(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110929125 B
(45)授权公告日 2023.07.11

(21)申请号 201911126486.2
(22)申请日 2019.11.15
(65)同一申请的已公布的文献号
申请公布号 CN 110929125 A
(43)申请公布日 2020.03.27
(73)专利权人 腾讯科技(深圳)有限公司
地址 518000 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 陈诚 冯帅 邓威 王军伟 方高林 郑楚涛 郑黄晓为
(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 贾允
(51)Int.Cl.
G06F 16/951(2019.01)
G06F 16/953(2019.01)

(56)对比文件
CN 103177075 A,2013.06.26
US 2016125003 A1,2016.05.05
CN 107943919 A,2018.04.20
WO 2018040503 A1,2018.03.08
US 2017220687 A1,2017.08.03
CN 107491518 A,2017.12.19
US 2010010970 A1,2010.01.14
US 2008306908 A1,2008.12.11
JP 2003256472 A,2003.09.12
US 2015081654 A1,2015.03.19
CN 104715065 A,2015.06.17
US 2016041986 A1,2016.02.11
US 5544049 A,1996.08.06

审查员 谢晓琦

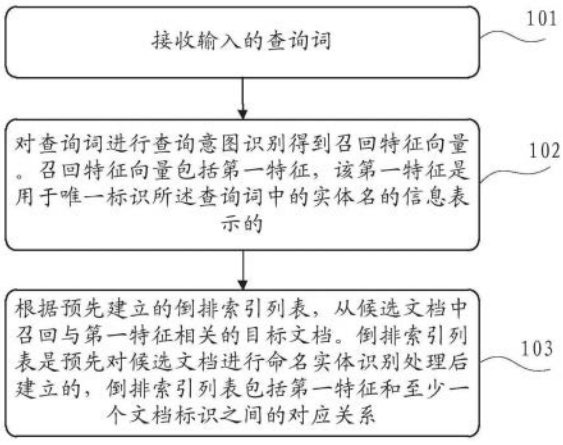
权利要求书2页 说明书11页 附图7页

(54)发明名称

搜索召回方法、装置、设备及其存储介质

(57)摘要

本申请公开了一种搜索召回方法、装置、设备及其存储介质。该方法包括:接收输入的查询词;对查询词进行查询意图识别得到召回特征向量,该召回特征向量包括第一特征,该第一特征是用于唯一标识查询词中的实体名的信息表示的;根据预先建立的倒排索引列表,从候选文档中召回与第一特征相关的目标文档,倒排索引列表是预先对候选文档进行命名实体识别处理后建立的,倒排索引列表包括第一特征和至少一个文档标识之间的对应关系。根据本申请实施例的技术方案,通过用于唯一标识查询词中的实体名的信息表示查询词中的实体名,基于这个唯一标识的信息查找预先建立的倒排索引表,有效地提高召回结果的准确性。



CN 110929125 B

1.一种搜索召回方法,其特征在于,其包括以下步骤:

接收输入的查询词;

对所述查询词进行查询意图识别得到召回特征向量,所述召回特征向量包括第一特征,所述第一特征是用于唯一标识所述查询词中的实体名的信息表示的;

根据预先建立的倒排索引列表,从候选文档中召回与所述第一特征相关的目标文档,所述倒排索引列表是预先对所述候选文档进行命名实体识别处理后建立的,所述倒排索引列表包括第一特征和至少一个文档标识之间的对应关系;

其中,所述倒排索引列表还包括第二特征和至少一个文档标识之间的对应关系,所述第二特征是用于唯一标识所述查询词中的实体名的信息和否定成分表示的,所述否定成分表示用于唯一标识所述查询词中的实体名的信息为假,则所述方法还包括:

对所述查询词进行查询意图识别得到召回特征向量,所述召回特征向量包括所述第二特征;

根据预先建立的倒排索引列表,从候选文档中召回与所述第二特征相关的目标文档。

2.根据权利要求1所述的搜索召回方法,其特征在于,在对所述查询词进行查询意图识别得到召回特征向量之后,所述方法还包括:

获取所述召回特征向量包含所述第一特征的第一数值;

获取每篇文档包含所述第一特征的第二数值,所述每篇文档是从所述候选文档中查找到的与所述第一特征相关的文档;

