



(12) 发明专利

(10) 授权公告号 CN 114625859 B

(45) 授权公告日 2026. 06. 05

(21) 申请号 202210303330.2

G06F 16/36 (2019.01)

(22) 申请日 2022.03.24

G06F 40/279 (2020.01)

G06F 40/30 (2020.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 114625859 A

(56) 对比文件

CN 110196901 A, 2019.09.03

CN 105677664 A, 2016.06.15

(43) 申请公布日 2022.06.14

(73) 专利权人 北京明略昭辉科技有限公司

胡小荣等.融合多策略的短语识别方法.《情报科学》.2019, 37(6), 第49-54页.

地址 100098 北京市海淀区北三环西路25

号27号楼二层2020室

审查员 李维

(72) 发明人 林阔

(74) 专利代理机构 北京华夏泰和知识产权代理

有限公司 11662

专利代理师 沈园园

(51) Int. Cl.

G06F 16/332 (2025.01)

G06F 16/335 (2019.01)

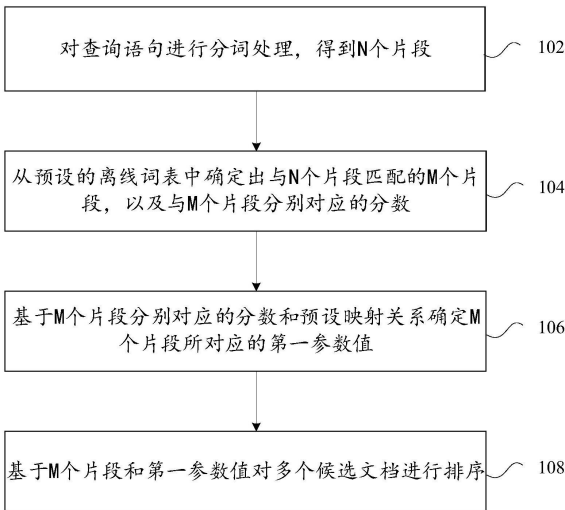
权利要求书2页 说明书9页 附图8页

(54) 发明名称

一种文档的排序方法及装置

(57) 摘要

本申请提供了一种文档的排序方法及装置,其中,该方法包括:对查询语句进行分词处理,得到N个片段;从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;基于M个片段和第一参数值对多个候选文档进行排序。通过本申请,解决了现有技术中以查询语句中两个相邻词在文档中的距离作为紧密度评价标准导致的召回结果不精确的技术问题。



1. 一种文档的排序方法,其特征在于,包括:

对查询语句进行分词处理,得到N个片段;

从预设的离线词表中确定出与所述N个片段匹配的M个片段,以及与所述M个片段分别对应的分数;其中,所述离线词表中包括多个目标候选片段以及与所述多个目标候选片段对应的分数,所述N为正整数,M为小于或等于N的正整数;其中,所述目标候选片段为分数大于第一预设阈值的候选片段,所述候选片段的分数通过如下方式确定:基于点间互信息计算所述候选片段对应的第一特征值;基于信息熵计算所述候选片段对应的第二特征值;基于词频计算所述候选片段对应的第三特征值;基于逆文档频率计算所述候选片段对应的第四特征值;基于所述第一特征值、所述第二特征值、所述第三特征值、所述第四特征值确定所述候选片段对应的分数;

基于所述M个片段分别对应的分数和预设映射关系确定所述M个片段所对应的第一参数值;其中,所述预设映射关系用于表征片段的分数与第一参数值之间的映射关系,所述第一参数值用于表征相隔所述参数值距离的词与待查询片段匹配;第一参数值是指在使用ES进行短语匹配时的slop值,slop值的大小取决于最少需要将词移动多少次才能得到待查询片段,slop值越小则片段的紧密度越高,对应的分数也越高;

基于所述M个片段和所述第一参数值对多个候选文档进行排序。

2. 根据权利要求1所述的方法,其特征在于,在从预设的离线词表中确定出与所述N个片段匹配的M个片段,以及与所述M个片段分别对应的分数之前,所述方法还包括:

获取离线文档库;

对所述离线文档库中的文档进行分词处理,得到第一分词结果;其中,所述第一分词结果中包括所述多个候选片段;

对所述多个候选片段进行计算,得到与所述多个候选片段分别对应的分数;

从所述多个候选片段中筛选出分数大于第一预设阈值的多个所述目标候选片段。

3. 根据权利要求1所述的方法,其特征在于,所述基于点间互信息计算所述候选片段对应的第一特征值,包括:

通过以下公式计算所述候选片段对应的第一特征值:

$$PMI(x, y) = P(y|x)/P(x);$$

其中,所述x和所述y为所述候选片段中的词组,所述PMI是指在所述x出现的情况下,所述y出现的概率;所述 $P(y|x)$ 是指所述x和所述y一起出现的概率;所述 $P(x)$ 是指所述x出现的概率。

4. 根据权利要求1所述的方法,其特征在于,基于信息熵计算所述候选片段对应的第二特征值,包括:

通过以下公式计算所述候选片段对应的所述第二特征值,所述第二特征值为 E_L 和 E_R 中的较小值:

$$E_L = - \sum_{a \in Z} P(aW|W) \cdot \log_2 P(aW|W) - \sum_{a \in N} P(aW|W) \cdot \log_2 P(1|W)$$

$$E_R = - \sum_{b \in Z} P(Wb|W) \cdot \log_2 P(Wb|W) - \sum_{b \in N} P(Wb|W) \cdot \log_2 P(1|W)$$

其中,所述 E_L 是指左熵,所述 E_R 是指右熵; W 是指所述候选短语;所述 z 是指所述候选短语 W 的上下文,所述 a 是指代表所述候选短语 W 左侧的词,所述 b 代表所述候选短语 W 右侧的词;所述 N 是指上下文缺失,所述 aW 是指上文缺失,所述 Wb 是指下文缺失。

5.根据权利要求1所述的方法,其特征在于,所述对所述第一特征值、所述第二特征值、所述第三特征值、所述第四特征值确定所述候选片段对应的分数包括:

对所述第一特征值、所述第二特征值、所述第三特征值、所述第四特征值进行归一化处理;

将归一化后的所述第一特征值、归一化后的所述第二特征值、归一化后的所述第三特征值、归一化后的所述第四特征值求和得到所述候选片段对应的分数。

6.根据权利要求2所述的方法,其特征在于,所述基于所述 M 个片段和所述第一参数值对多个候选文档进行排序包括:

从多个目标文档中筛选出所述多个候选文档;其中,所述候选文档包括所述 M 个片段,且所述 M 个片段所对应的第一参数值小于或等于预设阈值;

基于所述第一参数值对所述多个候选文档进行排序。

7.一种文档的排序装置,其特征在于,包括:

第一处理模块,用于对查询语句进行分词处理,得到 N 个片段;

