文本信息之间的相似度；根据相似度，确定文本信息中与目标话题不相关的信息。

[0005] 根据第二方面，提供了一种处理信息的装置，包括：信息获取模块，用于获取目标话题和针对目标话题的文本信息；关键词提取模块，用于从目标话题中提取关键词，得到针对目标话题的关键词；相似度确定模块，用于确定针对目标话题的关键词与文本信息之间的相似度；以及信息确定模块，用于根据相似度，确定文本信息中与目标话题不相关的信息。

[0006] 根据第三方面，提供了一种电子设备，包括：至少一个处理器；以及与至少一个处理器通信连接的存储器；其中，存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能够执行本公开提供的处理信息的方法。

[0007] 根据第四方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，其中，计算机指令用于使计算机执行本公开提供的处理信息的方法。

[0008] 根据第五方面，提供了一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现本公开提供的处理信息的方法。

[0009] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0010] 附图用于更好地理解本方案，不构成对本公开的限定。其中：

[0011] 图1是根据本公开实施例的处理信息的方法、装置、设备和存储介质的应用场景示意图；

[0012] 图2示意性示出了根据本公开实施例的处理信息的方法的流程图；

[0013] 图3示意性示出了根据本公开实施例的获取目标话题的原理图；

[0014] 图4示意性示出了根据本公开实施例的获取属于目标话题的文本信息的原理图；

[0015] 图5示意性示出了根据本公开实施例的得到针对目标话题的关键词的原理图；

[0016] 图6示意性示出了根据本公开实施例的确定文本信息中与目标话题不相关的信息的原理图；

[0017] 图7是根据本公开实施例的处理信息的装置的结构框图；以及

[0018] 图8是用来实现本公开实施例的处理信息的方法的电子设备的框图。

具体实施方式

[0019] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0020] 本公开提供了一种处理信息的方法，该方法包括信息获取阶段、关键词提取阶段、相似度确定阶段和信息确定阶段。在信息获取阶段中，获取目标话题和针对目标话题的文本信息。在关键词提取阶段中，从目标话题中提取关键词，得到针对目标话题的关键词。在相似度确定阶段，确定针对目标话题的关键词与文本信息之间的相似度。在信息确定阶段，根据相似度，确定文本信息中与目标话题不相关的信息。

[0021] 以下将结合图1对本公开提供的方法和装置的应用场景进行描述。

[0022] 图1是根据本公开实施例的处理信息的方法、装置、设备、介质和程序产品的应用场景图。

[0023] 如图1所示，该实施例的应用场景100可以包括终端设备110和服务器120。该终端设备110和服务器120之间例如可以通过网络通信，网络例如可以包括有线或无线通信网络。

[0024] 终端设备110可以安装有各种客户端应用，例如购物类应用、网页浏览器应用、搜索类应用、网盘类应用、邮箱客户端、社交类应用等(仅为示例)。终端设备110可以为具有显示屏并且具有处理功能的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便捷计算机和台式计算机等等。

[0025] 示例性地，用户可以使用终端设备110通过网络与服务器120进行信息交互。服务器120例如可以为应用服务器，用于对终端设备110运行的客户端应用提供支持。在该实施例中，终端设备110可以响应于用户操作向服务器120发起话题信息校验请求。服务器120例如可以根据该话题信息校验请求对针对话题的信息进行校验，以校验针对话题的信息与话题之间的相关性。以便于剔除针对话题的信息中与话题不相关的信息，提高客户端应用维护的针对各话题的信息的质量。

[0026] 在一实施例中，服务器120例如可以是结合了区块链的服务器。或者，服务器120还可以为虚拟服务器或云服务器等。该服务器120例如可以将校验得到的与话题不相关的信息130反馈给终端设备110，以供用户进行进一步的人工复核。

[0027] 在一实施例中，该应用场景100还可以包括数据库140，该数据库140例如可以维护有客户端应用中针对所有话题的信息。服务器120在校验信息与话题之间的相关性时，例如可以根据话题标识从数据库140中获取针对话题的信息150。

[0028] 需要说明的是，本公开提供的处理信息的方法可以由服务器120执行。相应地，本公开提供的处理信息的装置可以设置在服务器120中。

[0029] 应该理解，图1中的终端设备、服务器和数据库仅仅是示意性的。根据实现需要，可以具有任意类型的终端设备、展示页面和服务器。

[0030] 以下将结合图1描述的应用场景，通过图2～图6对本公开提供的处理信息的方法进行详细描述。

[0031] 图2是根据本公开实施例的处理信息的方法的流程示意图。如图2所示，该实施例的处理信息的方法200包括操作S210、操作S230、操作S250和操作S270。

[0032] 在操作S210，获取目标话题和针对目标话题的文本信息。

[0033] 根据本公开的实施例，目标话题例如可以为响应于请求确定的话题。该请求例如可以由终端设备响应于用户操作而生成，该请求中可以包括话题名、话题标识等能够唯一标识话题的属性信息。

[0034] 根据本公开的实施例，可以从维护的所有话题中筛选出非泛化话题作为目标话题。这是由于针对泛化话题的信息很广泛，一般不要求信息与话题之间具有相关性。在筛选非泛化话题时，例如可以对话题名进行专有名词识别。在话题名包括专有名词时，则确定话题为非泛化话题，否则确定话题为泛化话题。专有名词例如可以包括人名、地名、公司名、机构名等能表示特定的或独一无二的人或物的名词。可以理解的是，非泛化话题一般指对特定事件的内容聚合而生成的话题，该特定事件例如可以为在社会中非普遍存在的事件。

[0035] 根据本公开的实施例，针对目标话题的文本信息例如可以为包括目标话题的话题名的信息。在获取针对目标话题的文本信息时，还可以从包括目标话题的话题名的信息中剔除话题名等字符，以避免该话题名对相关性判定的影响。示例性地，包括目标话题的话题名的信息例如为用户在社交类应用中发布的信息。例如，若该信息为XXXX#YYY#，YYY为话题名，则获取的针对目标话题的文本信息为XXXX。