则所述根据预先建立的倒排索引列表,从候选文档中召回与所述第一特征相关的目标文档,还包括以下步骤:

在所述第一数值小于等于所述第二数值时,召回与所述第一特征相关的文档作为所述目标文档。

3.根据权利要求1或2所述的搜索召回方法,其特征在于,所述用于唯一标识所述查询词中的实体名的信息为股票代码,则在从候选文档中召回与所述第一特征相关的目标文档之后,所述方法还包括:

将所述目标文档构成召回文档列表;

基于所述召回特征向量抽取用户查询特征向量;

从所述召回文档列表中抽取文档特征向量;

将所述用户查询特征向量、所述文档特征向量、排序特征输入到预先训练建立的重排模型,输出重排序后的目标文档,其中,排序特征是根据所述查询词中包含股票代码的个数与待选文档中包含股票代码的个数计算得到的。

4.根据权利要求1或2所述的搜索召回方法,其特征在于,所述对所述查询词进行查询意图识别得到召回特征向量包括以下步骤:

对所述查询词进行分词处理得到至少一个分词;

对每个所述分词进行改写处理;

对上述处理后的所述分词进行命名实体识别得到至少一个实体名,并确定每个所述实体名是由所述第一特征来表示,或者是由所述第二特征来表示。

5.根据权利要求1或2所述的搜索召回方法,其特征在于,所述预先建立的倒排索引列表包括以下步骤:

获取所述候选文档；

对所述候选文档的标题和正文进行分词和关键词抽取处理，得到至少一个分词和至少一个关键词；

对所述分词和所述关键词进行命名实体识别，得到至少一个实体名；

并确定每个所述实体名是由所述第一特征来表示，或者是由所述第二特征来表示。

6.根据权利要求1所述的搜索召回方法，其特征在于，该方法还包括：

将所述倒排索引列表存储至区块链网络。

7.一种搜索召回装置，其特征在于，其包括：

接收单元，用于接收输入的查询词；

识别单元，用于对所述查询词进行查询意图识别得到召回特征向量，所述召回特征向量包括第一特征，所述第一特征是用于唯一标识所述查询词中的实体名的信息表示的；

召回单元，用于根据预先建立的倒排索引列表，从候选文档中召回与所述第一特征相关的目标文档，所述倒排索引列表是预先对所述候选文档进行命名实体识别处理后建立的，所述倒排索引列表包括第一特征和至少一个文档标识之间的对应关系；

其中，所述倒排索引列表还包括第二特征和至少一个文档标识之间的对应关系，所述第二特征是用于唯一标识所述查询词中的实体名的信息和否定成分表示的，所述否定成分表示用于唯一标识所述查询词中的实体名的信息为假，则所述装置还包括：

识别单元，还用于对所述查询词进行查询意图识别得到召回特征向量，所述召回特征向量包括所述第二特征；

召回单元，还用于根据预先建立的倒排索引列表，从候选文档中召回与所述第二特征相关的目标文档。

8.一种计算机设备，包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1-6中任一项所述的方法。

9.一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求1-6中任一项所述的方法。

搜索召回方法、装置、设备及其存储介质

技术领域

[0001] 本申请涉及互联网技术领域，尤其涉及搜索召回方法、装置、设备及其存储介质。

背景技术

[0002] 新闻资讯搜索功能为用户提供获取资讯结果的快捷渠道。搜索引擎根据用户输入的查询词语，在网络中召回与查询词语相关的查询结果，然后对查询结果进行排序，将排序靠前的查询结果展示给用户。

[0003] 在搜索过程中，用户获得的结果虽然形式上是与查询词语相关联的，但是其实质内容确与用户查询目的不匹配。特别是，期望搜索与专业领域相关的查询结果时，基于查询词语获取的查询结果精准度不高。

发明内容

[0004] 鉴于现有技术中的上述缺陷或不足，期望提供一种搜索召回方法、装置、设备及其存储介质，在资讯搜索过程中通过唯一标识资讯目标的方式，提高召回结果的准确性。