第一确定模块,用于从预设的离线词表中确定出与所述 N 个片段匹配的 M 个片段,以及与所述 M 个片段分别对应的分数;其中,所述离线词表中包括多个目标候选片段以及与所述多个目标候选片段对应的分数,所述 N 为正整数, M 为小于或等于 N 的正整数;其中,所述目标候选片段为分数大于第一预设阈值的候选片段,所述候选片段的分数通过如下方式确定:基于点间互信息计算所述候选片段对应的第一特征值;基于信息熵计算所述候选片段对应的第二特征值;基于词频计算所述候选片段对应的第三特征值;基于逆文档频率计算所述候选片段对应的第四特征值;基于所述第一特征值、所述第二特征值、所述第三特征值、所述第四特征值确定所述候选片段对应的分数;

第二确定模块,用于基于所述 M 个片段分别对应的分数和预设映射关系确定所述 M 个片段所对应的第一参数值;其中,所述预设映射关系用于表征片段的分数与第一参数之间的映射关系,所述第一参数值用于表征相隔所述参数值距离的词与待查询片段匹配;第一参数值是指在使用ES进行短语匹配时的 $slop$ 值, $slop$ 值的大小取决于最少需要将词移动多少次才能得到待查询片段, $slop$ 值越小则片段的紧密度越高,对应的分数也越高;

排序模块,用于基于所述 M 个片段和所述第一参数值对多个候选文档进行排序。

8.一种电子设备,其特征在于,包括处理器、通信接口、存储器和通信总线,其中,处理器,通信接口,存储器通过通信总线完成相互间的通信;

存储器,用于存放计算机程序;

处理器,用于执行存储器上所存放的程序时,实现权利要求1-6任一所述的方法。

9.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现如权利要求1-6中任一所述的方法。

一种文档的排序方法及装置

技术领域

[0001] 本申请涉及信息检索领域,特别是涉及一种文档的排序方法及装置。

背景技术

[0002] Elasticsearch(简称ES)是一个分布式、高扩展、高实时的搜索与数据分析引擎,可以用于搜索各种文档。在使用ES进行文档搜索的过程中,最重要的一个环节是召回。召回是指根据输入的查询语句,能够高效的获取查询语句相关的候选文档集合的过程。现有技术存在以查询语句中两个相邻词在文档中的距离作为紧密度评价标准导致的召回结果不精确的技术问题。

发明内容

[0003] 本申请实施例的目的在于提供一种文档的排序方法及装置,解决了现有技术中存在的以查询语句中两个相邻词在文档中的距离作为紧密度评价标准导致的召回结果不精确的技术问题。具体技术方案如下:

[0004] 在本申请实施的第一方面,首先提供了一种文档的排序方法,所述方法包括:对查询语句进行分词处理,得到N个片段;从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;其中,离线词表中包括多个目标候选片段以及与多个目标候选片段对应的分数,N为正整数,M为小于或等于N的正整数;基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;其中,预设映射关系用于表征片段的分数与第一参数之间的映射关系,第一参数值用于表征相隔参数值距离的词与待查询片段匹配;基于M个片段和第一参数值对多个候选文档进行排序。

[0005] 在本申请实施的第二方面,还提供了一种文档的排序装置,所述装置包括:第一处理模块,用于对查询语句进行分词处理,得到N个片段;第一确定模块,用于从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;其中,离线词表中包括多个目标候选片段以及与多个目标候选片段对应的分数,N为正整数,M为小于或等于N的正整数;第二确定模块,用于基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;其中,预设映射关系用于表征片段的分数与第一参数之间的映射关系,第一参数值用于表征相隔参数值距离的词与待查询片段匹配;排序模块,用于基于M个片段和第一参数值对多个候选文档进行排序。

[0006] 在本申请实施的第三方面,还提供了一种电子设备,包括处理器、通信接口、存储器和通信总线,其中,处理器,通信接口,存储器通过通信总线完成相互间的通信;存储器,用于存放计算机程序;处理器,用于执行存储器上所存放的程序时,实现第一方面所述的方法步骤。

[0007] 在本申请实施的第四方面,还提供了一种计算机可读存储介质,所述计算机可读存储介质中存储有指令,当其在计算机上运行时,使得计算机执行上述第一方面所述的文档的排序方法。

[0008] 本申请可以应用于信息检索领域用于对ES召回结果进行优化。本申请实施例提供的文档的排序方法及装置,通过对查询语句进行分词处理,得到N个片段;从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;其中,离线词表中包括多个目标候选片段以及与多个目标候选片段对应的分数,N为正整数,M为小于或等于N的正整数;基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;其中,预设映射关系用于表征片段的分数与第一参数值之间的映射关系,第一参数值用于表征相隔参数值距离的词与待查询片段匹配;基于M个片段和第一参数值对多个候选文档进行排序;也就是说,通过建立离线词表对文档库中的词进行紧密度评价并与待查询词匹配,从而解决了现有技术中以查询语句中两个相邻词在文档中的距离作为紧密度评价标准导致的召回结果不精确的技术问题。

附图说明

[0009] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍。

- [0010] 图1为本申请实施例中文档的排序方法流程图之一;
- [0011] 图2为本申请实施例中文档的排序方法流程图之二;
- [0012] 图3为本申请实施例中文档的排序方法流程图之三;
- [0013] 图4为本申请实施例中文档的排序方法流程图之四;
- [0014] 图5为本申请实施例中文档的排序方法流程图之五;
- [0015] 图6为本申请实施例中文档的排序方法一示范例流程图;
- [0016] 图7为本申请实施例中文档的排序装置结构示意图之一;
- [0017] 图8为本申请实施例中文档的排序装置结构示意图之二;
- [0018] 图9为本申请实施例中文档的排序装置结构示意图之三;
- [0019] 图10为本申请实施例中文档的排序装置结构示意图之四;
- [0020] 图11为本申请实施例中文档的排序装置结构示意图之五;
- [0021] 图12为本申请实施例中电子设备的结构示意图。

具体实施方式

[0022] 为使本申请实施例的目的、技术方案和优点更加清楚,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,所描述地实施例是本申请的一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动的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。

[0023] 在后续的描述中,使用用于表示元件的诸如“模块”、“单元”的后缀仅为了有利于本申请的说明,其本身并没有特定的意义。因此,“模块”与“部件”可以混合地使用。

[0024] 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描述。本申请实施例提供了一种文档的排序方法,如图1所示,该方法包括以下步骤:

[0025] 步骤102:对查询语句进行分词处理,得到N个片段;

[0026] 其中,需要说明的是,分词处理是指将连续的字序列按照一定的规范重新组合成语义独立词序列的过程,包括基于字符串匹配的分词方法、基于理解的分词方法和基于统

计的分词方法;N为正整数;片段可以是指词语,也可以是单个的字;在一示范例中,对查询语句“知识图谱”进行分词,得到“知识”和“图谱”两个片段。

[0027] 步骤104:从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;其中,离线词表中包括多个目标候选片段以及与多个目标候选片段对应的分数,N为正整数,M为小于或等于N的正整数;