[0036] 在操作S230，从目标话题中提取关键词，得到针对目标话题的关键词。

[0037] 根据本公开的实施例，可以对目标话题的话题名进行关键词提取，将提取的关键词作为针对目标话题的关键词。

[0038] 示例性地，在提取关键词时，可以先对话题名进行分词处理。将分词处理得到的多个词与关键词词库中的词进行比对，将多个词中属于关键词词库的词作为针对目标话题的关键词。关键词词库可以根据实际需求预先构建，例如可以由百科词条、专有名词、输入法细胞词库等构成，本公开对此不做限定。

[0039] 示例性地，可以采用词频-逆文本频率指数技术(Term Frequency-Inverse Document Frequency，TF-IDF)、基于序列标注模型的方法等来进行关键词的提取。

[0040] 在操作S250，确定针对目标话题的关键词与文本信息之间的相似度。

[0041] 根据本公开的实施例，可以采用word2vec方法分别将针对目标话题的关键词和文本信息转换为词向量。将转换得到的两个词向量之间的皮尔逊相关系数、斯皮尔曼相关系数或杰卡德相似系数等作为关键词与文本信息之间的相似度。或者，可以采用词频-逆文本频率指数技术、概率主题(Latent Dirichlet Allocation，LDA)模型(又称隐含狄利克雷分布模型)等来计算关键词与文本信息之间的相似度。可以理解的是，上述计算相似度的方法仅作为示例以利于理解本公开，本公开对此不做限定。

[0042] 在操作S270,根据相似度,确定文本信息中与目标话题不相关的信息。

[0043] 根据本公开的实施例,可以将文本信息中与针对目标话题的关键词的相似度低于相似度阈值的信息,作为与目标话题不相关的信息。

[0044] 根据本公开的实施例,若针对目标话题的关键词为多个,可以在多个关键词中每个关键词与文本信息之间的相似度均小于相似度阈值的情况下,确定该文本信息为与目标话题不相关的信息。或者,可以在多个关键词与文本信息之间的相似度的平均值小于相似度阈值的情况下,确定该文本信息为与目标话题不相关的信息。可以理解的是,相似度阈值可以根据实际需求进行设定,本公开对此不做限定。

[0045] 该实施例通过对话题进行关键词提取,并根据提取的关键词与文本信息之间的相似度来从文本信息中筛选与目标话题不相关的信息,可以实现对话题与话题内容之间相关性的自动判定。从而便于从针对话题的信息中挖掘与话题不相关的低质信息,提升话题下信息的质量,提高用户体验。

[0046] 图3示意性示出了根据本公开实施例的获取目标话题的原理图。

[0047] 根据本公开的实施例,在获取目标话题时,例如可以从话题库中获取非泛化话题作为目标话题,以此实现话题内容审核校验的准确性和有效性。如图3所示,可以先执行操作S311,从预定话题库中获取多个话题,随后从该多个话题中挑选非泛化话题。

[0048] 根据本公开的实施例,在从多个话题中挑选非泛化话题时,例如可以对多个话题中每个话题的话题名进行词法分析,得到针对每个话题的目标词。根据得到的目标词来确定该每个话题是否为非泛化话题,从而从多个话题中挑选出目标话题。

[0049] 示例性地,在对每个话题的话题名进行词法分析时,可以通过执行操作S312来实现,即对话题名进行分词并进行词性标注。该实施例可以采用词法分析工具来分析话题名,实现分词处理和词性标注处理,根据词性标注结果可以得到话题名包括的名词和动词,并将话题名包括的名词和动词作为针对目标话题的目标词。

[0050] 示例性地,词法分析工具例如可以为百度的词法分析工具(Lexical Analysis of Chinese 2.0,LAC2.0)、语言技术平台(Language Technology Platform,LTP)、清华大学词法分析器(THU Lexical Analyzer for Chinese,THULAC)等。该词法分析工具可以实现中文分词、词性标注和专有名词识别等功能。上述词法分析工具的类型仅作为示例以利于理解本公开,根据实际需求,可以采用任意类型的词法分析工具,本公开对此不做限定。

[0051] 在得到这对每个话题的目标词后,如图3所示,可以先执行操作S313,判断针对每个话题的目标词中是否有专有名词。若有专有名词,则执行操作S315,确定该每个话题为目标话题。若没有专有名词,则执行操作S314,判断针对每个话题的目标词中包括的名词个数是否大于等于第三预定值和针对每个话题的目标词中包括的动词个数是否大于等于第三预定值。若名词个数大于等于第三预定值或者动词个数大于等于第三预定值,则执行操作S315,确定该每个话题为目标话题。若名词个数和动词个数均小于第三预定值,且目标词不包括专有名词,则返回对从预定话题库中获取的另一话题进行分析,以确定该另一话题是否为目标话题,直至挑选到预定话题库中的所有目标话题。第三预定值例如可以根据实际需求进行设定,例如可以设定为3,本公开对此不做限定。

[0052] 图4示意性示出了根据本公开实施例的获取属于目标话题的文本信息的原理图。

[0053] 根据本公开的实施例,在得到目标话题后,可以获取针对目标话题的信息,并从获

取的信息中提取文本信息。

[0054] 示例性地，该实施例400可以以目标话题的话题标识410为索引，从信息库420中获取针对目标话题的信息。话题标识410可以为话题名等唯一指示目标话题的信息。信息库420中维护有针对所有话题的信息。

[0055] 示例性地，针对目标话题的信息例如可以包括文本、图片、视频等多种类型的信息。该实施例在获取针对目标话题的文本信息时，可以先从获取的多种类型的信息中过滤掉仅包括图片和/或视频的信息(即不包括文本的信息)，将剩余的信息作为针对目标话题的图文信息430。在过滤不包括文本的信息时，例如可以对信息的存储格式进行识别，将以图片格式和视频格式进行存储的信息过滤掉。

[0056] 根据本公开的实施例，在获取到针对目标话题的图文信息后，可以从该图文信息中提取文本字段。在提取文本字段时，如图4所示，可以先依据图文信息的标识查询下游服务440，以从该下游服务440中获取该图文信息的详情内容450。该下游服务例如可以为存储详情内容450的数据库，该数据库中的详情内容以图文信息的标识为索引。在获取到详情内容后，可以识别并提取详情内容中的文本字段460，该文本字段例如可以为content字段等。在提取到文本字段460后，例如可以剔除文本字段460中的特殊字段，将剔除了特殊字段的文本字段作为针对目标话题的文本信息470。

