(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110765239 B
(45)授权公告日 2023.03.28

(21)申请号 201911035607.2
(22)申请日 2019.10.29
(65)同一申请的已公布的文献号
申请公布号 CN 110765239 A
(43)申请公布日 2020.02.07
(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 王策 杜东
(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 邢惠童
(51)Int.Cl.
G06F 16/33(2019.01)
G06F 40/289(2020.01)
G06F 40/30(2020.01)

(56)对比文件
CN 101246499 A,2008.08.20
CN 103136258 A,2013.06.05
CN 103294664 A,2013.09.11
CN 103678670 A,2014.03.26
CN 104462551 A,2015.03.25
CN 104598583 A,2015.05.06
CN 104679738 A,2015.06.03
CN 107016999 A,2017.08.04
CN 107066589 A,2017.08.18
CN 107153658 A,2017.09.12
CN 107330022 A,2017.11.07
CN 107423444 A,2017.12.01
CN 108027814 A,2018.05.11
CN 109739367 A,2019.05.10
CN 110286775 A,2019.09.27
CN 110377916 A,2019.10.25
US 2014136523 A1,2014.05.15
刘荣,王奕凯.利用统计量和语言学规则提取多字词表达.《太原理工大学学报》.2011,第42卷(第42期),133-137.
李渝勤,孙丽华. 面向互联网舆情的热词分析技术.《中文信息学报》.2011,第25卷(第25期),48-53+59.
(续)

审查员 陈茜

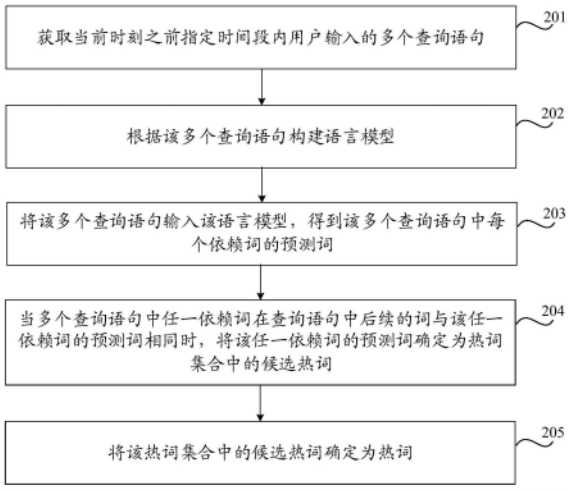
权利要求书2页 说明书11页 附图5页

(54)发明名称

热词识别方法、装置及存储介质

(57)摘要

本申请公开了一种热词识别方法、装置及存储介质,属于人工智能领域,涉及人工智能领域中的自然语言处理技术。所述方法通过将多个查询语句输入根据该多个查询语句构建的语言模型,可以得到该多个查询语句中每个依赖词的预测词,当该任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时,将该任一依赖词的预测词确定为热词集合中的候选热词,根据该候选热词确定热词。也即是根据用户的查询语句即可确定热词,热词的识别过程中,无需分析大量的文章,减少了热词识别过程中的计算量,解决了相关技术中热词识别的耗时较长的问题,提高了热词识别的效率。



CN 110765239 B

[转续页]

[接上页]

(56)对比文件

余一骄,尹燕飞,刘芹. 基于大规模语料库的高频汉字串互信息分布规律分析.《计算机科学》.2014,第41卷(第41期),276-282.

1.一种热词识别方法,其特征在于,所述方法包括:

获取当前时刻之前指定时间段内用户输入的多个查询语句;

根据所述多个查询语句构建N-Gram矩阵,所述N-Gram矩阵用于记录所述多个查询语句中任意两个词之间相邻的次数;

对所述N-Gram矩阵进行归一化;

对归一化后的N-Gram矩阵进行降维,得到语言模型,所述语言模型用于根据依赖词输出所述依赖词后续对应的预测词,所述依赖词包括至少一个词;

将所述多个查询语句输入所述语言模型,得到所述多个查询语句中每个依赖词的预测词;

当多个查询语句中任一依赖词在查询语句中后续的词与所述任一依赖词的预测词相同时,将所述任一依赖词的预测词确定为热词集合中的候选热词;

将所述热词集合中的候选热词确定为热词。

2.根据权利要求1所述的方法,其特征在于,所述对归一化后的N-Gram矩阵进行降维,得到所述语言模型,包括:

对归一化后的N-Gram矩阵中的每个值进行中心化;

获取进行所述中心化后的N-Gram矩阵的协方差矩阵;

对所述协方差矩阵进行特征值分解,得到多个特征值以及这多个特征值对应的特征向量;

将所述多个特征值中最大的n个特征值对应的特征向量构成所述语言模型,所述n为大于或等于1的整数。

3.根据权利要求1所述的方法,其特征在于,所述将所述热词集合中的候选热词确定为热词之前,所述方法还包括:

根据每个所述候选热词的边界熵以及作为查询语句的搜索次数确定每个所述候选热词的权重,所述边界熵与所述权重负相关,所述搜索次数与所述权重正相关;

去除所述热词集合中所述权重小于指定阈值的候选热词。

4.根据权利要求3所述的方法,其特征在于,所述根据每个所述候选热词的边界熵以及作为查询语句的搜索次数确定每个所述候选热词的权重,包括:

将所述多个候选热词划分为多个簇,每个簇包括至少一个所述候选热词,且同一个簇的候选热词所属的查询语句对应于同一个文档;

在每个所述簇中,根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重。

5.根据权利要求1所述的方法,其特征在于,所述将所述热词集合中的候选热词确定为热词之前,所述方法还包括:

去除所述热词集合中满足指定条件的候选热词,所述指定条件包括以停用词开头或结尾、以空格为开头或结尾,或者字符数大于指定值。

6.一种热词识别装置,其特征在于,所述热词识别装置包括:

获取模块,用于获取当前时刻之前指定时间段内用户输入的多个查询语句;

模型构建模块,用于根据所述多个查询语句构建N-Gram矩阵,所述N-Gram矩阵用于记录所述多个查询语句中任意两个词之间相邻的次数;

对所述N-Gram矩阵进行归一化；

对归一化后的N-Gram矩阵进行降维，得到语言模型，所述语言模型用于根据依赖词输出所述依赖词后续对应的预测词，所述依赖词包括至少一个词；

输入模块，用于将所述多个查询语句输入所述语言模型，得到所述多个查询语句中每个依赖词的预测词；

集合建立模块，用于当多个查询语句中任一依赖词在查询语句中后续的词与所述任一依赖词的预测词相同时，将所述任一依赖词的预测词确定为热词集合中的候选热词；

热词确定模块，用于将所述热词集合中的候选热词确定为热词。

7.一种服务器，其特征在于，所述服务器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权利要求1至5任一所述的热词识别方法。