[0028] 其中,需要说明的是,离线词表中的片段根据离线文档库中的文档进行确定,并根据相应片的特征进行打分。

[0029] 步骤106:基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;其中,预设映射关系用于表征片段的分数与第一参数之间的映射关系,第一参数值用于表征相隔参数值距离的词与待查询片段匹配;

[0030] 其中,需要说明的是,第一参数值是指在使用ES进行短语匹配时的slop值;短语匹配用于匹配含有正确片段出现顺序的文档,且在片段之间没有插入别的片段;在ES技术中的短语匹配是使用特定于域的语言(Domain Specific Language,简称DSL)语句完成的,其中,ES提供了基于JSON的完整查询DSL来定义查询;在第一个示范例中,使用“知识图谱”来做短语匹配,必须保证匹配的文档中“图谱”出现在“知识”后面一位,只含有“图谱知识”的文档则无法进行匹配;slop值的大小取决于最少需要将词移动多少次才能得到待查询片段,slop值越小则片段的紧密度越高,对应的分数也越高;在第二个示范例中,待查询片段为“解决方案”的情况下,在第一参数值为1时,“解决的方案”通过移动一次词即可得到“解决方案”,因此“解决的方案”可以与待查询片段匹配;在第三个示范例中,待查询片段为“知识图谱”的情况下,在第一参数值为0时,文档中必须出现“知识图谱”这一片段才可以进行匹配;映射是指指两个元素的集之间元素相互对应的关系;在第四个示范例中,片段“知识的图谱”对应的分数为10,片段“知识构建图谱”对应的分数为5分,通过预设映射关系确定的片段“知识的图谱”对应的第一参数值为1,通过预设映射关系确定的片段“知识构建图谱”对应的第一参数值为2。

[0031] 步骤108:基于M个片段和第一参数值对多个候选文档进行排序。

[0032] 通过本申请实施例的上述步骤102至步骤108,对查询语句进行分词处理,得到N个片段;从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;其中,离线词表中包括多个目标候选片段以及与多个目标候选片段对应的分数,N为正整数,M为小于或等于N的正整数;基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;其中,预设映射关系用于表征片段的分数与第一参数之间的映射关系,第一参数值用于表征相隔参数值距离的词与待查询片段匹配;基于M个片段和第一参数值对多个候选文档进行排序;也就是说,通过建立离线词表对文档库中的词进行紧密度评价并与待查询词匹配,从而解决了现有技术中存在的以查询语句中两个相邻词在文档中的距离作为紧密度评价标准导致的召回结果不精确的技术问题。

[0033] 在本申请实施例的可选实施方式中,本申请步骤104中涉及到的在从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数之前,如图2所示,包括:

[0034] 步骤202:获取离线文档库;

[0035] 步骤204:对离线文档库中的文档进行分词处理,得到第一分词结果,其中,第一分

词结果中包括多个候选片段;

[0036] 其中,需要说明的是,对文档进行分词处理是指将连续的语句转化为中文词汇的集合,可以通过分词器进行;第一分词结果通过将相邻的几个中文词汇组合得到。

[0037] 步骤206:对多个候选片段进行计算,得到与多个候选片段分别对应的分数;

[0038] 其中,需要说明的是,对多个候选片段进行计算是指对文档库的文档进行离线的片段挖掘,并根据片段的特征打分并存入离线词表中。

[0039] 步骤208:从多个候选片段中筛选出分数大于第一预设阈值的多个目标候选片段。

[0040] 其中,需要说明的是,第一预设阈值可以根据用户需求或者候选片段的得分情况进行设置;在一示范例中,分数低于5的片段紧密度过低,其中,紧密度是指候选片段之间的紧密程度,则可以将第一预设阈值设置为5,只筛选分数大于5的候选片段作为目标候选片段。

[0041] 可见,本申请实施例提供的文档的排序方法根据离线文档中片段的紧密度特征进行打分并筛选,不仅仅考察片段在文档中的距离,提高了召回的精确性。

[0042] 本申请实施例提供的文档的排序方法,步骤206中涉及的到的对多个候选片段进行计算,得到与多个候选片段分别对应的分数,如图3所示,包括:

[0043] 步骤302:基于点间互信息计算候选片段对应的第一特征值;

[0044] 其中,需要说明的是,点间互信息是指一种用来衡量两个具体事件关联强度的统计量;点间互信息的值越大,表明两个事件的关联性越强,点间互信息的值越小表明两个事件关联性越弱。

[0045] 步骤304:基于信息熵计算候选片段对应的第二特征值;

[0046] 其中,需要说明的是,信息熵是信息量的度量指标,片段两侧的信息熵越高,说明该片段两侧的信息量越大,该片段两侧可组合的词比较丰富,更接近于一个可以独立使用的短语。

[0047] 步骤306:基于词频计算候选片段对应的第三特征值;

[0048] 其中,需要说明的是,词频是指的是某一个片段在文档中出现的次数;可以通过考察词频来确定片段的可信性,片段出现的频率越高则说明片段中的词语和字可以作为固定搭配的概率越高,更加可信。

[0049] 步骤308:基于逆文档频率计算候选片段对应的第四特征值;

[0050] 其中,需要说明的是,逆文档频率又称反文档频率,是文档频率的倒数。

[0051] 步骤310:基于第一特征值、第二特征值、第三特征值、第四特征值确定候选片段对应的分数。

[0052] 其中,需要说明的是,将第一特征值、第二特征值、第三特征值、第四特征值归一化处理到同一量级,并将归一化后的第一特征值、归一化后的第二特征值、归一化后的第三特征值、归一化后的第四特征值相加得到候选片段对应的分数。

[0053] 可见,本申请实施例提供的文档的排序方法基于点间互信息、信息熵、词频、逆文档频率对文档中的片段进行紧密度打分,使得ES的召回结果更精确。

[0054] 本申请实施例提供的文档的排序方法,步骤302中涉及到的基于点间互信息计算候选片段对应的第一特征值,包括:

[0055] 通过以下公式计算候选片段对应的第一特征值:

[0056] $PMI(x, y) = P(y|x) / P(x)$;

[0057] 其中, x 和 y 为候选片段中的词组, PMI是指在 x 出现的情况下, y 出现的概率; $P(y|x)$ 是指 x 和 y 一起出现的概率; $P(x)$ 是指 x 出现的概率。

[0058] 可见, 本申请实施例提供的文档的排序方法可以通过公式计算片段之间的点间互信息, 判断两个片段之间是否足够紧密。

[0059] 本申请实施例提供的文档的排序方法, 步骤304中涉及到的基于信息熵计算候选片段对应的第二特征值, 包括:

[0060] 通过以下公式计算候选片段对应的第二特征值, 第二特征值为 E_L 和 E_R 中的较小值:

$$[0061] \quad E_L = - \sum_{a \in z} P(aW|W) \cdot \log_2 P(aW|W) - \sum_{a \in N} P(aW|W) \cdot \log_2 P(1|W)$$

$$[0062] \quad E_R = - \sum_{b \in z} P(Wb|W) \cdot \log_2 P(Wb|W) - \sum_{b \in N} P(Wb|W) \cdot \log_2 P(1|W)$$