[0057] 示例性地，特殊字段例如可以包括除了汉字、二十六个英文字母外的特殊字符、介于一对特殊字符之间的字段等。该特殊字段例如可以根据实际需求进行设定，本公开对此不做限定。在剔除特殊字段时，例如可以先对文本字段进行特殊字符识别，得到文本字段包括的所有特殊字符。随后将所有特殊字符中位置相邻且相同的两个字符作为一对特殊字符。

[0058] 该实施例通过从图文信息中提取文本信息，并仅针对文本信息进行与目标话题相关的审核，可以提高审核的准确性。

[0059] 图5示意性示出了根据本公开实施例的得到针对目标话题的关键词的原理图。

[0060] 根据本公开的实施例，在确定针对目标话题的关键词时，可以从目标话题的话题名中提取。如图5所示，例如可以执行操作S531，从话题名中提取专有名词作为关键词，这是由于专有名词一般可以表示唯一的人或物，因此能够更好的表征该目标话题。

[0061] 示例性地，例如可以采用前文描述的词法分析工具来分析目标话题的话题名，从而得到话题名中的专有名词。

[0062] 在得到专有名词后，例如还可以执行操作S532，判断专有名词个数是否小于第一预定值。若小于第一预定值，例如还可以从话题名中提取除专有名词外的其他词，并将专有名词和该其他词作为关键词。若大于等于第一预定值，则无需再提取其他词，将提取的专有名词作为最终确定的关键词。第一预定值例如可以根据实际需求进行设定，本公开对此不做限定。

[0063] 根据本公开的实施例，在提取除专有名词外的其他词时，例如可以执行操作S533来实现。在操作S533，采用词权重计算算法从话题名中提取权重满足预定条件的词，作为针对目标话题的关键词。

[0064] 示例性地，词权重计算算法例如可以包括TF-IDF算法、文本排序算法(TextRank)、信息增益算法、条件随机场(Conditional Random Field,CRF)模型或百度自然语言处理云

平台的term重要性算子等。满足预定条件的词可以为计算得到的权重大于预定权重的词。或者，该满足预定条件的词可以为计算得到的权重较大的预定数量的词。预定数量例如可以根据专有名词的个数来确定，以使得该专有名词的个数与预定数量之和为固定值，该固定值例如可以为前述的第一预定值。

[0065] 该实施例通过在专有名词个数较少时提取权重满足预定条件的词，可以使得从话题名中提取的关键词能够更为充分的表达话题，便于提高最终确定的文本信息与目标话题之间的相似度的准确性。并因此可以提高最终确定的不相关信息的准确性，降低对相关信息的误伤。

[0066] 根据本公开的实施例，在从话题名中提取到关键词后，例如还可以对关键词进行扩充，并将扩充到的词作为针对目标话题的关键词。以此避免在确定相似度时，因表示相同含义的不同词的存在导致的相似度较低的情况，并因此可以进一步提高确定的关键词和文本信息之间相似度的准确性。

[0067] 示例性地，可以通过操作S534来实现关键词的扩充。在操作S534，采用邻近算法确定针对权重满足预定条件的词的邻近词，作为针对目标话题的关键词。邻近算法例如可以包括百度自然语言处理云平台提供的邻近词扩展算子、K最近邻分类算法(K-NearestNeighbor,KNN)等。

[0068] 示例性地，本公开例如可以维护有词库，在确定邻近词时，可以将权重满足预定条件的词作为中心，采用邻近算法确定词库中与该权重满足预定条件的词最接近的预定数量的词作为邻近词。预定数量可以根据实际需求进行设定，本公开对此不做限定。

[0069] 通过上述方法，可以在专有名词个数大于等于第一预定值时，将提取的专有名词作为针对目标话题的关键词，从而完成操作S535。在专有名词个数小于第一预定值时，可以将提取的专有名词和权重满足预定条件的词作为针对目标话题的关键词，或者将提取的专有名词、权重满足预定条件的词和针对权重满足预定条件的词的邻近词作为针对目标话题的关键词，从而完成操作S535。

[0070] 图6示意性示出了根据本公开实施例的确定文本信息中与目标话题不相关的信息的原理图。

[0071] 根据本公开的实施例，在确定文本信息中与目标话题不相关的信息时，例如可以先根据文本信息的长度来选择匹配的相似度模型。这是由于不同的相似度模型所擅长处理的语句长度有差别，通过根据文本信息的长度选择相似度模型，并采用选择的相似度模型来计算文本信息与关键词的相似度，可以提高确定的与目标话题不相关的信息的准确性。

[0072] 示例性地，如图6所示，在确定关键词与文本信息之间的相似度，并根据相似度确定文本信息是否与目标话题相关时，可以先通过操作S601对文本信息进行分句处理，并确定文本信息包括的语句个数。在对文本信息进行分句处理时，例如可以先识别文本信息中的标点符号，并将特定标点符号所在位置作为语句划分点。特定标点符号例如可以包括句子结束符号“。”、“!”、“?”和“.”等，还可以包括语句间隔符“,”、“、”和“;”等。可以理解的是，该对文本信息进行分句处理的方法仅作为示例以利于理解本公开，本公开对此不做限定。

[0073] 在得到语句个数后，可以确定与语句个数具有映射关系的相似度模型，以此来确定针对目标话题的关键词与文本信息之间的相似度。本公开实施例可以预先维护有语句个数与相似度模型的映射关系。该映射关系中相似度模型的类型和个数可以根据实际需求进

行设定，本公开对此不做限定。例如，可以根据语句个数设定三种相似度模型，以分别用于确定长语句与关键词之间的相似度、中长语句与关键词之间的相似度和短语句与关键词之间的相似度。

[0074] 示例性地，可以设定两种相似度模型，以分别用于确定较长语句和较短语句与关键词的相似度。在该实施例中，在得到语句个数后，可以先执行操作S602，判断语句个数是否小于第二预定值。若小于第二预定值，则确定文本信息为短语句，否则确定文本信息为长语句。第二预定值例如可以根据实际需求进行设定，例如，该第二预定值可以为4，本公开对此不做限定。