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权利要求1至5任一所述的热词识别方法。

热词识别方法、装置及存储介质

技术领域

[0001] 本申请涉及人工智能领域，特别涉及一种热词识别方法、装置及存储介质。

背景技术

[0002] 热词即为热门词汇，其能够反应当前一段时间人们普遍关注的问题和事务。因此，如何准确快速的识别热词是目前人工智能领域中自然语言处理(Nature Language processing,NLP)技术的一个重要发展方向。

[0003] 相关技术提供了一种热词识别方法，该方法首先获取当前一段时间发布的多篇文章，之后对这多篇文章进行文本分析，确定这多篇文章中每个词的出现次数，再将其中出现次数最多的词确定为热词。

[0004] 但是，上述方法对多篇文章进行文本分析的计算量较大，进而导致热词识别的耗时较长。

发明内容

[0005] 本申请实施例提供了一种热词识别方法、装置及存储介质，能够解决相关技术中热词识别的耗时较长的问题。所述技术方案如下：

[0006] 根据本申请的第一方面，提供了一种热词识别方法，所述方法包括：

[0007] 获取当前时刻之前指定时间段内用户输入的多个查询语句；

[0008] 根据所述多个查询语句构建语言模型，所述语言模型用于根据依赖词输出所述依赖词后续对应的预测词，所述依赖词包括至少一个词；

[0009] 将所述多个查询语句输入所述语言模型，得到所述多个查询语句中每个依赖词的预测词；

[0010] 当多个查询语句中任一依赖词在查询语句中后续的词与所述任一依赖词的预测词相同时，将所述任一依赖词的预测词确定为热词集合中的候选热词；

[0011] 将所述热词集合中的候选热词确定为热词。

[0012] 另一方面，提供了一种热词识别装置，所述热词识别装置包括：

[0013] 获取模块，用于获取当前时刻之前指定时间段内用户输入的多个查询语句；

[0014] 模型构建模块，用于根据所述多个查询语句构建语言模型，所述语言模型用于根据依赖词输出所述依赖词后续对应的预测词，所述依赖词包括至少一个词；

[0015] 输入模块，用于将所述多个查询语句输入所述语言模型，得到所述多个查询语句中每个依赖词的预测词；

[0016] 集合建立模块，用于当多个查询语句中任一依赖词在查询语句中后续的词与所述任一依赖词的预测词相同时，将所述任一依赖词的预测词确定为热词集合中的候选热词；

[0017] 热词确定模块，用于将所述热词集合中的候选热词确定为热词。

[0018] 另一方面，提供了一种服务器，所述服务器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前述一方面所述的热词识别方法。

[0019] 另一方面，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、所述代码集或指令集由所述处理器加载并执行以实现如前述一方面所述的热词识别方法。

[0020] 本申请实施例提供的技术方案带来的有益效果至少包括：

[0021] 通过将多个查询语句输入根据该多个查询语句构建的语言模型，可以得到该多个查询语句中每个依赖词的预测词，当该任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词，根据该候选热词确定热词。也即是根据用户的查询语句即可确定热词，热词的识别过程中，无需分析大量的文章，减少了热词识别过程中的计算量，解决了相关技术中热词识别的耗时较长的问题，提高了热词识别的效率。

附图说明

[0022] 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。

[0023] 图1是本申请实施例提供的一种热词识别方法所涉及的实施环境示意图；

[0024] 图2是本申请实施例提供的一种热词识别方法的流程图；

[0025] 图3是本申请实施例提供的另一种热词识别方法的流程图；

[0026] 图4是本申请实施例提供的一种用根据多个查询语句构建语言模型的流程图；

[0027] 图5是本申请实施例提供的一种根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重的流程图；

[0028] 图6是本申请实施例提供的一种热词识别装置的框图；

[0029] 图7是本申请实施例提供的一种服务器的框图。

[0030] 通过上述附图，已示出本申请明确的实施例，后文中将有更详细的描述。这些附图和文字描述并不是为了通过任何方式限制本申请构思的范围，而是通过参考特定实施例为本领域技术人员说明本申请的概念。

具体实施方式

[0031] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0032] 人工智能(Artificial Intelligence,AI)技术是利用数字计算机或者利用由数字计算机控制的机器来模拟、延伸和扩展人的智能。人工智能技术可以感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。人工智能是计算机科学的一个综合技术，其用于研究人类智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，使该智能机器具有感知、推理与决策的功能。人工智能技术包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习等研究方向。

[0033] 随着人工智能技术研究和进步，人工智能技术已在多个领域展开研究和应用，其中，自然语言处理技术作为其重要的分支，可以用于实现人与计算机之间用自然语言进行有效通信，自然语言即人们日常使用的语言，自然语言处理技术可以包括文本处理、语义理解、机器翻译、机器人问答以及知识图谱等技术。本申请实施例提供的方案涉及人工智能的自然语言处理技术，用于实现一种热词的识别方法。在对本申请实施例进行说明之前，首先对现有技术进行说明：

[0034] 随着互联网的高速发展，用户可以在网络上随时随地针对某些问题或者事务发表自己的看法。对于当前一段时间用户普遍关注的问题和事务，通常会引发用户之间频繁的互动以及讨论，在互动以及讨论的过程中，会涌现一些针对该关注的问题和事务的热门词汇，即热词。例如在春节期间，“春节联欢晚会”为大多数用户所关注的事件，相应的，“春节联欢晚会”或者“春晚”为春节期间出现的热词。伴随着热词通常会出现一些新词，新词可以是新的词语、词组或者旧词新用等。由于网络上的用户趋于年轻化，因此，新词被频繁地制造并流行。在本申请实施例中，可以将新词等同于热词。

[0035] 识别当前网络中流行的热词，可以有助于实现文本语义分析以进行有效地数据挖掘等人工智能任务。但是，由于用户在网络上会活跃于不同的社交圈内，例如，用户可以活跃于与运动有关的社交圈、与电影有关的社交圈或者与音乐有关的社交圈等，而不同的社交圈具有不同的话题，如此使得不同的社交圈内热词的分布也不同。因此，基于特定社交圈(或特定领域)的语料库形成专门的热词，才能实现更精确的文本语义分析。

[0036] 相关技术提供的热词识别方法中，由于识别热词的过程是基于大规模的语料库的，需要对多篇文章进行文本分析，将该多篇文章中出现频率高的词作为热词，因此，导致热词识别的耗时较长。而本申请实施例提供了一种热词识别方法，可以有效缩短热词识别的耗时。需要提前说明的是，在本申请实施例中，热词可以是中文或者英文等任一语言的词汇，本申请实施例以热词为英文进行举例说明。