[0005] 一方面，本申请实施例提供了一种搜索召回方法，其包括以下步骤：

[0006] 接收输入的查询词；

[0007] 对查询词进行查询意图识别得到召回特征向量，该召回特征向量包括第一特征，该第一特征是用于唯一标识查询词中的实体名的信息表示的；

[0008] 根据预先建立的倒排索引列表，从候选文档中召回与第一特征相关的目标文档，倒排索引列表是预先对候选文档进行命名实体识别处理后建立的，倒排索引列表包括第一特征和至少一个文档标识之间的对应关系。

[0009] 一方面，本申请实施例提供了一种搜索召回装置，其包括：

[0010] 接收单元，用于接收输入的查询词；

[0011] 识别单元，用于对查询词进行查询意图识别得到召回特征向量，该召回特征向量包括第一特征，该第一特征是用于唯一标识查询词中的实体名的信息表示的；

[0012] 召回单元，用于根据预先建立的倒排索引列表，从候选文档中召回与第一特征相关的目标文档，倒排索引列表是预先对候选文档进行命名实体识别处理后建立的，倒排索引列表包括第一特征和至少一个文档标识之间的对应关系。

[0013] 一方面，本申请实施例提供了一种计算机设备，包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，该处理器执行该程序时实现如本申请实施例描述的方法。

[0014] 一方面，本申请实施例提供了一种计算机可读存储介质，其上存储有计算机程序，该计算机程序用于：

[0015] 该计算机程序被处理器执行时实现如本申请实施例描述的方法。

[0016] 本申请实施例提供的搜索召回方法、装置及其设备和存储介质，通过对接收的查询词，进行查询意图识别，构建查询词中包含的实体名的统一标注，即通过用于唯一标识查

询词中的实体名的信息表示查询词中的实体名，基于这个唯一标识的信息查找预先建立的倒排索引表，该倒排索引列表也是预先基于命名实体识别处理后建立的，通过这种统一标注方式，有效地提高召回结果的准确性。

[0017] 可选地，在排序阶段引用上述统一标注的排序特征，通过该排序特征可以将召回结果进行优化排序后提供给用户，提高了展示效率。

附图说明

[0018] 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，本申请的其它特征、目的和优点将会变得更明显：

[0019] 图1示出了本申请实施例提供的搜索召回方法所涉及的实施环境的结构示意图；

[0020] 图2示出了本申请实施例提供的搜索召回方法的流程示意图；

[0021] 图3示出了本申请实施例提供的搜索召回方法的流程示意图；

[0022] 图4示出了本申请实施例提供的一搜索召回方法的流程示意图；

[0023] 图5示出了本申请实施例提供的倒排索引列表的数据结构示意图；

[0024] 图6示出了本申请实施例提供的搜索召回装置500的结构示意图；

[0025] 图7示出了根据本申请一实施例提供的搜索召回装置600的示例性结构框图；

[0026] 图8示出了本申请实施例提供的搜索召回方法的完整流程示意图；

[0027] 图9示出了适于用来实现本申请实施例的计算机设备的计算机系统的结构示意图。

具体实施方式

[0028] 下面结合附图和实施例对本申请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公开，而非对该公开的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与公开相关的部分。

[0029] 需要说明的是，在不冲突的情况下，本申请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申请。

[0030] 下面先对本申请实施例提供的搜索召回方法所涉及的实施环境进行介绍。请参考图1，图1示出了本申请实施例提供的搜索召回方法所涉及的实施环境的结构示意图。如图1所示，该实施环境中包括终端11以及服务器12。其中，终端11的类型包括但不限于智能手机、台式电脑、笔记本电脑、平板电脑、可穿戴设备、多媒体播放设备等，终端上可以安装有各种应用程序，如新闻资讯软件、股票资讯软件或者其他资讯软件等，本申请实施例对此不进行具体限定。

[0031] 在本申请实施例中，终端11用于获取用户输入的查询词，并将获取到的查询词以网络请求的方式发送给服务器12，而服务器12用于根据终端11发送的查询词，返回与查询词相关的结果给终端11，进而由终端11将结果展示给用户。服务器可以是一台独立的服务器、或由若干台服务器组成的服务器集群、或云计算中心。服务器能够为终端提供查询处理服务。服务器可以是应用程序的后台服务器，例如：服务器可以是中间服务器，终端可以通过应用程序与服务器进行交互，从而实现查询处理流程。终端可通过有线或无线方式与服务器进行交互，从而实现查询处理流程。