[0075] 在确定文本信息为长语句时，执行操作S603，确定相似度模型为基于广义回归神经网络(General Regression Neural Network，GRNN)的相似度模型。

[0076] 在确定文本信息为短语句时，执行操作S604，确定相似度模型为基于潜在狄利克雷分布算法(Latent Dirichlet Allocation，LDA)的相似度模型。

[0077] 在确定了相似度模型后，即可采用相似度模型来确定文本信息与针对目标话题的关键词中每个关键词的相似度。根据本公开的实施例，在针对目标话题的文本信息为多个时，针对多个文本信息中的每个文本信息，均可以通过操作S601～操作S604来确定相似度模型，并采用确定的相似度模型计算与针对目标话题的关键词中每个关键词的相似度。

[0078] 在采用基于广义回归神经网络的相似度模型来计算相似度后，可以执行操作S605，判断文本信息与每个关键词的相似度是否小于第一相似度阈值，若与针对目标话题的所有关键词的相似度均小于第一相似度阈值，则执行操作S607，确定该文本信息与目标话题不相关。若该文本信息与针对目标话题的某个关键词的相似度大于等于第一相似度阈值，则执行操作S608，确定该文本信息与目标话题相关。

[0079] 类似地，在采用基于潜在狄利克雷分布算法的相似度模型来计算相似度后，可以执行操作S606，判断文本信息与每个关键词的相似度是否小于第二相似度阈值，若与针对目标话题的所有关键词的相似度均小于第二相似度阈值，则执行操作S607，确定该文本信息与目标话题不相关。若该文本信息与针对目标话题的某个关键词的相似度大于等于第二相似度阈值，则执行操作S608，确定该文本信息与目标话题相关。

[0080] 可以理解的是，前述第一相似度阈值和第二相似度阈值可以根据实际需求进行设定。例如，在一实施例中，第一相似度阈值和第二相似度阈值可以为逼近于0的任意值，例如，第一相似度阈值为10^{-5}，第二相似度阈值为10^{-2}。本公开对该第一相似度阈值和第二相似度阈值的取值不做限定。

[0081] 基于前文描述的处理信息的方法，本公开还提供了一种处理信息的装置。以下将结合图7对该装置进行详细描述。

[0082] 图7是根据本公开实施例的处理信息的装置的结构框图。

[0083] 如图7所示，该实施例的处理信息的装置700可以包括信息获取模块710、关键词提取模块730、相似度确定模块750和信息确定模块770。

[0084] 信息获取模块710用于获取目标话题和针对目标话题的文本信息。在一实施例中，信息获取模块710例如可以用于执行前文描述的操作S210，在此不再赘述。

[0085] 关键词提取模块730用于从目标话题中提取关键词，得到针对目标话题的关键词。在一实施例中，关键词提取模块730例如可以用于执行前文描述的操作S230，在此不再赘

述。

[0086] 相似度确定模块750用于确定针对目标话题的关键词与文本信息之间的相似度。在一实施例中，相似度确定模块750例如可以用于执行前文描述的操作S250，在此不再赘述。

[0087] 信息确定模块770用于根据相似度，确定文本信息中与目标话题不相关的信息。在一实施例中，信息确定模块770例如可以用于执行前文描述的操作S270，在此不再赘述。

[0088] 根据本公开的实施例，上述关键词提取模块730例如可以用于采用词法分析工具分析目标话题的话题名，得到话题名中的专有名词，作为针对目标话题的关键词；并用于在目标话题中的专有名词个数小于第一预定值的情况下，采用词权重计算算法从目标话题的话题名中提取权重满足预定条件的词，作为针对目标话题的关键词。

[0089] 根据本公开的实施例，上述关键词提取模块730例如还可以用于采用邻近算法确定针对权重满足预定条件的词的邻近词，作为针对目标话题的关键词。

[0090] 根据本公开的实施例，上述相似度确定模块750例如可以包括语句个数确定子模块、模型确定子模块和相似度确定子模块。语句个数确定子模块用于确定文本信息包括的语句个数。模型确定子模块用于确定与语句个数具有映射关系的相似度模型。相似度确定子模块用于采用相似度模型确定针对目标话题的关键词与文本信息之间的相似度。

[0091] 根据本公开的实施例，上述模型确定子模块具体用于在语句个数小于第二预定值的情况下，确定相似度模型为基于广义回归神经网络的相似度模型；在语句个数大于等于第二预定值的情况下，确定相似度模型为基于潜在狄利克雷分布算法的相似度模型。

[0092] 根据本公开的实施例，上述信息获取模块710例如可以包括话题获取子模块、目标词确定子模块和目标话题确定子模块。话题获取子模块用于获取预定话题库中的多个话题。词法分析子模块用于对多个话题中每个话题的话题名进行词法分析，得到针对每个话题的目标词。目标话题确定子模块用于根据针对每个话题的目标词，确定多个话题中的目标话题。

[0093] 根据本公开的实施例，上述目标词确定子模块例如可以用于采用词法分析工具分析每个话题的话题名，得到话题名包括的名词和动词，以作为针对每个话题的目标词。上述目标话题确定子模块例如可以在针对每个话题的目标词中包括专有名词的情况下，确定每个话题为目标话题；在针对每个话题的目标词中不包括专有名词，且包括的名词个数或包括的动词个数大于等于第三预定值的情况下，确定每个话题为目标话题。

[0094] 根据本公开的实施例，上述信息获取模块710例如可以包括图文信息获取子模块、文本字段提取子模块和字符剔除子模块。图文信息获取子模块用于获取针对目标话题的图文信息。文本字段提取子模块用于从图文信息中提取文本字段。字符剔除子模块用于剔除文本字段中的特殊字符，得到针对目标话题的文本信息。

[0095] 根据本公开的实施例，针对目标话题的关键词为多个，文本信息为多个，上述信息确定模块770具体可以用于针对多个文本信息中的每个文本信息，在多个关键词中每个关键词与每个文本信息之间的相似度均小于相似度阈值的情况下，确定每个文本信息为与目标话题不相关的信息。