[0037] 图1是本申请实施例提供的一种热词识别方法所涉及的实施环境示意图。如图1所示，该实施环境可以包括：至少一个终端110以及服务器120，该至少一个终端110与服务器120可以通过有线或者无线网络进行连接。服务器120可以为一个服务器或服务器集群。终端110可以为计算机、笔记本电脑或者智能手机等，图1以该终端110为电脑为例进行说明。服务器120可以用于执行本申请实施例所提供的热词识别方法。至少一个用户可以通过该多个终端110输入多个查询语句以供服务器120获取。图1中是以两个终端110为例进行说明。

[0038] 图2是本申请实施例示出的一种热词识别方法的流程图，该热词识别方法可以由热词识别装置执行，该热词识别装置可以以硬件或者软件的形式设置于图1所示的服务器120中，该热词识别方法可以包括：

[0039] 步骤201、获取当前时刻之前指定时间段内用户输入的多个查询语句。

[0040] 步骤202、根据该多个查询语句构建语言模型。

[0041] 该语言模型用于根据依赖词输出该依赖词后续对应的预测词，该依赖词包括至少一个词。

[0042] 步骤203、将该多个查询语句输入该语言模型，得到该多个查询语句中每个依赖词的预测词。

[0043] 步骤204、当多个查询语句中任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词。

[0044] 步骤205、将该热词集合中的候选热词确定为热词。

[0045] 综上所述，本申请实施例提供的热词识别方法，通过将多个查询语句输入根据该多个查询语句构建的语言模型，可以得到该多个查询语句中每个依赖词的预测词，当该任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词，根据该候选热词确定热词。也即是根据用户的查询语句即可确定热词，热词的识别过程中，无需分析大量的文章，减少了热词识别过程中的计算量，解决了相关技术中热词识别的耗时较长的问题，提高了热词识别的效率。

[0046] 相关技术中还提供了一种热词识别的方法，该方法首先对文本进行分词，然后将未能成功匹配(即未能成功分词)的剩余片段确定为热词。对文本进行分词指的是将由汉字序列组成的文本切分成一个个单独的词。该方法中，分词的准确性依赖于语料库的完整性，但是语料库难以做到及时更新以适应网络的迅猛发展。当语料库更新不及时则会导致该语料库中缺失部分热词，导致对文本进行分词的可信度较差，进而导致热词识别的准确率较低。而本申请实施例提供的热词识别方法，由于语言模型是基于当前时刻之前指定时间段内用户输入的多个查询语句构建的，保证了该语言模型的时效性，使得基于该语言模型识别的热词准确率较高。

[0047] 进一步的，请参考图3，其示出了本申请实施例提供的另一种热词识别方法的流程图，该方法可以由热词识别装置执行，该热词识别装置可以以硬件或者软件的形式设置于图1所示的服务器120中，该热词识别方法可以包括：

[0048] 步骤301、获取当前时刻之前指定时间段内用户输入的多个查询语句。

[0049] 与热词识别装置连接的终端可以提供用于输入该多个查询语句的入口，该入口可以以搜索框或输入框的形式呈现。该入口可以为搜索引擎的搜索入口，热词识别装置通过该搜索入口获取用户输入的多个查询语句。

[0050] 该多个查询语句可以为由一个用户输入的，或者由多个用户输入的。该查询语句可以包括至少一个词。该指定时间段可以包括当前时刻或者不包括当前时刻。当该指定时间段包括当前时刻时，也即是，该指定时间段为历史时刻至当前时刻对应的时间段，则获取的用户输入的多个查询语句的时效性更强，使得最终识别的热词的准确性更高。

[0051] 步骤302、根据该多个查询语句构建语言模型。

[0052] 该语言模型可以用于根据依赖词输出该依赖词后续对应的预测词，依赖词包括至少一个词。语言模型可以为N-Gram模型，该N-Gram模型可以根据前N-1个词(item)来预测第N个词。N-Gram模型可以有多种，例如，如果一个词的出现依赖于该词前面出现的一个词，则该N-Gram模型为bi-gram模型；如果一个词的出现依赖于该词前面出现的两个词，则该N-Gram模型为tri-gram模型。该N-Gram模型可以通过N-Gram矩阵进行描述。

[0053] 则如图4所示，步骤302中根据多个查询语句构建语言模型的过程可以包括：

[0054] 步骤3021、根据该多个查询语句构建N-Gram矩阵。

[0055] 该N-Gram矩阵用于记录该多个查询语句中任意两个词之间相邻的次数。

[0056] 表1示意性地示出了用户在当前时刻之前指定时间段内输入的多个查询语句中每个词出现的次数：

[0057] 表1

[0058]

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	146	158	1093	341	278

[0059] 从表1中可以看出，在用户在当前时刻之前指定时间段内输入的多个查询语句(例如用户输入了3000个查询语句)中，词“i”出现的次数为2533，词“want”出现的次数为927，词“to”出现的次数为2417，词“eat”出现的次数为146，词“chinese”出现的次数为158，词“food”出现的次数为1093，词“lunch”出现的次数为341，词“spend”出现的次数为278。当然，表1中的词仅为示意性举例。

[0060] 表2示意性示出了一种根据多个查询语句构建的bi-gram矩阵，也即是当N为2时的N-Gram矩阵。

[0061] 表2

[0062]

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

[0063] 表2示出的bi-gram矩阵中，示意性示出了“i”、“want”、“to”、“eat”、“chinese”、“food”、“lunch”以及“spend”这8个词在多个查询语句中，两两相邻出现的次数。其中，最左一列的词为两个词中位于前面位置的词，最上一行的词为两个词中位于后面位置的词，其余位置的数值均为两个词相邻出现的次数。对于任意一个位置的数值，表示其同一行的词后续出现其同一列的词的次数，例如，左起第三列第二行的位置的数值827，表示其同一行的词“i”后续出现其同一列的词“want”的次数。

[0064] 表2仅示出了一种N-Gram矩阵的形式，但N-Gram矩阵还可以为其它形式，本申请实施例不进行限制。

[0065] 步骤3022、对N-Gram矩阵进行归一化。

[0066] 为了提高运算速度，可以在构建N-Gram矩阵之后对该N-Gram矩阵进行归一化(也称标准化)处理。

[0067] 表3示意性地示出了在对表2示出的Bi-gram矩阵进行归一化后的频率分布表格。

[0068] 表3

[0069]

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0

chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

[0070] 该表3为在上述表1以及表2的基础上确定的。表3中最左一列的词为两个词中位于前面位置的词，最上一行的词为两个词中位于后面位置的词，其余位置的数值均为两个词相邻出现的频率，该频率通过将两个词相邻出现的次数除以两个词中位于前面位置的词在用户在当前时刻之前指定时间段内输入的多个查询语句中出现的次数来确定。例如，左起第四列第三行的位置的数值0.66，为通过将数值608除以数值927近似确定，其中，数值608表示其同一行的词“want”后续出现其同一列的词“to”的次数，数值927表示用户在当前时刻之前指定时间段内输入的多个查询语句中“want”出现的次数。