[0063] 其中, E_L 是指左熵, E_R 是指右熵; W 是指候选短语; z 是指候选短语 W 的上下文, a 是指代表候选短语 W 左侧的词, b 代表候选短语 W 右侧的词; N 是指上下文缺失, aW 是指上文缺失, Wb 是指下文缺失。

[0064] 可见, 本申请实施例提供的文档的排序方法可以通过公式确定片段的信息熵, 考察片段是否独立和是否为常用的短语。

[0065] 本申请实施例提供的文档的排序方法, 步骤310中涉及到的对第一特征值、第二特征值、第三特征值、第四特征值确定候选片段对应的分数, 如图4所示, 包括:

[0066] 步骤402: 对第一特征值、第二特征值、第三特征值、第四特征值进行归一化处理;

[0067] 其中, 需要说明的是, 对第一特征值、第二特征值、第三特征值、第四特征值进行归一化处理是指将第一特征值、第二特征值、第三特征值、第四特征值处理为同一个量级的值。

[0068] 步骤404: 将归一化后的第一特征值、归一化后的第二特征值、归一化后的第三特征值、归一化后的第四特征值求和得到候选片段对应的分数。

[0069] 可见, 本申请实施例提供的文档的排序方法基于点间互信息、信息熵、词频、逆文档频率对文档中的片段进行紧密度打分, 使得ES的召回结果更精确。

[0070] 本申请实施例提供的文档的排序方法, 步骤108中涉及到的基于 M 个片段和第一参数值对多个候选文档进行排序, 如图5所示, 包括:

[0071] 步骤502: 从多个目标文档中筛选出多个候选文档; 其中, 候选文档包括 M 个片段, 且 M 个片段所对应的第一参数值小于或等于预设阈值;

[0072] 步骤504: 基于第一参数值对多个候选文档进行排序。

[0073] 其中, 需要说明的是, 排序可以是第一参数值越大的候选文档排序越靠前, 也可以是第一参数值越大的候选文档排序越靠后。

[0074] 可见, 本申请实施例提供的文档的排序方法可以通过设置查询片段和对应的 $slop$ 值, 得到紧密度从高到低或从低到高的候选文档。

[0075] 在本申请实施例的一示范例中, 本申请实施例提供的文档的排序方法的查询语句为“知识图谱”, 流程图如图7所示, 本申请实施例的步骤包括:

[0076] 步骤601: 对查询语句“知识图谱”进行分词处理, 得到“知识”和“图谱”两个片段;

[0077] 步骤602:获取离线文档库;

[0078] 步骤603:对离线文档库中的文档进行分词处理,得到第一分词结果,第一分词结果中包括“知识的图谱”和“知识构建图谱”;

[0079] 步骤604:基于点间互信息计算候选片段“知识的图谱”对应的第一特征值为20,片段“知识构建图谱”对应的第一特征值为10;

[0080] 步骤605:基于信息熵计算候选片段“知识的图谱”对应的第二特征值为30,片段“知识构建图谱”对应的第二特征值为15;

[0081] 步骤606:基于词频计算候选片段“知识的图谱”对应的第三特征值为30,候选片段“知识构建图谱”对应的第三特征值为15;

[0082] 步骤607:基于逆文档频率计算候选片段“知识的图谱”对应的第四特征值为40,候选片段“知识构建图谱”对应的第四特征值为20;

[0083] 步骤608:对候选片段“知识的图谱”对应的第一特征值、第二特征值、第三特征值、第四特征值进行归一化处理得到归一化后的第一特征值为4,归一化后的第二特征值为6,归一化后的第三特征值为6,归一化后的第四特征值为8;对候选片段“知识构建图谱”对应的第一特征值、第二特征值、第三特征值、第四特征值进行归一化处理得到归一化后的第一特征值为2,归一化后的第二特征值为3,归一化后的第三特征值为3,归一化后的第四特征值为4;

[0084] 步骤609:将归一化后的第一特征值、归一化后的第二特征值、归一化后的第三特征值、归一化后的第四特征值求和得到候选片段对应的分数,候选片段“知识的图谱”得分为24,候选片段“知识构建图谱”得分为12;

[0085] 步骤610:第一预设阈值为10,将得分大于10的候选片段“知识的图谱”和候选片段“知识构建图谱”确定为目标候选片段;

[0086] 步骤611:与“知识”和“图谱”两个片段匹配的片段为目标候选片段“知识的图谱”和“知识构建图谱”;

[0087] 步骤612:通过预设映射关系确定目标候选片段“知识的图谱”对应的第一参数值为1,目标候选片段“知识构建图谱”对应的第一参数值为2;

[0088] 步骤613:第一参数值小越小的排序越靠前,召回结果中,含有目标候选片段“知识的图谱”的文档排在含有目标候选片段“知识构建图谱”的文档前面。

[0089] 可见,本申请实施例所提供的文档的排序方法,可以通过建立离线词表对文档库中的词进行紧密度评价并与待查询词匹配,从而解决现有技术中存在的以查询语句中两个相邻词在文档中的距离作为紧密度评价标准导致的召回结果不精确的技术问题。

[0090] 本申请实施例提供了一种文档的排序装置,如图7所示,该装置包括:

[0091] 第一处理模块72,用于对查询语句进行分词处理,得到N个片段;

[0092] 第一确定模块74,用于从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;其中,离线词表中包括多个目标候选片段以及与多个目标候选片段对应的分数,N为正整数,M为小于或等于N的正整数;

[0093] 第二确定模块76,用于基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;其中,预设映射关系用于表征片段的分数与第一参数之间的映射关系,第一参数值用于表征相隔参数值距离的词与待查询片段匹配;

[0094] 排序模块78,用于基于M个片段和第一参数值对多个候选文档进行排序。

[0095] 通过本申请实施例提供的文档的排序装置,通过第一处理模块对查询语句进行分词处理,得到N个片段;通过第一确定模块从预设的离线词表中确定出与N个片段匹配的M个片段,以及与M个片段分别对应的分数;其中,离线词表中包括多个目标候选片段以及与多个目标候选片段对应的分数,N为正整数,M为小于或等于N的正整数;通过第二确定模块基于M个片段分别对应的分数和预设映射关系确定M个片段所对应的第一参数值;其中,预设映射关系用于表征片段的分数与第一参数之间的映射关系,第一参数值用于表征相隔参数值距离的词与待查询片段匹配;通过排序模块基于M个片段和第一参数值对多个候选文档进行排序;也就是说,通过建立离线词表对文档库中的词进行紧密度评价并与待查询词匹配,从而解决了现有技术中存在的以查询语句中两个相邻词在文档中的距离作为紧密度评价标准导致的召回结果不精确的技术问题。

[0096] 本申请实施例提供了一种文档的排序装置,如图8所示,该装置还包括:

[0097] 获取模块82,用于获取离线文档库;

[0098] 第二处理模块84,用于对离线文档库中的文档进行分词处理,得到第一分词结果,其中,第一分词结果中包括多个候选片段;