[0096] 根据本公开的实施例，本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0097] 图8示出了可以用来实现本公开实施例的处理信息的方法的电子设备800的示意性框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0098] 如图8所示，设备800包括计算单元801，其可以根据存储在只读存储器(ROM)802中的计算机程序或者从存储单元808加载到随机访问存储器(RAM)803中的计算机程序，来执行各种适当的动作和处理。在RAM 803中，还可存储设备800操作所需的各种程序和数据。计算单元801、ROM 802以及RAM 803通过总线804彼此相连。输入/输出(I/O)接口805也连接至总线804。

[0099] 设备800中的多个部件连接至I/O接口805，包括：输入单元806，例如键盘、鼠标等；输出单元807，例如各种类型的显示器、扬声器等；存储单元808，例如磁盘、光盘等；以及通信单元809，例如网卡、调制解调器、无线通信收发机等。通信单元809允许设备800通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0100] 计算单元801可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元801的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元801执行上文所描述的各个方法和处理，例如处理信息的方法。例如，在一些实施例中，处理信息的方法可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元808。在一些实施例中，计算机程序的部分或者全部可以经由ROM 802和/或通信单元809而被载入和/或安装到设备800上。当计算机程序加载到RAM 803并由计算单元801执行时，可以执行上文描述的处理信息的方法的一个或多个步骤。备选地，在其他实施例中，计算单元801可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行处理信息的方法。

[0101] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0102] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0103] 在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供

指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0104] 为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置(例如，CRT(阴极射线管)或者LCD(液晶显示器)监视器)；以及键盘和指向装置(例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈(例如，视觉反馈、听觉反馈、或者触觉反馈)；并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0105] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如，作为数据服务器)、或者包括中间件部件的计算系统(例如，应用服务器)、或者包括前端部件的计算系统(例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如，通信网络)来将系统的部件相互连接。通信网络的示例包括：局域网(LAN)、广域网(WAN)和互联网。

[0106] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0107] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本公开公开的技术方案所期望的结果，本文在此不进行限制。