[0071] 表3仅示出了对一种N-Gram矩阵的形式进行归一化处理后的形式，当N-Gram矩阵为其它形式时，归一化的该其他形式的N-Gram矩阵的方法可以参考上述，本申请实施例不进行限制。

[0072] 步骤3023、对归一化后的N-Gram矩阵进行降维，得到语言模型。

[0073] 对矩阵降维可以便于对矩阵中数据的计算以及可视化，能够将有效信息进行提取和综合而将无效信息进行摒弃。因此，在本申请实施例中，通过对归一化后的N-Gram矩阵进行降维处理，使得得到的语言模型可以有效进行热词的识别。

[0074] 对归一化后的N-Gram矩阵进行降维的方式可以有多种，例如，可以采用主成分分析(Principal Component Analysis，PCA)或者奇异值分解(Singular Value Decomposition，SCD)等矩阵降维方式。本申请实施例以采用PCA对归一化后的N-Gram矩阵进行降维的方式进行说明。则步骤3023可以包括：

[0075] 步骤S1、对归一化后的N-Gram矩阵中的每个值进行中心化。

[0076] 对归一化后的N-Gram矩阵中的每个值进行中心化的过程可以：确定该归一化后的N-Gram矩阵中的所有值的均值，将归一化后的N-Gram矩阵中的每个值均减去该均值则得到中心化后的每个值。

[0077] 步骤S2、获取进行中心化后的N-Gram矩阵的协方差矩阵。

[0078] 步骤S3、对协方差矩阵进行特征值分解，得到多个特征值以及这多个特征值对应的特征向量。

[0079] 步骤S4、将多个特征值中最大的n个特征值对应的特征向量构成语言模型，n为大于或等于1的整数。

[0080] 如此构成的语言模型可以保留重要的词过滤掉相对不重要的词。由于步骤3023中是对归一化的N-Gram矩阵进行降维处理得到的语言模型，而N-Gram矩阵以及归一化的N-Gram矩阵可以分别用于描述某一依赖词后出现某一词的次数或者概率，因此该语言模型可以用于根据依赖词以及该依赖词后出现某一词的次数或者频率输出该依赖词后续对应的预测词。

[0081] 在步骤302中构建的语言模型，依赖词所包括的词的个数由N-Gram模型中N的取值决定，例如Bi-Gram模型中，依赖词包括一个词，Tri-Gram模型中，依赖词包括两个词。预测词可以为N-Gram矩阵中依赖词后续出现次数最多的至少一个词，或者，预测词可以为归一

化的N-Gram矩阵中依赖词后续出现频率最高的至少一个词。

[0082] 步骤303、将多个查询语句输入该语言模型，得到该多个查询语句中每个依赖词的预测词。

[0083] 每个依赖词的预测词的个数可以为一个或者多个。

[0084] 步骤304、当多个查询语句中任一依赖词在查询语句中后续的词与任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词。

[0085] 为了保证该热词集合中的多个候选热词均为热词，可以对该热词集合中的多个候选热词进行筛选。步骤305和步骤307分别描述了两种去除热词集合中部分候选热词的方式，可以择一方式进行执行，或者两种方式均执行，本申请实施例对此不进行限制。需要说明的是，图3仅示意性示出了两种方式均执行的流程图，图3并不对其他实现方式进行限制。当两种方式都执行时，热词可以为该两种方式所确定的热词的并集或者交集等。

[0086] 步骤305、根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重。

[0087] 其中，每个候选热词的权重用来指示该候选热词被确定为热词的可能性，权重越大，则该候选热词被确定为热词的可能性越大，相反，权重越小，则该候选热词被确定为热词的可能性越小。

[0088] 边界熵与权重负相关，也即是，边界熵越大，权重越小，而边界熵越小，权重越大；搜索次数与权重正相关，也即是，搜索次数越多，权重越大，而搜索次数越少，权重越小。

[0089] 边界熵也称为信息熵，某个候选热词的边界熵越大，则说明该候选热词与其他词组合的不确定性越高；某个候选热词的边界熵越小，则说明该候选热词与其他词组合的确定性越高，即该候选热词的上下文更固定。由于热词具有上下文较为固定的特性，因此，若某一候选热词的边界熵越小，该候选热词被确定为热词的可能性越大。除此之外，由于热词针对关注的问题和事务，因此，会引发大量用户针对该热词的搜索，因此，若某一候选热词被独立搜索次数较多，则说明该候选热词被确定为热词的可能性越大。

[0090] 可选地，请参考图5，步骤305中，根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重的过程可以包括：

[0091] 步骤3051、将该多个候选热词划分为多个簇，每个簇包括至少一个候选热词，且同一个簇的候选热词所属的查询语句对应于同一个文档。

[0092] 步骤3052、在每个簇中，根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重。

[0093] 文档可以通过统一资源定位符(Uniform Resource Locator，URL)进行访问。由于一个文档描述的可以是当前关注的某一问题或者某一事务，该某一问题或者某一事务可以对应多个候选热词，该多个候选热词之间可以具有关联性或者该多个候选热词可以为相似的概念。因此，以文档为单位来对热词集合中的候选热词进行筛选，可以提高识别的热词的准确性以及效率。

[0094] 需要说明的是，可以预先存储有多组由查询语句(Query)以及URL组成的Query-URL对。一个URL可以对应一个或者多个查询语句，相应的，一个查询语句也可以对应一个或者多个URL，也即是，查询语句与URL为多对多的对应关系。当用户在输入多个查询语句时，热词识别装置可以通过该Query-URL对向用户返回包含该查询语句的文档。当用户输入多

个不同的查询语句，但是该多个不同的查询语句均对应一个相同的URL时，则会形成了一个二分图，该二分图中左边是查询语句，右边是URL。当用户点击了一个URL，则可以认为与该URL成对的至少一个查询语句是相关的。除了如步骤3052中所描述的，在每个簇中确定每个候选热词的权重之外，也可以在一个簇找到最近一段时间用户搜索的高频热词及其相关词，从而将该高频热词及其相关词补充到语料库中，提高热词的准确性。

[0095] 步骤306、去除热词集合中权重小于指定阈值的候选热词。

[0096] 步骤307、去除热词集合中满足指定条件的候选热词，该指定条件包括以停用词开头或结尾、以空格为开头或结尾，或者字符数大于指定值。

[0097] 针对不同的使用目的，设置的停用词可以不同。在本申请实施例中，由于停用词用来去除热词集合中部分候选热词，因此，该停用词可以包括没有实际含义的功能词，例如介词或者代词等。当某一候选热词以停用词开头或结尾，则可以将其从热词集合中去除。除此之外，当某一候选热词以空格为开头或结尾，或者该某一候选热词的字符数大于指定值(例如该某一候选热词的单词长度大于10)，则可以将其从热词集合中去除。