[0099] 第三处理模块86,用于对多个候选片段进行计算,得到与多个候选片段分别对应的分数;

[0100] 筛选模块88,用于从多个候选片段中筛选出分数大于第一预设阈值的多个目标候选片段。

[0101] 在本申请实施例的可选实施方式中,本申请实施例提供的第三处理模块86,如图9所示,包括:

[0102] 第一计算单元902,用于基于点间互信息计算候选片段对应的第一特征值;

[0103] 第二计算单元904,用于基于信息熵计算候选片段对应的第二特征值;

[0104] 第三计算单元906,用于基于词频计算候选片段对应的第三特征值;

[0105] 第四计算单元908,用于基于逆文档频率计算候选片段对应的第四特征值;

[0106] 确定单元910,用于基于第一特征值、第二特征值、第三特征值、第四特征值确定候选片段对应的分数。

[0107] 在本申请实施例的可选实施方式中,本申请实施例提供的第一计算单元902,包括:第一计算子单元,用于通过以下公式计算候选片段对应的第一特征值:

[0108] $PMI(x, y) = P(y|x) / P(x)$;

[0109] 其中,x和y为候选片段中的词组,PMI是指在x出现的情况下,y出现的概率; $P(y|x)$ 是指x和y一起出现的概率; $P(x)$ 是指x出现的概率。

[0110] 在本申请实施例的可选实施方式中,本申请实施例提供的第二计算单元904,包括:第二计算子单元,用于通过以下公式计算候选片段对应的第二特征值,第二特征值为 E_L 和 E_R 中的较小值:

$$[0111] \quad E_L = - \sum_{a \in Z} P(aW|W) \cdot \log_2 P(aW|W) - \sum_{a \in N} P(aW|W) \cdot \log_2 P(1|W)$$

$$[0112] \quad E_R = - \sum_{b \in Z} P(Wb|W) \cdot \log_2 P(Wb|W) - \sum_{b \in N} P(Wb|W) \cdot \log_2 P(1|W)$$

[0113] 其中, E_L 是指左熵, E_R 是指右熵; W 是指候选短语; z 是指候选短语 W 的上下文, a 是指代表候选短语 W 左侧的词, b 代表候选短语 W 右侧的词; N 是指上下文缺失, aW 是指上文缺失, Wb 是指下文缺失。

[0114] 在本申请实施例的可选实施方式中,本申请实施例提供的确定单元910,如图10所示,包括:

[0115] 第一处理子单元1002,用于对第一特征值、第二特征值、第三特征值、第四特征值进行归一化处理;

[0116] 第二处理子单元1004,将归一化后的第一特征值、归一化后的第二特征值、归一化后的第三特征值、归一化后的第四特征值求和得到候选片段对应的分数。

[0117] 在本申请实施例的可选实施方式中,本申请实施例提供的排序模块78,如图11所示,包括:

[0118] 筛选单元1102,用于从多个目标文档中筛选出多个候选文档;其中,候选文档包括 M 个片段,且 M 个片段所对应的第一参数值小于或等于预设阈值;

[0119] 排序单元1104,用于基于第一参数值对多个候选文档进行排序。

[0120] 本申请实施例还提供了一种电子设备,如图12所示,包括处理器1201、通信接口1202、存储器1203和通信总线1204,其中,处理器1201,通信接口1202,存储器1203通过通信总线1204完成相互间的通信,

[0121] 存储器1203,用于存放计算机程序;

[0122] 处理器1201,用于执行存储器1203上所存放的程序时,实现图1中的方法步骤,其所起到的作用与图1中的方法步骤一样,在此不再赘述。

[0123] 上述终端提到的通信总线可以是外设部件互连标准(Peripheral Component Interconnect,简称PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture,简称EISA)总线等。该通信总线可以分为地址总线、数据总线、控制总线等。为便于表示,图12中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0124] 通信接口用于上述终端与其他设备之间的通信。

[0125] 存储器可以包括随机存取存储器(Random Access Memory,简称RAM),也可以包括非易失性存储器(non-volatile memory),例如至少一个磁盘存储器。可选的,存储器还可以是至少一个位于远离前述处理器的存储装置。

[0126] 上述的处理器可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、网络处理器(Network Processor,简称NP)等;还可以是数字信号处理器(Digital Signal Processing,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)、现场可编程门阵列(Field-Programmable Gate Array,简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

[0127] 在本申请提供的又一实施例中,还提供了一种计算机可读存储介质,该计算机可读存储介质中存储有指令,当其在计算机上运行时,使得计算机执行上述实施例中任一所述的文档的排序方法。

[0128] 在本申请提供的又一实施例中,还提供了一种包含指令的计算机程序产品,当其在计算机上运行时,使得计算机执行上述实施例中任一所述的文档的排序方法。

[0129] 在上述实施例中,可以全部或部分地通过软件、硬件、固件或者其任意组合来实

现。当使用软件实现时,可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时,全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中,或者从一个计算机可读存储介质向另一个计算机可读存储介质传输,例如,所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质,(例如,软盘、硬盘、磁带)、光介质(例如,DVD)、或者半导体介质(例如固态硬盘 Solid State Disk(SSD))等。

[0130] 需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0131] 本说明书中的各个实施例均采用相关的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于系统实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0132] 以上所述仅为本申请的较佳实施例而已,并非用于限定本申请的保护范围。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,均包含在本申请的保护范围内。