[0032] 本申请实施例提供的搜索召回方法，可以由搜索召回装置作为执行主体来执行。搜索召回装置可以集成在终端或服务器等计算机设备中，搜索召回装置可以是硬件也可以是软件模块。也可以由单一的终端或服务器执行，或者二者配合起来执行。

[0033] 请参考图2，图2示出了本申请实施例提供的搜索召回方法的流程示意图。该方法可以由搜索召回装置执行。

[0034] 步骤101，接收输入的查询词；

[0035] 接收输入的查询词即query过程。查询词是指用户在搜索界面的输入区域输入的词语，即query的内容。查询词可以是词语、句子、数字，英文字母，也可以上述各种形式的组合。

[0036] 步骤102，对查询词进行查询意图识别得到召回特征向量。召回特征向量包括第一特征，该第一特征是用于唯一标识所述查询词中的实体名的信息表示的。

[0037] 在本步骤中，查询意图识别，也可以称为用户意图识别，其可以理解用户搜索意图。

[0038] 对搜索意图的理解，可以结合用户历史行为对查询词进行各种分析处理，例如，在查询词进行分词处理后，再进行改写处理。改写处理可以理解包括对查询词进行纠错处理、扩展处理等。纠错处理，例如可以对查询词进行繁写体到简写体的改写，对全角符号和半角符号进行识别改写，对英文字符的大小写进行统一，对查询词进行去除标点和末尾语气词。还可以基于拼音纠错和基于字形纠错，基于搜索日志的会话分析进行纠错结果的调整。扩展处理，是将与用户查询词相近、相关的词语进行扩展。优选地，可以基于实体名的标识属性进行扩展，标识属性例如可以是实体名的股票代码、证券名称、英文名称、拼音缩写、公司简称、公司全称等。还可以进一步包括上市公司的董事长、创始人、首席执行官等。通过多个维度来识别股票实体。

[0039] 查询意图识别还可以包括基于用户输入的查询词对搜索意图进行分类。对搜索意图进行分类，可以明确用户是搜索方向。例如，可以识别出用户是意图了解与查询词相关的资讯，还是用户想获取与查询词相关的需求。查询意图识别是期望对用户输入的查询词进行理解而获得最相关的信息。查询意图识别可以通过命名实体识别算法(也称为实体识别、实体分块和实体提取)对信息进行提取，旨在将文本中的命名实体定位并分类为预先定义的类别，如股票代码、证券名称、英文名称、拼音缩写、公司简称、公司全称等。也可以通过条件随机场(conditional random field algorithm，缩写CRF)算法、神经网络类算法、BERT(Bidirectional Encoder Representations from Transformers的缩写)及其改进算法来实现命名实体的识别。

[0040] 召回特征向量是对用户输入的查询词进行改写处理之后，根据命名实体规则识别得到的至少一个实体名，然后对每个实体名进行标注后的结果，使得多个实体名可以统一标注为唯一的表达方式。

[0041] 召回特征向量还可以包括排序特征，排序特征可以由改写处理后的特征之间的相关性来表征。召回特征向量可以包括第一特征或者第二特征。第一特征是由用于唯一标识实体名的信息来标注的，可以表示实体名是股票实体。第二特征是由用于唯一标识实体名的信息和否定成分来标注的，可以表示实体名不是股票实体。例如，用于唯一标识实体名的信息，例如可以是股票代码，或者是对股票代码加密映射后的信息，或者是利用股票代码与

股票名称和股票实体的法人信息共同生成一个标识信息，该标识信息也可以作为用于唯一标识实体名的信息。即第一特征可以用股票代码来标注，第二特征可以用股票代码和否定成分来标注。

[0042] 对查询词进行查询意图识别得到召回特征向量包括以下步骤：

[0043] 对查询词进行分词处理得到至少一个分词；

[0044] 对每个分词进行改写处理；

[0045] 对上述处理后的所述分词进行命名实体识别得到至少一个实体名，并确定每个实体名是由第一特征来表表示，或者是由第二特征来表示。

[0046] 步骤103，根据预先建立的倒排索引列表，从候选文档中召回与第一特征相关的目标文档。倒排索引列表是预先对候选文档进行命名实体识别处理后建立的，倒排索引列表包括第一特征和至少一个文档标识之间的对应关系。