[0098] 步骤308、将热词集合中的候选热词确定为热词。

[0099] 综上所述，本申请实施例提供的热词识别方法，通过将多个查询语句输入根据该多个查询语句构建的语言模型，可以得到该多个查询语句中每个依赖词的预测词，当该任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词，根据该候选热词确定热词。也即是根据用户的查询语句即可确定热词，热词的识别过程中，无需分析大量的文章，减少了热词识别过程中的计算量，解决了相关技术中热词识别的耗时较长的问题，提高了热词识别的效率。

[0100] 并且，再根据候选热词确定热词时，首先去除了热词集合中权重小于指定阈值的候选热词，和/或，去除了热词集合中满足指定条件的候选热词，使得最终根据该候选热词确定的热词更加准确，提高了热词识别的准确率。

[0101] 图6示出了本申请实施例提供的一种热词识别装置600的框图，该热词识别装置600包括：

[0102] 获取模块601，用于获取当前时刻之前指定时间段内用户输入的多个查询语句；

[0103] 模型构建模块602，用于根据所述多个查询语句构建语言模型，所述语言模型用于根据依赖词输出所述依赖词后续对应的预测词，所述依赖词包括至少一个词；

[0104] 输入模块603，用于将所述多个查询语句输入所述语言模型，得到所述多个查询语句中每个依赖词的预测词；

[0105] 集合建立模块604，用于当多个查询语句中任一依赖词在查询语句中后续的词与所述任一依赖词的预测词相同时，将所述任一依赖词的预测词确定为热词集合中的候选热词；

[0106] 热词确定模块605，用于将所述热词集合中的候选热词确定为热词。

[0107] 综上所述，本申请实施例提供的热词识别装置，通过将多个查询语句输入根据该多个查询语句构建的语言模型，可以得到该多个查询语句中每个依赖词的预测词，当该任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词，根据该候选热词确定热词。也即是根据用户的查询语句即可确定热词，热词的识别过程中，无需分析大量的文章，减少了热词识别过程中的计算

量,解决了相关技术中热词识别的耗时较长的问题,提高了热词识别的效率。

[0108] 可选的,模型构建模块602,用于:

[0109] 根据所述多个查询语句构建N-Gram矩阵,所述N-Gram矩阵用于记录所述多个查询语句中任意两个词之间相邻的次数;

[0110] 对所述N-Gram矩阵进行归一化;

[0111] 对归一化后的N-Gram矩阵进行降维,得到所述语言模型。

[0112] 可选的,模型构建模块602,还用于:

[0113] 对归一化后的N-Gram矩阵中的每个值进行中心化;

[0114] 获取进行所述中心化后的N-Gram矩阵的协方差矩阵;

[0115] 对所述协方差矩阵进行特征值分解,得到多个特征值以及这多个特征值对应的特征向量;

[0116] 将所述多个特征值中最大的n个特征值对应的特征向量构成所述语言模型,所述n为大于或等于1的整数。

[0117] 可选的,热词确定模块605,用于:

[0118] 根据每个所述候选热词的边界熵以及作为查询语句的搜索次数确定每个所述候选热词的权重,所述边界熵与所述权重负相关,所述搜索次数与所述权重正相关;

[0119] 去除所述热词集合中所述权重小于指定阈值的候选热词。

[0120] 可选的,热词确定模块605,用于:

[0121] 将所述多个候选热词划分为多个簇,每个簇包括至少一个所述候选热词,且同一个簇的候选热词所属的查询语句对应于同一个文档;

[0122] 在每个所述簇中,根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重。

[0123] 可选的,热词确定模块605,用于:

[0124] 去除所述热词集合中满足指定条件的候选热词,所述指定条件包括以停用词开头或结尾、以空格为开头或结尾,或者字符数大于指定值。

[0125] 综上所述,本申请实施例提供的热词识别装置,通过将多个查询语句输入根据该多个查询语句构建的语言模型,可以得到该多个查询语句中每个依赖词的预测词,当该任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时,将该任一依赖词的预测词确定为热词集合中的候选热词,根据该候选热词确定热词。也即是根据用户的查询语句即可确定热词,热词的识别过程中,无需分析大量的文章,减少了热词识别过程中的计算量,解决了相关技术中热词识别的耗时较长的问题,提高了热词识别的效率。

[0126] 并且,再根据候选热词确定热词时,首先去除了热词集合中权重小于指定阈值的候选热词,和/或,去除了热词集合中满足指定条件的候选热词,使得最终根据该候选热词确定的热词更加准确,提高了热词识别的准确率。

[0127] 图7示出了本申请一个实施例提供的服务器的结构示意图,该服务器可以是图1所示实施环境中的服务器120。

[0128] 服务器700包括中央处理单元(Central Processing Unit,CPU)701、包括随机存取存储器(Random Access Memory,RAM)702和只读存储器(Read-Only Memory,ROM)703的系统存储器704,以及连接系统存储器704和中央处理单元701的系统总线705。服务器700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(Input/output system,I/O系统)706,和用于存储操作系统713、应用程序714和其他程序模块715的大容量存储设备707。

[0129] 基本输入/输出系统706包括有用于显示信息的显示器708和用于用户输入信息的诸如鼠标、键盘之类的输入设备709。其中显示器708和输入设备809都通过连接到系统总线705的输入输出控制器710连接到中央处理单元701。基本输入/输出系统706还可以包括输入输出控制器710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710还提供输出到显示屏、打印机或其他类型的输出设备。

[0130] 大容量存储设备707通过连接到系统总线705的大容量存储控制器(未示出)连接到中央处理单元701。大容量存储设备707及其相关联的计算机可读介质为服务器700提供非易失性存储。也就是说,大容量存储设备707可以包括诸如硬盘或者只读光盘(Compact Disc Read-Only Memory,CD-ROM)驱动器之类的计算机可读介质(未示出)。

[0131] 不失一般性,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、可擦除可编程只读存储器(Erasable Programmable Read Only Memory,EPROM)、带电可擦可编程只读存储器(Electrically Erasable Programmable Read Only Memory,EEPROM)/闪存或其他固态存储其技术,CD-ROM、数字通用光盘(Digital Versatile Disc,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计算机存储介质不局限于上述几种。上述的系统存储器704和大容量存储设备707可以统称为存储器。

[0132] 根据本申请的各种实施例,服务器700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。也即服务器700可以通过连接在系统总线705上的网络接口单元711连接到网络712,或者说,也可以使用网络接口单元711来连接到其他类型的网络或远程计算机系统(未示出)。