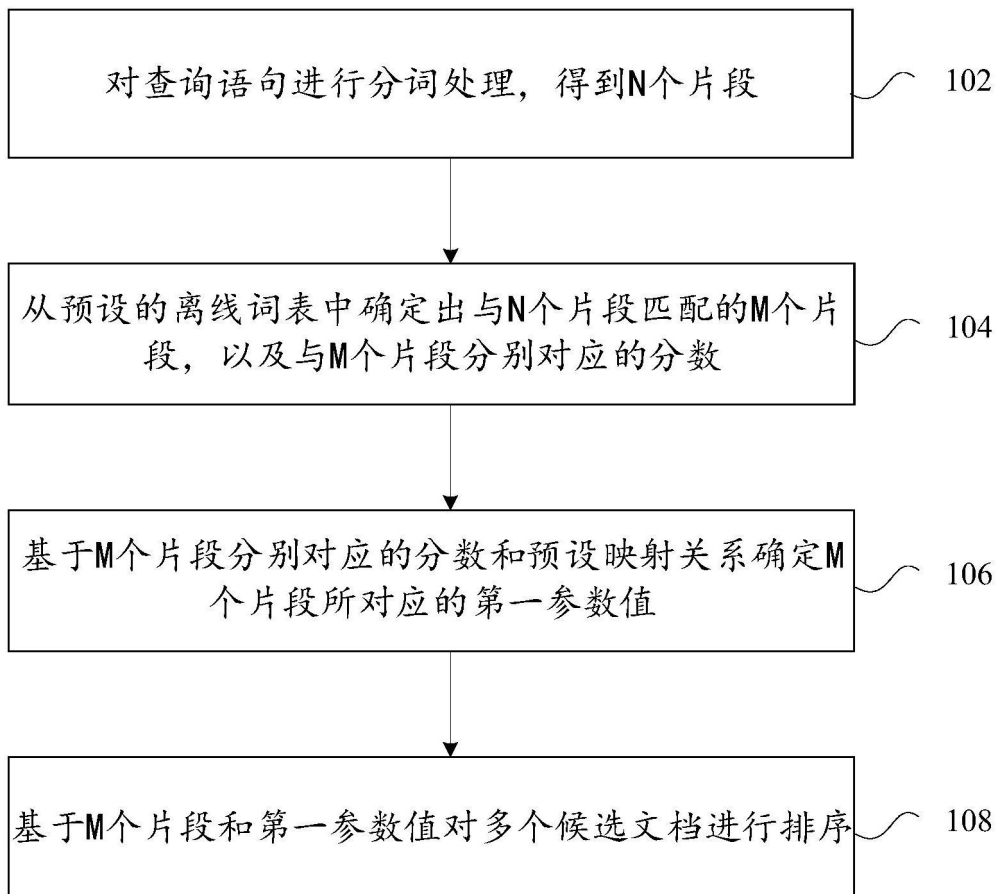


图1

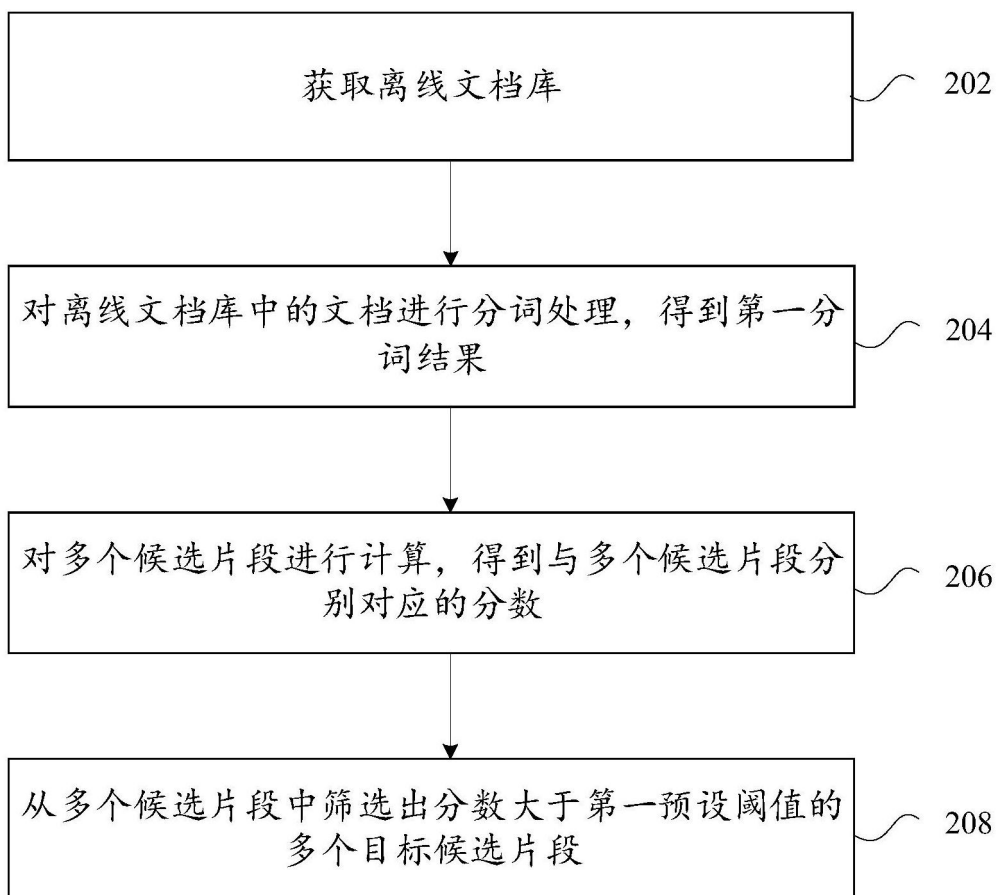


图2

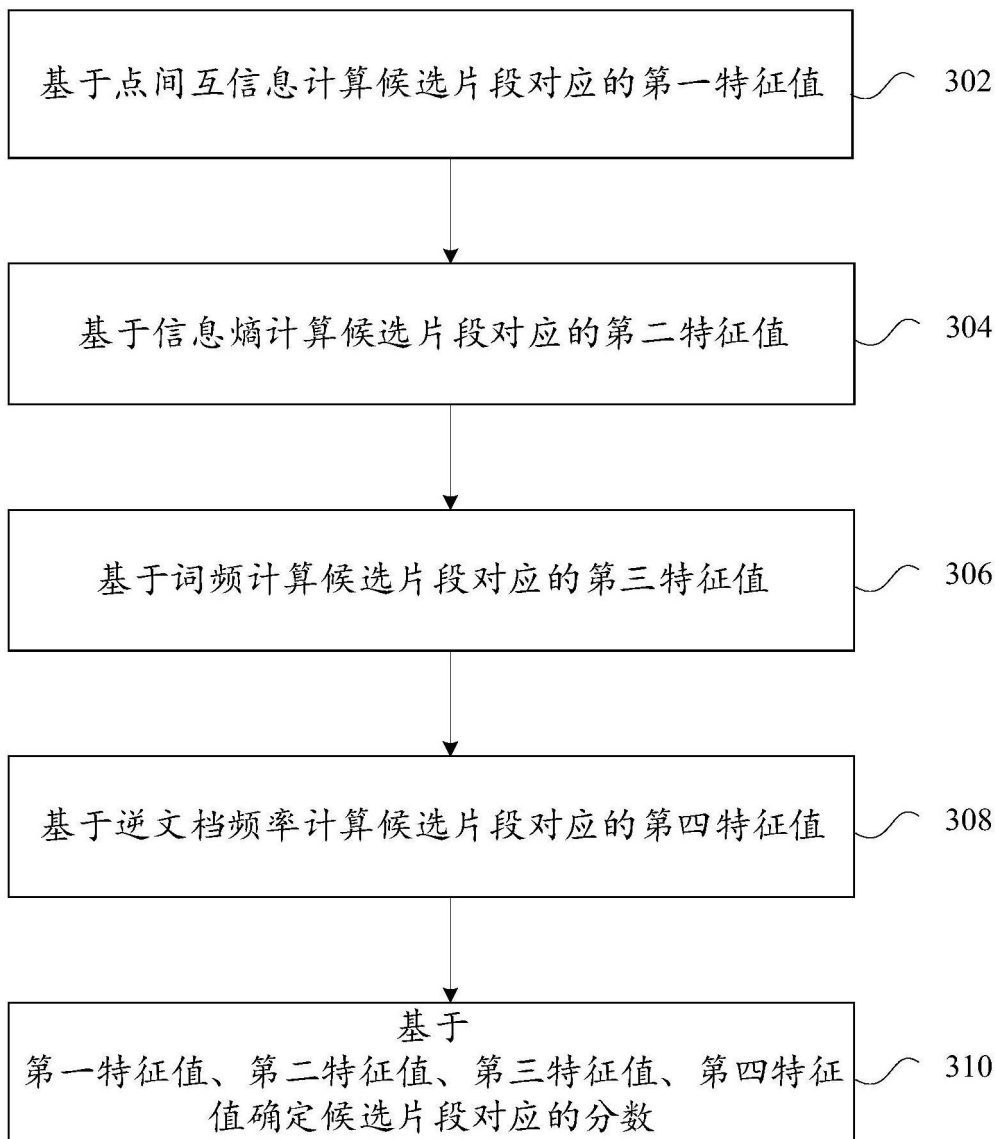


图3

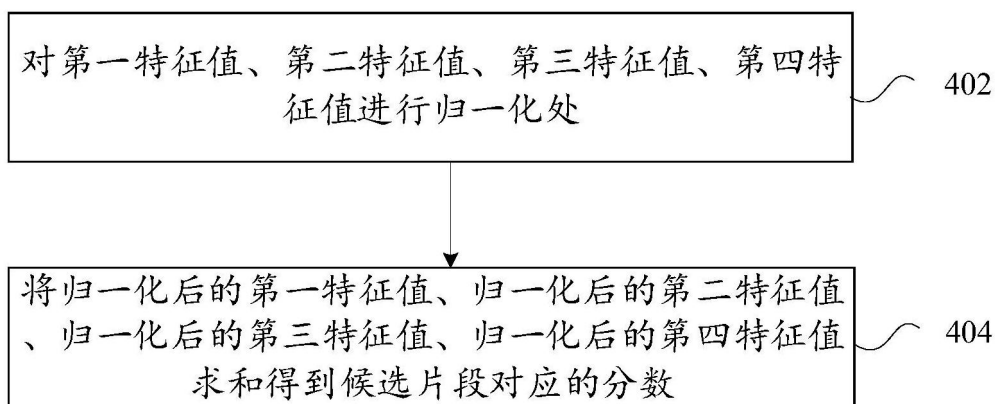


图4