[0108] 上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等，均应包含在本公开保护范围之内。

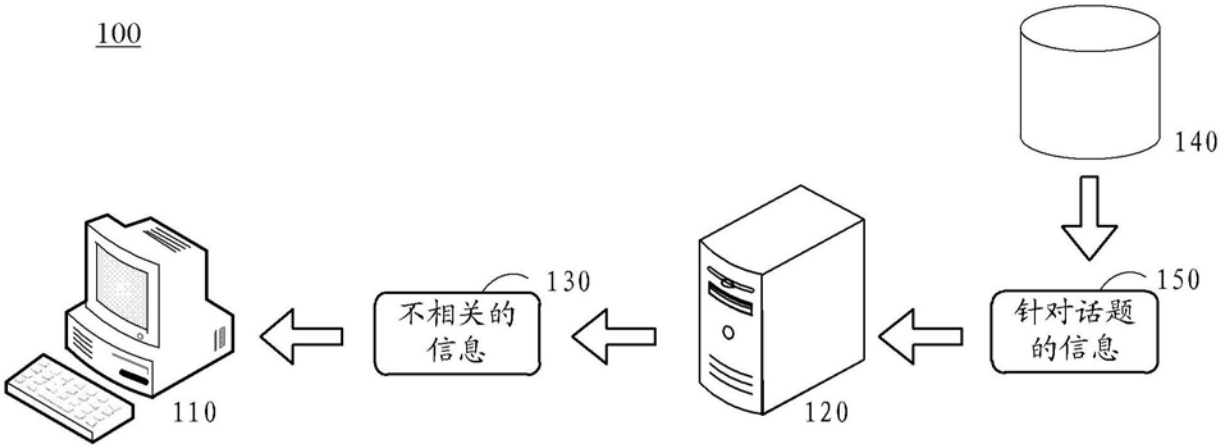

100
140
150
针对话题的信息
130
不相关的信息
110
120

图1

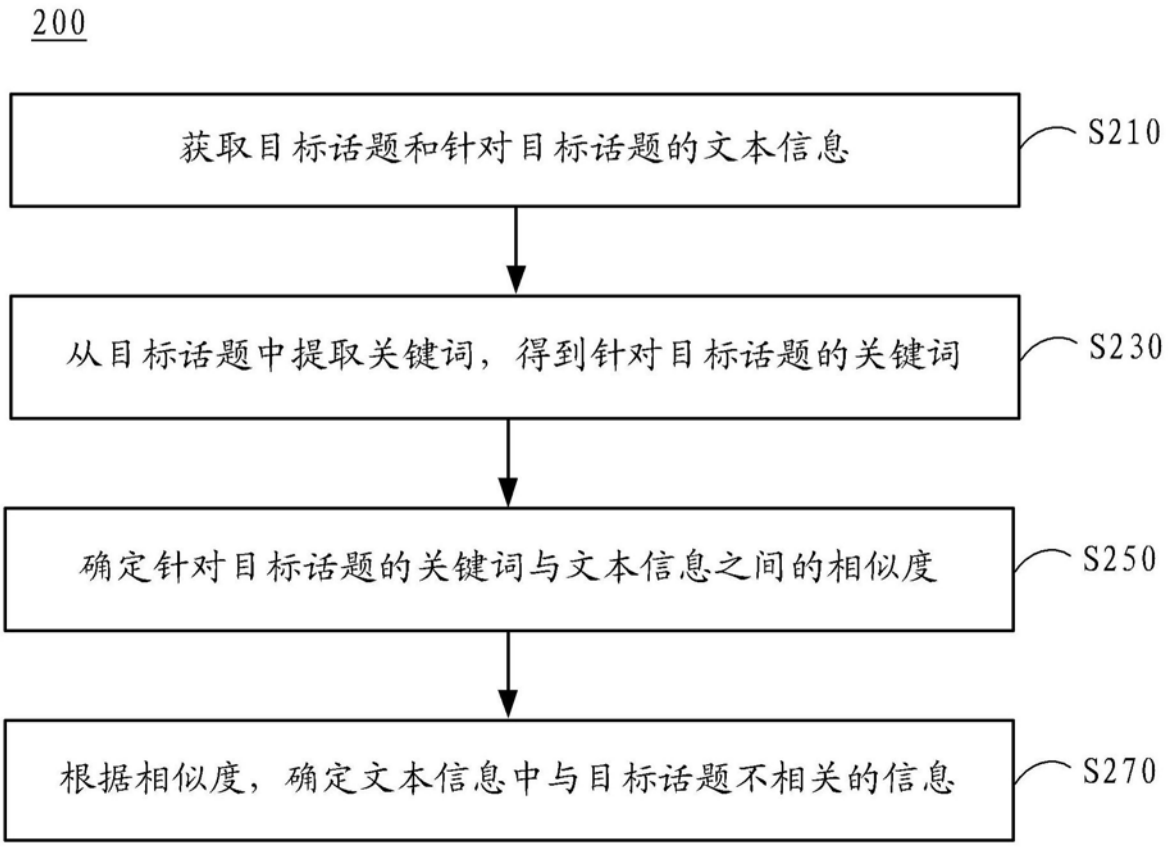

200
获取目标话题和针对目标话题的文本信息
S210
从目标话题中提取关键词，得到针对目标话题的关键词
S230
确定针对目标话题的关键词与文本信息之间的相似度
S250
根据相似度，确定文本信息中与目标话题不相关的信息
S270

图2

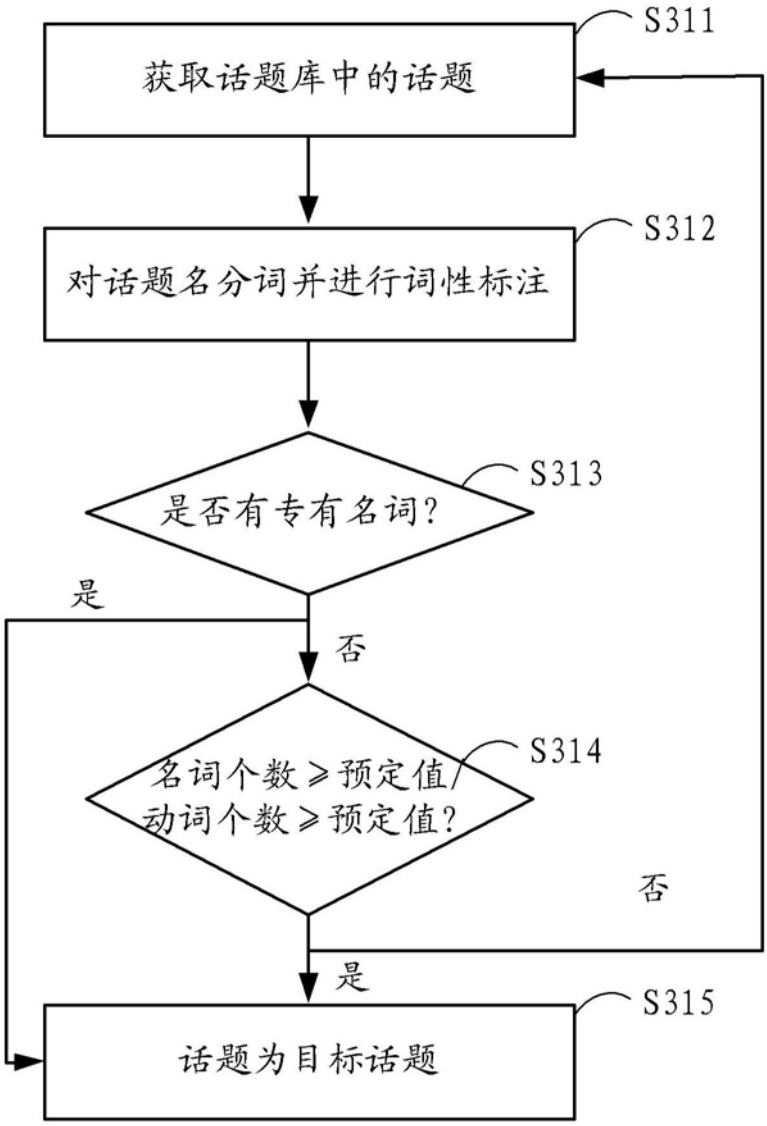

S311
获取话题库中的话题
S312
对话题名分词并进行词性标注
S313
是否有专有名词?
是
否
S314
名词个数≥预定值/
动词个数≥预定值?
否
是
S315
话题为目标话题