[0133] 上述存储器还包括一个或者一个以上的程序,一个或者一个以上程序存储于存储器中,被配置由CPU执行。

[0134] 本申请还提供了一种服务器,该服务器包括: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上述各实施例提供的热词识别方法。

[0135] 本申请还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所述处理器加载并执行,以实现上述各实施例提供的热词识别方法。

[0136] 本申请中术语“和/或”,仅仅是一种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。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0137] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0138] 以上所述仅为本申请的可选实施例，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。

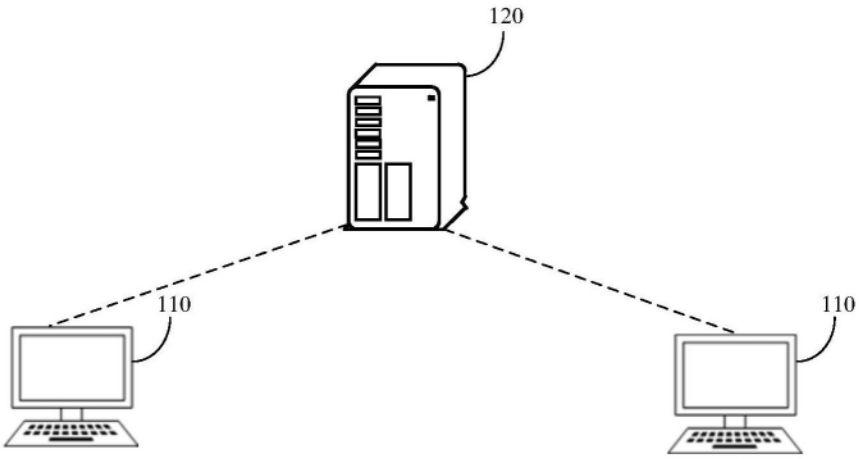

120
110
110

图1

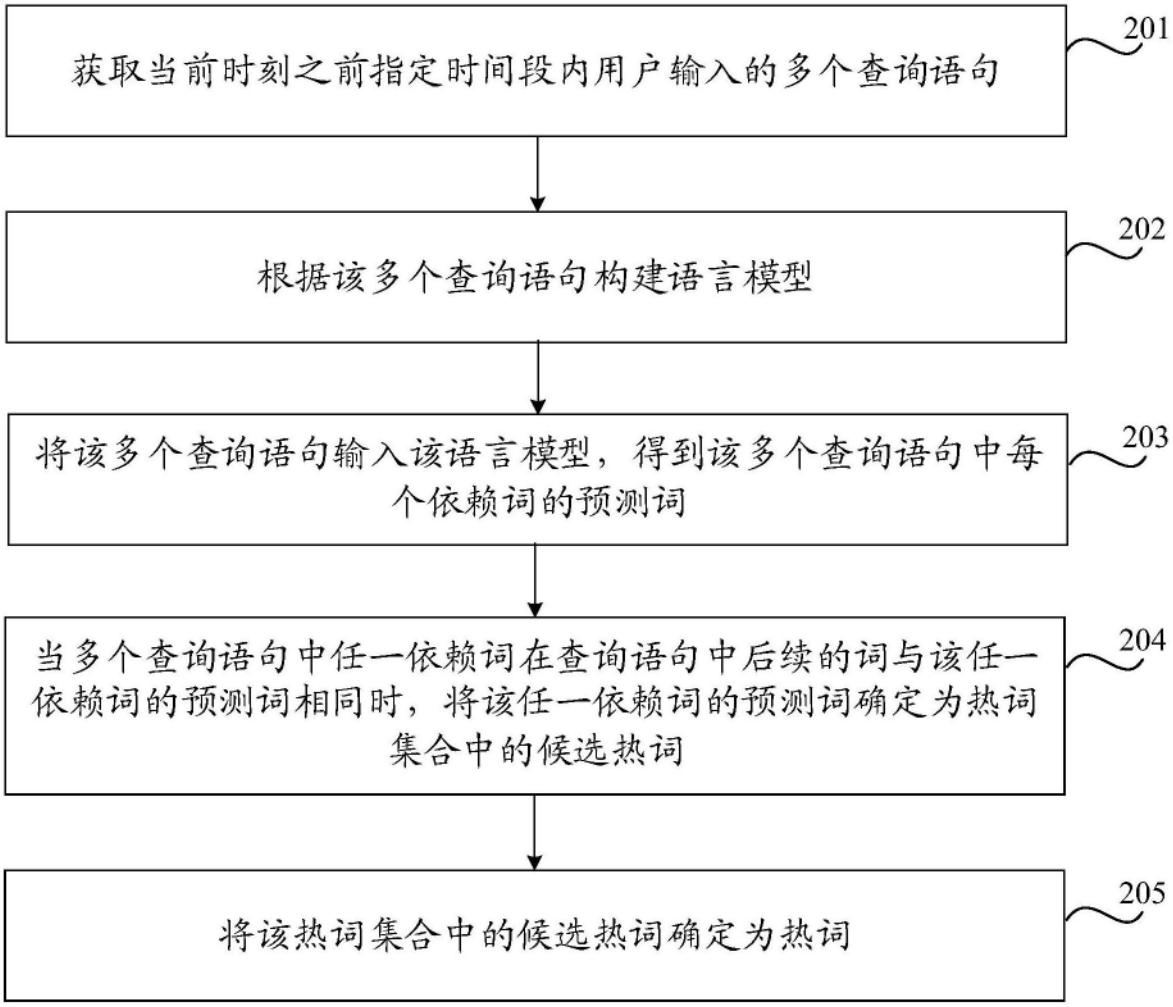

获取当前时刻之前指定时间段内用户输入的多个查询语句
201
根据该多个查询语句构建语言模型
202
将该多个查询语句输入该语言模型，得到该多个查询语句中每个依赖词的预测词
203
当多个查询语句中任一依赖词在查询语句中后续的词与该任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词
204
将该热词集合中的候选热词确定为热词
205