[0047] 在本步骤中，倒排索引列表是一种预先建立的数据结构，该数据结构可以包括第一特征和至少一个文档标识之间的对应关系，还可以包括第二特征和至少一个文档标识之间的对应关系。如图5所示，图5示出了本申请实施例提供的倒排索引列表的数据结构示意图。其中401表示第一特征，402表示文档标识，403表示否定成分，其与401组合构成第二特征。401例如可以是：股票代码1，股票代码2，股票代码3，股票代码4；与股票代码1对应的402可以包括文档1，文档2，…，文档N，N为自然数。与股票代码2对应的402可以包括文档3，文档5，…，文档X，X为自然数。与股票代码3对应的402可以包括文档1，文档4，文档N，N为自然数。其中，403表示否定成分，其与股票代码4组合在一起，表示不是股票代码4对应的402可以包括文档2，文档3。403可以直接采用NO+股票代码4，或者10+股票代码4。

[0048] 候选文档是通过爬行和抓取技术获取的文档，或者是包含查询目标的其他文档，例如新闻资讯，公告研报，资料摘要等。

[0049] 在对用户输入的查询词进行查询意图识别后得到召回特征向量，召回特征向量中可以包括第一特征。根据第一特征，查找预先建立的倒排索引列表来得到目标文档。

[0050] 预先建立的倒排索引列表包括以下步骤：

[0051] 获取候选文档；

[0052] 对候选文档的标题和正文进行分词和关键词抽取处理，得到至少一个分词和至少一个关键词；

[0053] 对上述分词和关键词进行命名实体识别，得到至少一个实体名；

[0054] 并确定每个实体名是由第一特征来表示，或者是由第二特征来表示。

[0055] 下面以A公司的股票代码XXXXX.HK为例，A公司的全称为ABCD公司，A公司的中文拼音缩写为ABCD，A公司的证券名称为AB控股，A公司的英文名称为TT，A公司简称为AB。

[0056] 假设用户在金融资讯界面中输入查询词为AB控股，经过分词、改写处理后，查询词扩展为{XXXXX.HK，ABCD，TT，AB…}。对改写处理后的各个词进行查询意图识别，可以确定用户输入的AB控股是股票实体，则用XXXXX.HK标注AB控股，作为召回特征向量的第一特征。

[0057] 基于XXXXX.HK在倒排索引表中查找与XXXXX.HK关联的至少一个文档作为目标文档。

[0058] 本申请实施例，通过在搜索资讯过程中索引、召回阶段中建立查询词与用于唯一标识实体名的信息之间的关联关系，可以有效地提高召回的精确度。

[0059] 现有技术的金融证券场景，如果输入A公司的英文名称，可能只能召回包含英文名称的资讯结果，如果在部分文档中仅涉及A公司的中文名称或者股票代码等内容，则该部分文档可能会被漏召回。或者，按照现有的对英文名称的扩展处理后，可能在分词过程中将英文名称进行拆分，按照拆分后的单词进行召回，这时召回结果可能与用户输入的A公司的英文名称毫不相关，则属于错误召回的结果。有时候，甚至还会出现利用A公司的拼音字母缩写进行查询时，搜不到相关结果的情况。

[0060] 本申请实施例，通过使用统一的用于唯一标识实体名的信息作为特征值，可以高效地搜索到与该唯一标识实体名的信息的所有相关结果，并能够避免错误召回不相关的新闻资讯，还可以满足用户多维度输入查询的需求，提升召回的精准性和效率。

[0061] 针对用户查询对象为非股票实体的查询场景，本申请实施例还提供一种搜索召回方法，来提升召回的精准性。

[0062] 请参考图3，图3示出了本申请实施例提供的搜索召回方法的流程示意图。该方法可以由搜索召回装置执行。

[0063] 步骤201，接收输入的查询词；

[0064] 步骤202，对查询词进行查询意图识别得到召回特征向量，该召回特征向量包括第二特征。第二特征是用于唯一标识查询词中的实体名的信息和否定成分表示的，该否定成分表示用于唯一标识查询词中的实体名的信息为假。