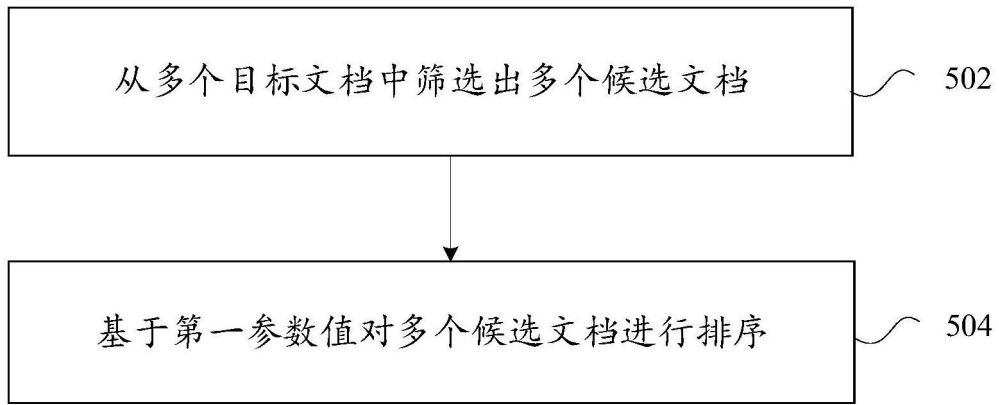


图5

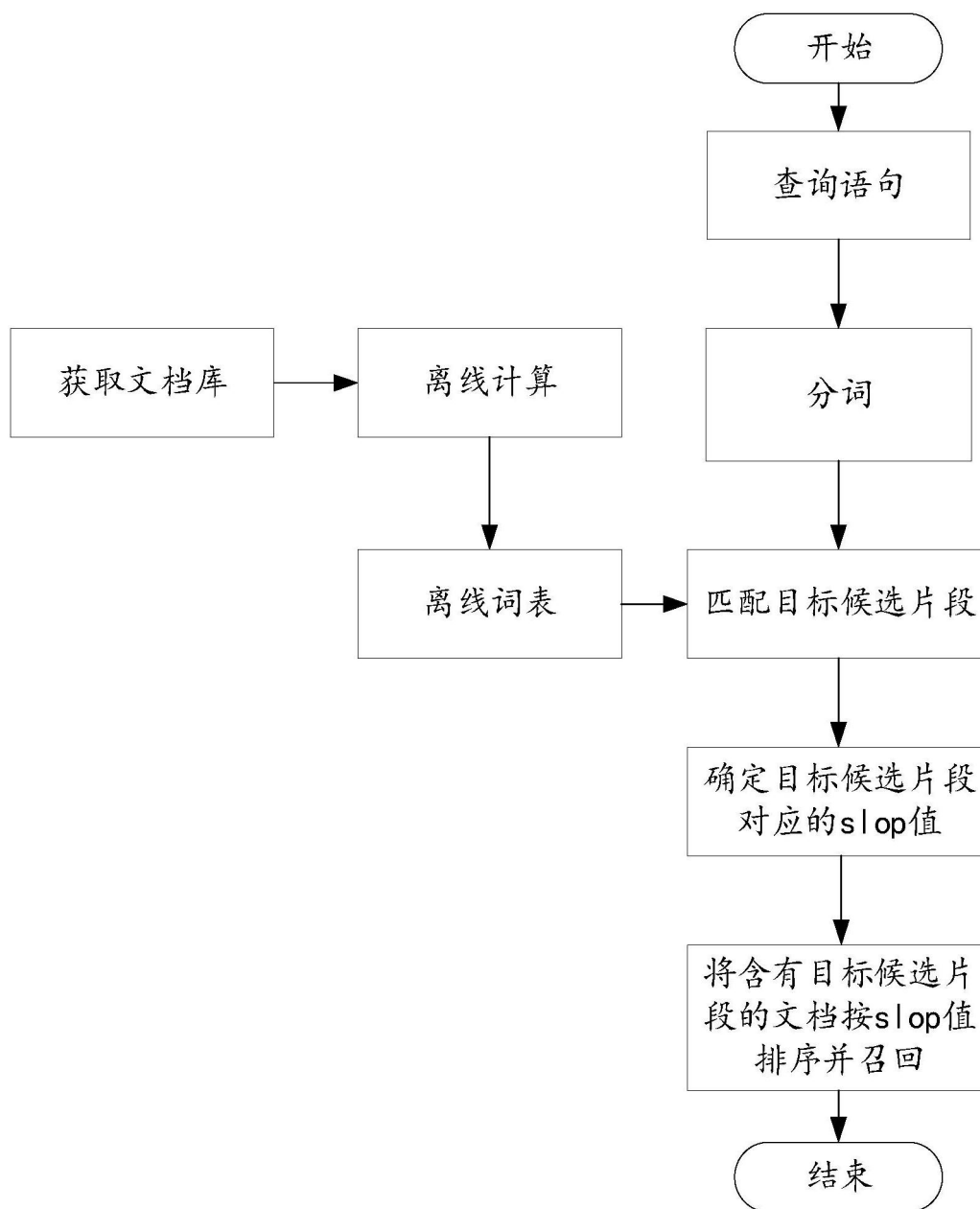


图6

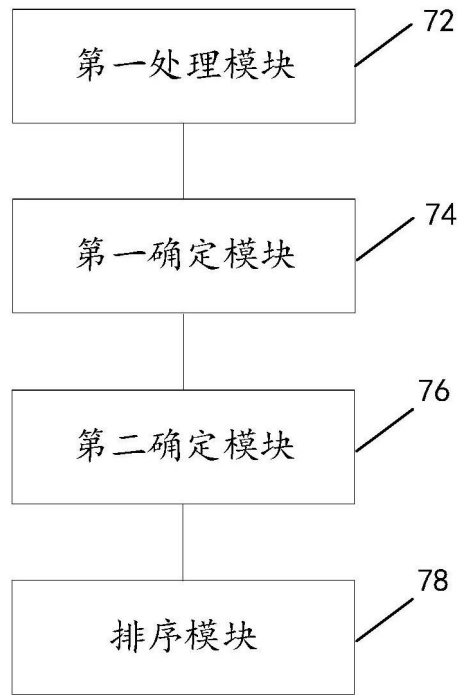


图7

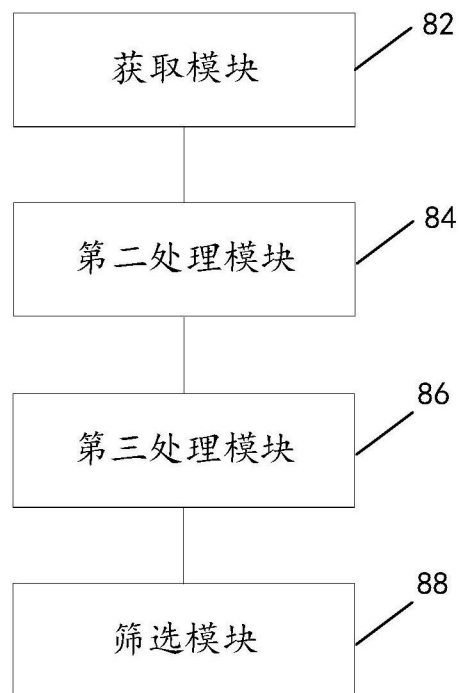


图8