图3

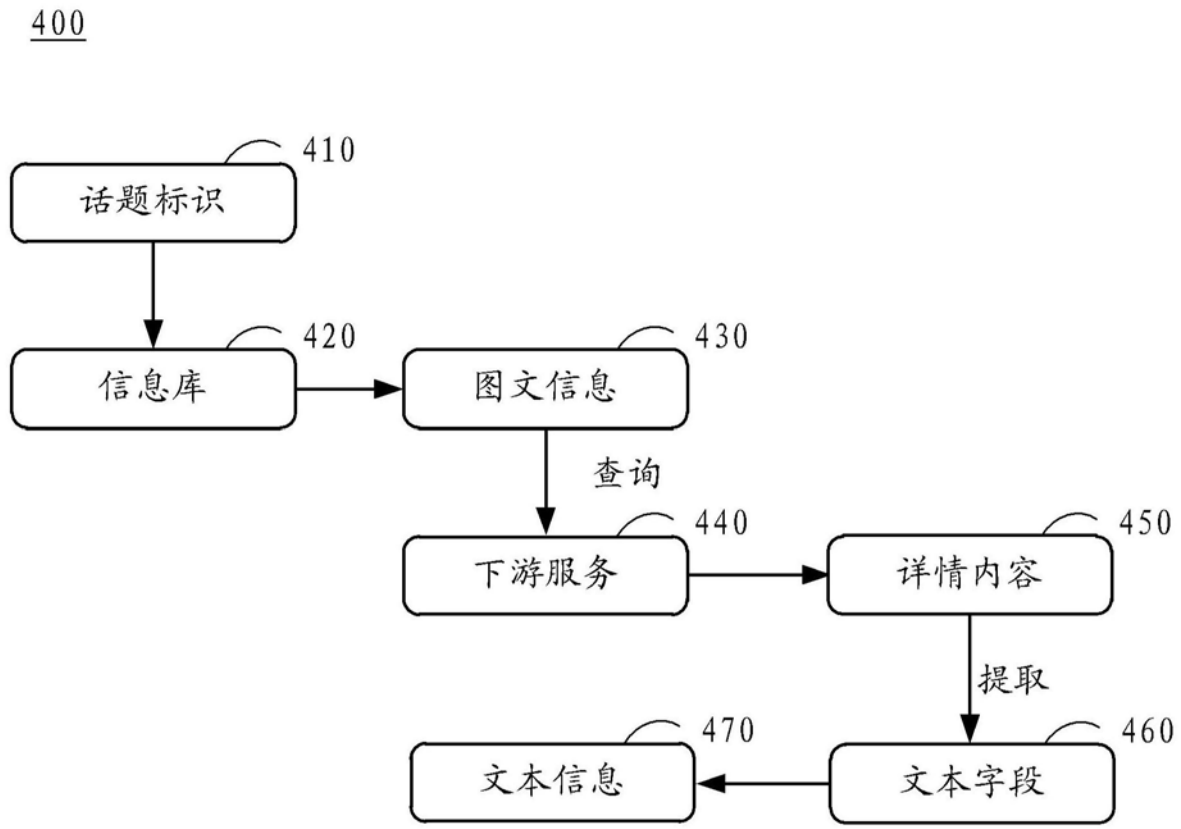

400
410
话题标识
420
信息库
430
图文信息
查询
440
下游服务
450
详情内容
提取
470
文本信息
460
文本字段

图4

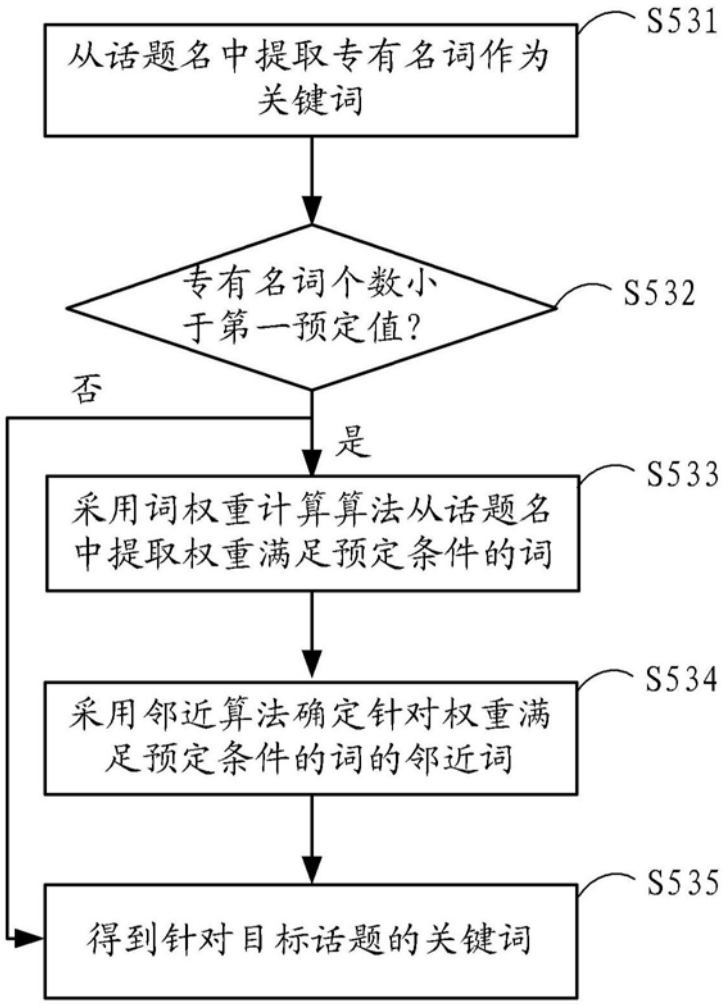

从话题名中提取专有名词作为关键词
S531
专有名词个数小于第一预定值?
S532
否
是
采用词权重计算算法从话题名中提取权重满足预定条件的词
S533
采用邻近算法确定针对权重满足预定条件的词的邻近词
S534
得到针对目标话题的关键词
S535