[0065] 步骤203，根据预先建立的倒排索引列表，从候选文档中召回与第二特征相关的目标文档。该倒排索引列表包括第一特征和至少一个文档标识之间的对应关系，还包括第二特征和至少一个文档标识之间的对应关系。

[0066] 在上述步骤中，接收用户输入的查询词，对查询词进行查询意图识别可以得到一个或者多个实体名，对于每个实体名进行标注，实体名被标注表示为股票实体，或者被标注表示为非股票实体。非股票实体是指非上市公司。例如，用户输入的查询词为某种水果，按照现有技术召回的结果可能包括该水果和以该水果命名的公司。本申请则通过查询意图识别可以对用户查询的水果的查询意图进行理解识别，以确定用户期望查询的是水果的本意，还是水果命名的公司。如果是水果命名的公司，将水果用该水果公司的股票代码来标注，如果是水果本身含义，则用水果公司的股票代码和否定成分来标注。例如，用户输入：吃水果，对其进行分词后进行命名实体识别可以得到水果，水果两种含义，第一种含义是公司名称，第二种含义是水果本身。结合上下文理解识别出吃水果是第二种含义，则召回特征向量包括水果公司的股票代码和否定成分，表示吃水果中出现的“水果”不是股票代码等，将用户意图找到的对象精确地限定在非股票代码的文档区域范围内。

[0067] 以召回特征向量中股票代码和否定成分，查找倒排索引列表中不与水果对应的股票代码相关的文档作为召回结果。

[0068] 本申请实施例中，在召回特征向量中包括第二特征，并通过预先建立第二特征与文档标识之间的对应关系，查找与第二特征关联的文档，将文档的检索范围缩小在第二特征关联的文档范围内，其有效地提高了召回的精准性。

[0069] 为了更好地展示基于第一特征召回的结果，本申请还提供了一种搜索召回方法。请参考图4，图4示出了本申请实施例提供的一搜索召回方法的流将程示意图。该方法可以由搜索召回装置执行。

[0070] 步骤301，接收输入的查询词；

[0071] 步骤302，对查询词进行查询意图识别得到召回特征向量。召回特征向量包括第一特征，第一特征是用于唯一标识所述查询词中的实体名的信息表示的；

[0072] 步骤303，根据预先建立的倒排索引列表，在候选文档中查找与第一特征相关的文档；

[0073] 步骤304，获取召回特征向量包含第一特征的第一数值和每篇文档包含所述第一特征的第二数值，该每篇文档是从候选文档中查找到的与第一特征相关的文档。

[0074] 步骤305，在第一数值小于等于第二数值时，召回与第一特征相关的文档作为目标文档。

[0075] 在获取目标文档之后，该方法还可以包括：

[0076] 步骤306，将召回的与第一特征相关的目标文档构成召回文档列表；

[0077] 步骤307，基于召回特征向量获取用户查询特征向量；

[0078] 步骤308，从召回文档列表中抽取文档特征向量；

[0079] 步骤309，将用户查询特征向量、文档特征向量、排序特征输入到预先训练建立的重排模型，输出重排序后的目标文档，其中，排序特征是根据查询词中包含股票代码的个数与待选文档中包含股票代码的个数计算得到的。

[0080] 在上述步骤中，接收查询词，对查询词进行查询意图识别得到召回特征向量，可以参见图2和图3相关内容的描述。

[0081] 在得到召回文档列表之前，处理器还可以获取召回特征向量中包含第一特征的第一数值，获取从候选文档中查找到的与第一特征相关的文档中每篇文档所包含第一特征的第二数值，基于第一数值和第二数值的比较结果，确定是否召回与第一特征相关的文档。在第一数值小于等于第二数值时，召回与第一特征相关的文档作为目标文档。例如，用户输入的查询词中包括A公司的股票代码，B公司的英文名称，则召回的文档至少要包括A公司的股票代码和B公司的股票代码。如果仅包括A公司的股票代码的文档，这类文档则不会被召回。