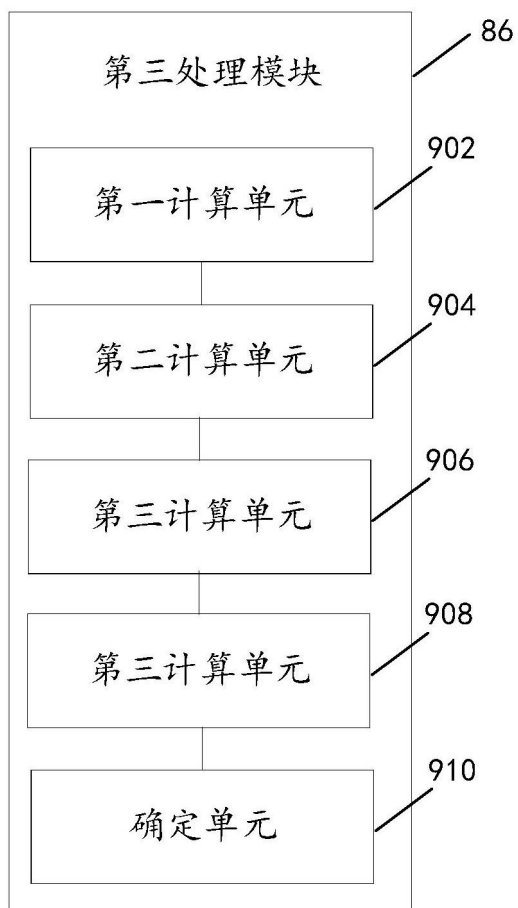


图9

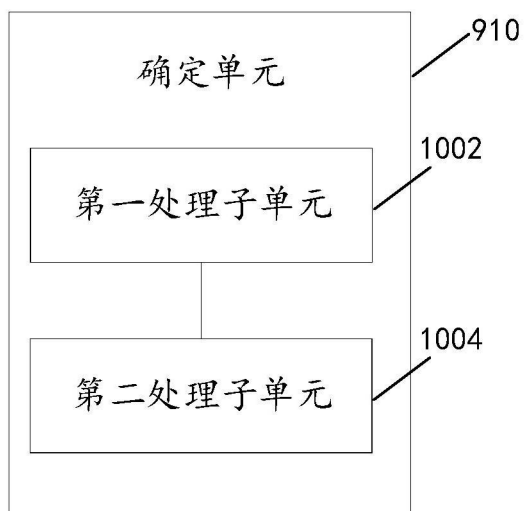


图10

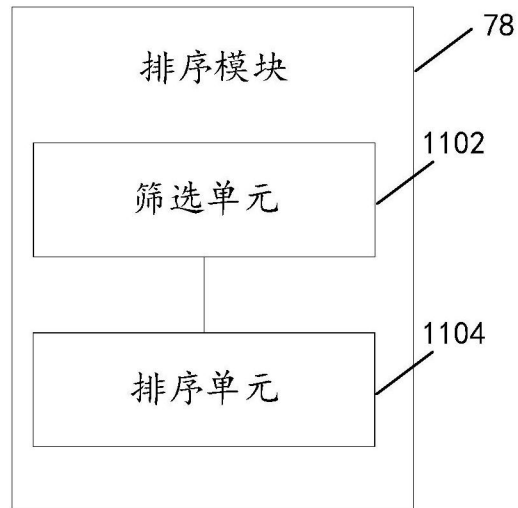


图11

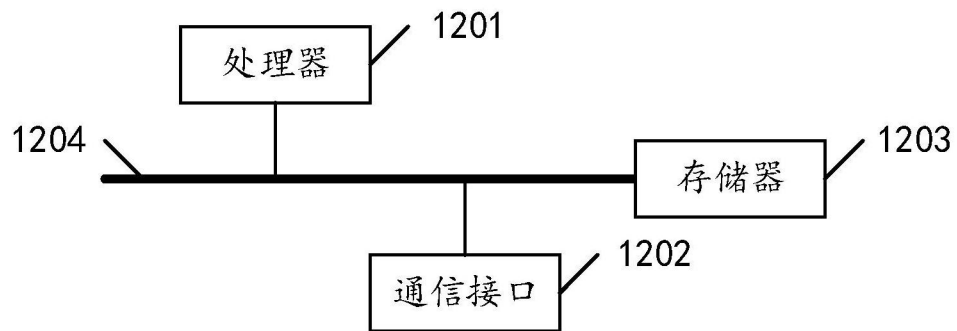


图12