图2

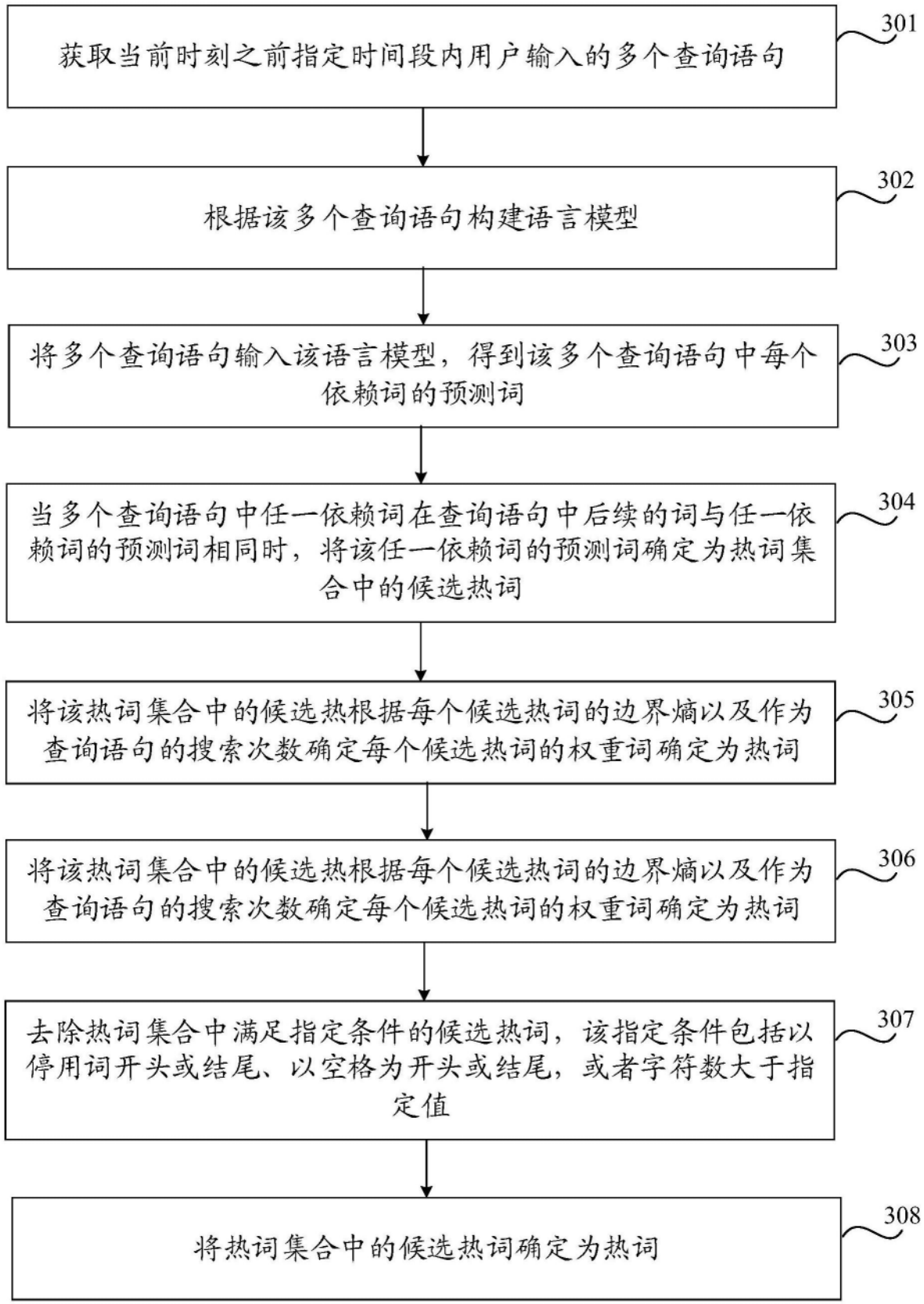

获取当前时刻之前指定时间段内用户输入的多个查询语句
301
根据该多个查询语句构建语言模型
302
将多个查询语句输入该语言模型，得到该多个查询语句中每个依赖词的预测词
303
当多个查询语句中任一依赖词在查询语句中后续的词与任一依赖词的预测词相同时，将该任一依赖词的预测词确定为热词集合中的候选热词
304
将该热词集合中的候选热根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重词确定为热词
305
将该热词集合中的候选热根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重词确定为热词
306
去除热词集合中满足指定条件的候选热词，该指定条件包括以停用词开头或结尾、以空格为开头或结尾，或者字符数大于指定值
307
将热词集合中的候选热词确定为热词
308