[0082] 上述实施例中，第一特征为股票代码时，基于召回特征向量中获取用户查询特征向量，用户查询特征向量用于表示用户查询词经过分词处理后，各个分词之间的相关性。召回特征向量所包含的是股票代码，或者包含的是股票代码和否定成分，会导致召回结果的范围不同。例如，召回特征向量所包含的是股票代码1，则按照股票代码1索引得到的召回文档列表至少包括文档1，文档2，文档N。基于召回特征向量中包含的股票代码的个数，例如查询词中仅包含股票代码1，假设召回文档列表中包含文档1，文档2，文档N。文档1中包含股票代码1，股票代码3，文档2包含股票代码1，文档N中包含股票代码1。此时，召回特征向量中所包含的股票代码的个数为1，每篇文档中所包含股票代码的个数可以是文档1为2，文档2为1，文档N为1。

[0083] 然后基于召回文档列表抽取文档特征向量，文档特征向量用于表示与股票代码关联的关键词与文档之间的相关性。在计算相关性之后，进一步计算排序特征。该排序特征是根据查询词中包含股票代码的个数与待选文档中包含股票代码的个数计算得到的。排序特征，可以按照如下公式计算得到：

[0084] $$\frac{\left|StockSet_{query}\right|+\alpha}{\left|StockSet_{query}\cup StockSet_{doc}\right|+\alpha}$$

[0085] 其中，$StockSet_{query}$表示查询词中包含股票代码的个数；

[0086] $StockSet_{doc}$表示文档中包含股票代码个数；

[0087] a表示一个小于1正数。

[0088] 根据上述公式计算得到排序特征，该特征可以影响召回文档列表中所有文档的排序结果，优先展示用户输入的查询词最相关的资讯结果。

[0089] 本申请实施例通过在排序阶段引入排序特征影响排序结果，从而优化排序结果，有效地提升了展示结果的精准性。

[0090] 优选地，本申请实施例还可以将倒排索引列表存储到区块链网络中。倒排索引列表包括第一特征和至少一个文档标识之间的对应关系，通常可以存储到磁盘的某个文件中，形成倒排文件。为了更好地共享数据，保持数据的一致性，优选地，可以将倒排索引列表存储至区块链网络中。

[0091] 区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain)，本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层。

[0092] 区块链底层平台可以包括用户管理、基础服务、智能合约以及运营监控等处理模块。其中，用户管理模块负责所有区块链参与者的身份信息管理，包括维护公私钥生成(账户管理)、密钥管理以及用户真实身份和区块链地址对应关系维护(权限管理)等，并且在授权的情况下，监管和审计某些真实身份的交易情况，提供风险控制的规则配置(风控审计)；基础服务模块部署在所有区块链节点设备上，用来验证业务请求的有效性，并对有效请求完成共识后记录到存储上，对于一个新的业务请求，基础服务先对接口适配解析和鉴权处理(接口适配)，然后通过共识算法将业务信息加密(共识管理)，在加密之后完整一致的传输至共享账本上(网络通信)，并进行记录存储；智能合约模块负责合约的注册发行以及合约触发和合约执行，开发人员可以通过某种编程语言定义合约逻辑，发布到区块链上(合约注册)，根据合约条款的逻辑，调用密钥或者其它的事件触发执行，完成合约逻辑，同时还提供对合约升级注销的功能；运营监控模块主要负责产品发布过程中的部署、配置的修改、合约设置、云适配以及产品运行中的实时状态的可视化输出，例如：告警、监控网络情况、监控节点设备健康状态等。

[0093] 平台产品服务层提供典型应用的基本能力和实现框架，开发人员可以基于这些基本能力，叠加业务的特性，完成业务逻辑的区块链实现。应用服务层提供基于区块链方案的应用服务给业务参与方进行使用。

[0094] 应当注意，尽管在上述附图中以特定顺序描述了本公开方法的操作，但是，这并非要求或者暗示必须按照该特定顺序来执行这些操作，或是必须执行全部所示的操作才能实现期望的结果。相反，流程图中描绘的步骤可以改变执行顺序。附加地或备选地，可以省略某些步骤，将多个步骤合并为一个步骤执行，和/或将一个步骤分解为多个步骤执行。

[0095] 上述方法步骤可以由与其对应的装置来执行，参考图6，图6示出了本申请实施例提供的搜索召回装置500的结构示意图。该装置500包括：

[0096] 接收单元501，用于接收输入的查询词；

[0097] 识别单元502，用于对查询词进行查询意图识别得到召回特征向量，召回特征向量包括第一特征，该第一特征是用于唯一标识查询词中的实体名的信息表示的；

[0098] 召回单元503，用于根据预先建立的倒排索引列表，从候选文档中召回与第一特征相关的目标文档，该倒排索引列表是预先对候选文档进行命名实体识别处理后建立的，该倒排索引列表包括第一特征和至少一个文档标识之间的对应关系。