图5

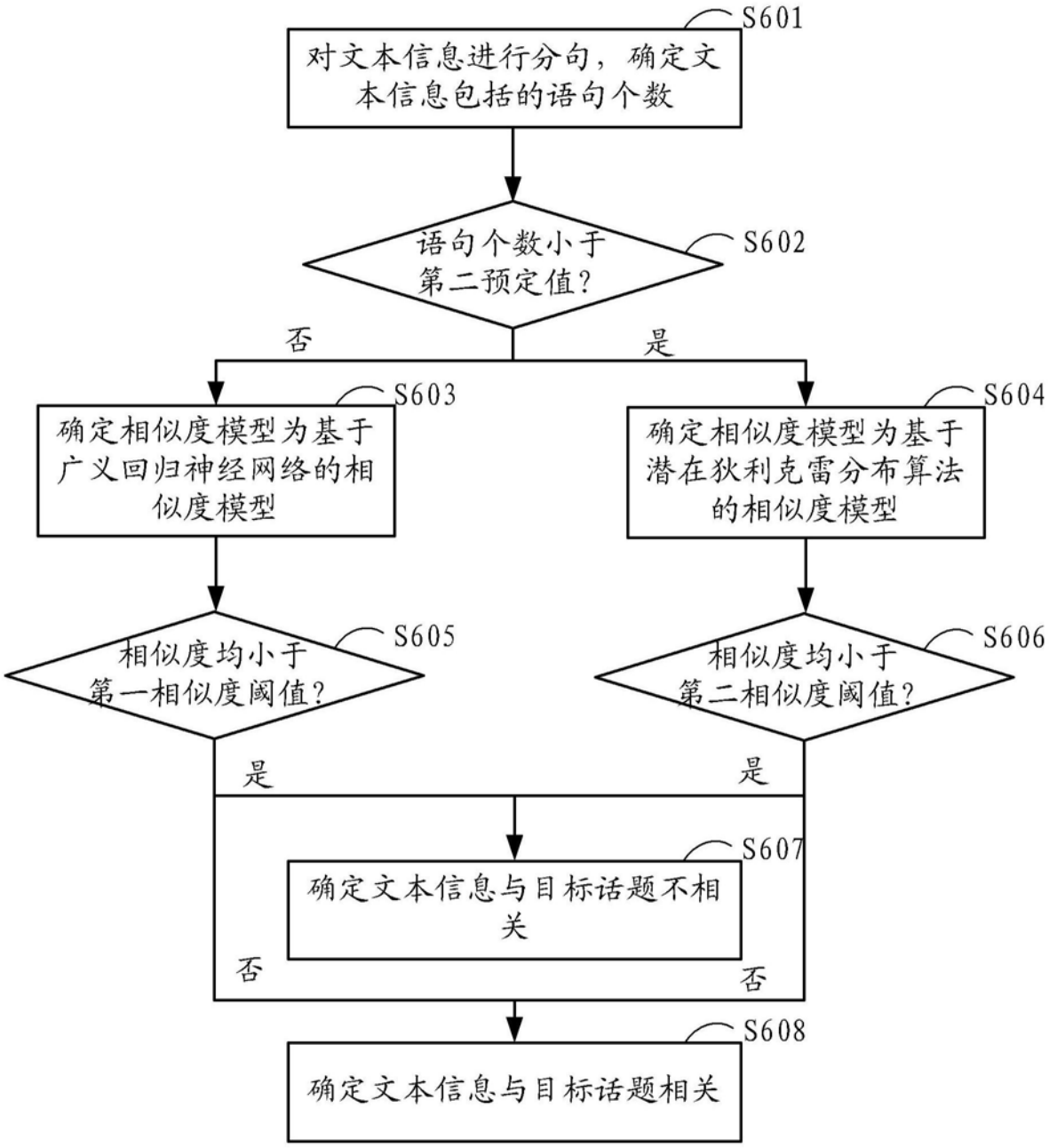

S601
对文本信息进行分句，确定文本信息包括的语句个数
语句个数小于第二预定值？
S602
否
是
S603
确定相似度模型为基于广义回归神经网络的相似度模型
S604
确定相似度模型为基于潜在狄利克雷分布算法的相似度模型
相似度均小于第一相似度阈值？
S605
相似度均小于第二相似度阈值？
S606
是
是
S607
确定文本信息与目标话题不相关
否
否
S608
确定文本信息与目标话题相关

图6

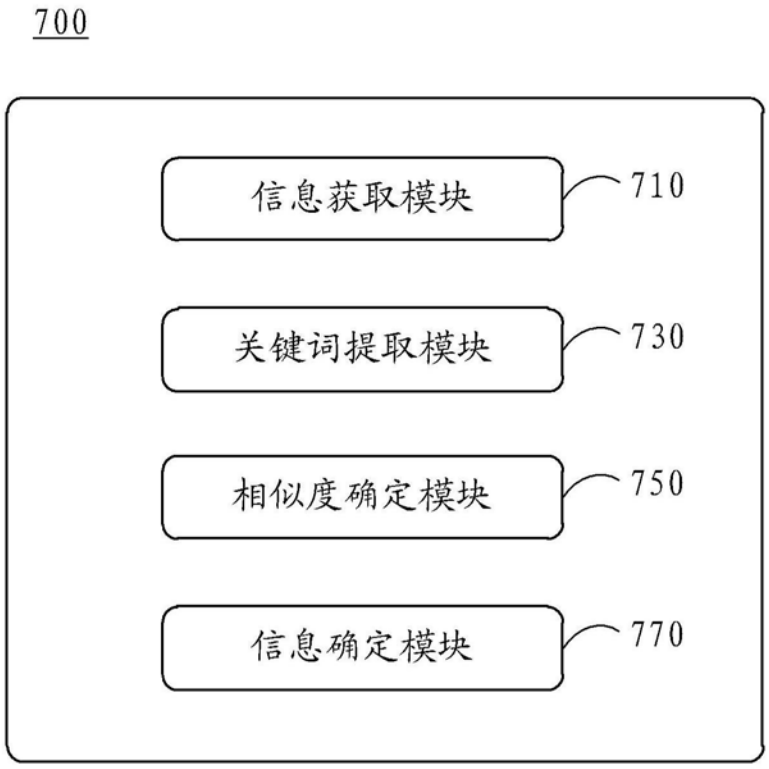

700
信息获取模块
710
关键词提取模块
730
相似度确定模块
750
信息确定模块
770

图7

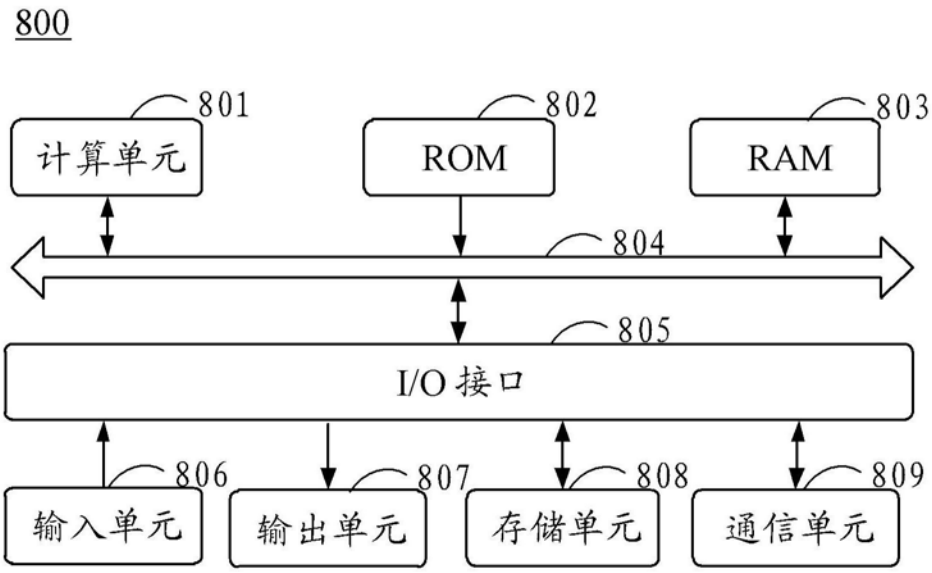

800
801
计算单元
802
ROM
803
RAM
804
805
I/O 接口
806
输入单元
807
输出单元
808
存储单元
809
通信单元

图8