图3

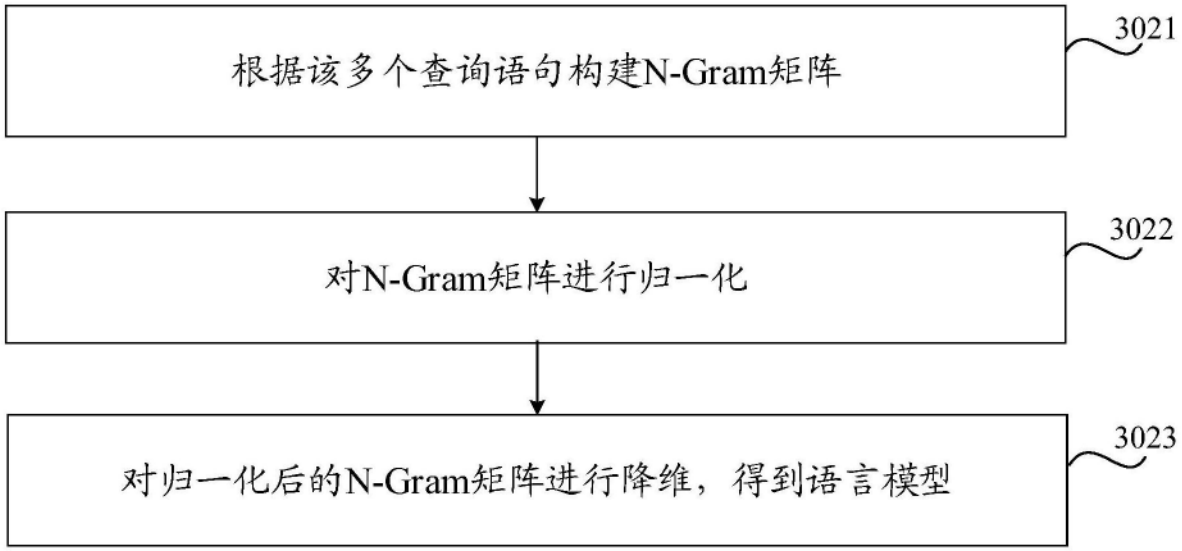

根据该多个查询语句构建N-Gram矩阵
3021
对N-Gram矩阵进行归一化
3022
对归一化后的N-Gram矩阵进行降维，得到语言模型
3023

图4

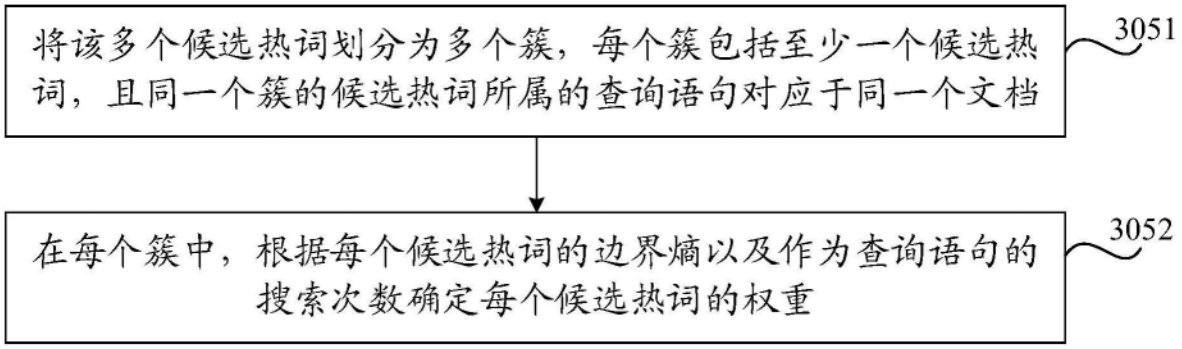

将该多个候选热词划分为多个簇，每个簇包括至少一个候选热词，且同一个簇的候选热词所属的查询语句对应于同一个文档
3051
在每个簇中，根据每个候选热词的边界熵以及作为查询语句的搜索次数确定每个候选热词的权重
3052

图5

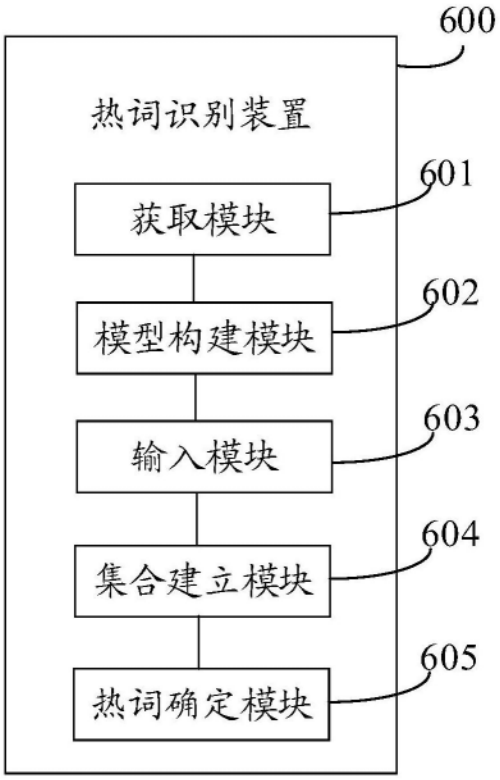

600
热词识别装置
601
获取模块
602
模型构建模块
603
输入模块
604
集合建立模块
605
热词确定模块

图6

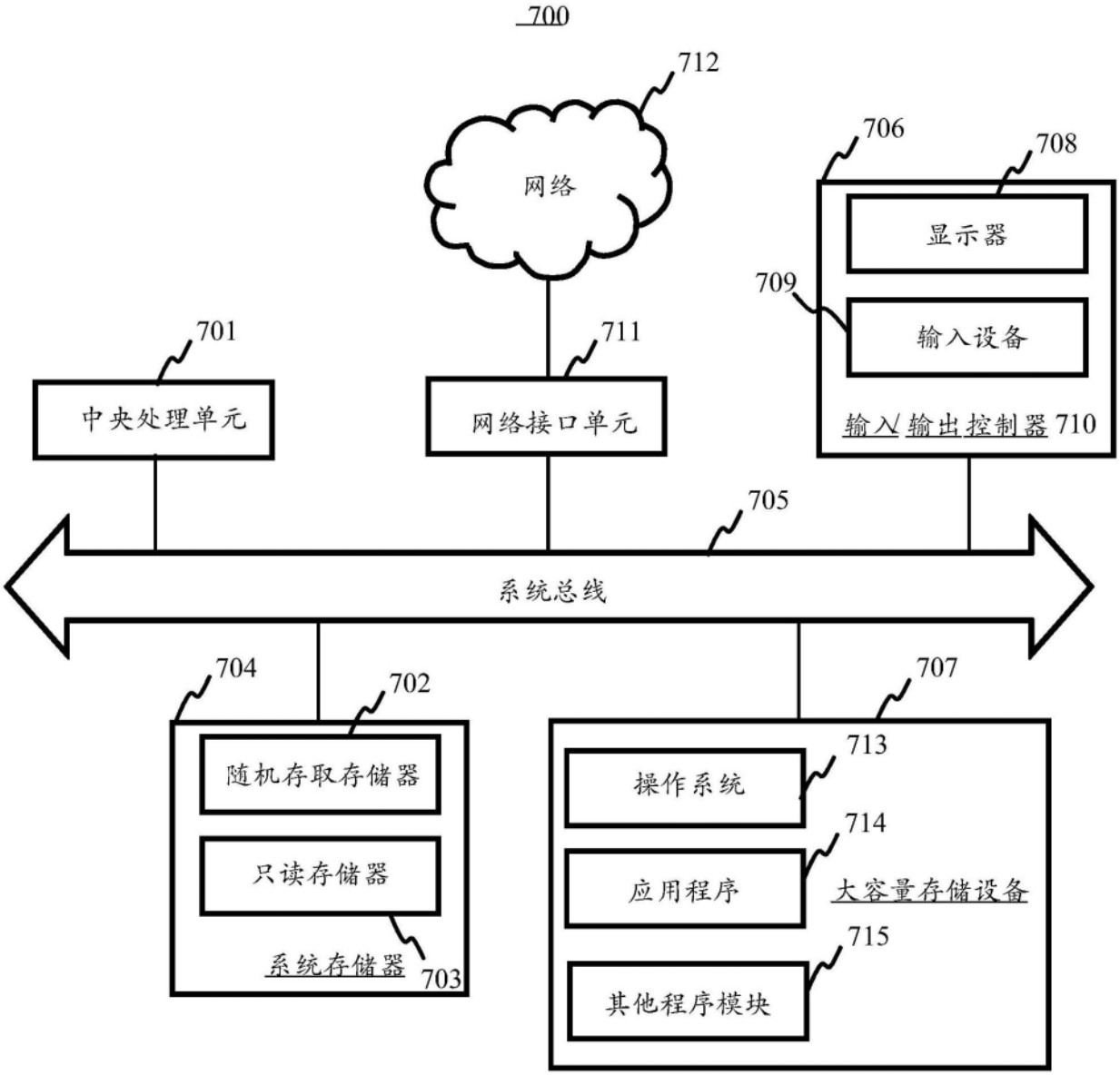

700
712
网络
706
708
显示器
709
输入设备
701
711
中央处理单元
网络接口单元
输入/输出控制器 710
705
系统总线
704
702
707
713
随机存取存储器
操作系统
714
只读存储器
应用程序
大容量存储设备
系统存储器
703
715
其他程序模块

图7