[0099] 在上述实施例基础上，倒排索引列表还包括第二特征和至少一个文档标识之间的对应关系，该第二特征是用于唯一标识所述查询词中的实体名的信息和否定成分表示的，该否定成分表示用于唯一标识所述查询词中的实体名的信息为假，则识别单元502，还用于对查询词进行查询意图识别得到召回特征向量，召回特征向量包括所述第二特征；

[0100] 召回单元503，还用于根据预先建立的倒排索引列表，从候选文档中召回与所述第二特征相关的目标文档。

[0101] 召回单元503还可以包括：

[0102] 获取子单元，用于获取召回特征向量包含第一特征的第一数值；获取每篇文档包含所述第一特征的第二数值，该每篇文档是从候选文档中查找到的与所述第一特征相关的文档；

[0103] 召回子单元，用于在第一数值小于等于所述第二数值时，召回与第一特征相关的文档作为目标文档。

[0104] 其中，识别单元502还可以包括：

[0105] 分词子单元，用于对查询词进行分词处理得到至少一个分词；

[0106] 改写子单元，用于对每个分词进行改写处理；

[0107] 第一实体名识别子单元，用于对上述处理后的分词进行命名实体识别得到至少一个实体名，并确定每个实体名是由第一特征来表示，或者是由第二特征来表示。

[0108] 装置500还可以包括倒排索引建立单元504，用于预先建立的倒排索引列表，其可以包括：

[0109] 文档获取子单元，用于获取候选文档；

[0110] 第一抽取子单元，用于对候选文档的标题和正文进行分词和关键词抽取处理，得到至少一个分词和至少一个关键词；

[0111] 第二实体名识别子单元，用于对分词和关键词进行命名实体识别，得到至少一个实体名；并确定每个实体名是由第一特征来表示，或者是由第二特征来表示。

[0112] 在上述实施例基础上，参考图7，图7示出了根据本申请又一实施例提供的搜索召回装置600的示例性结构框图。用于唯一标识查询词中的实体名的信息为股票代码，在装置500的基础上，装置600还包括：

[0113] 列表构成单元505，用于将召回的与第一特征相关的目标文档构成召回文档列表；

[0114] 第一抽取单元506，用于基于召回特征向量抽取用户查询特征向量；

[0115] 第二抽取单元507，用于从召回文档列表中抽取文档特征向量；

[0116] 排序单元508，用于将用户查询特征向量、文档特征向量、排序特征输入到预先训

练建立的重排模型，输出重排序后的目标文档，其中，排序特征是根据查询词中包含股票代码的个数与待选文档中包含股票代码的个数计算得到的。

[0117] 应当理解，装置500-600中记载的诸单元或模块与参考图1-3描述的方法中的各个步骤相对应。由此，上文针对方法描述的操作和特征同样适用于装置500-600及其中包含的单元，在此不再赘述。装置500-600中的相应单元可以与电子设备中的单元相互配合以实现本申请实施例的方案。

[0118] 在上文详细描述中提及的若干模块或者单元，这种划分并非强制性的。实际上，根据本公开的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0119] 为了清楚地理解本申请，以股票代码唯一以标识实体名，从而可以搜索到与该股票代码相关的全部咨询。该方法可以应用在金融/证券产品的资讯搜索功能中，也可以应用在包含金融/证券/股市相关的新闻资讯阅读平台的搜索功能中，还可以应用在与上市公司相关的文档搜索场景中。请参考图8，图8示出了本申请实施例提供的搜索召回方法的完整流程示意图。该方法可以包括三个阶段。

[0120] 资讯索引阶段，该阶段主要从候选文档集合中，通过股票命名实体识别算法对候选文档集合中每一篇候选文档进行实体名识别，并识别实体名是否统一采用股票代码标注。其中，股票命名实体识别算法是指通过自然语言处理算法识别文档中的上市公司股票实体，其识别维度可以包括股票代码、证券名称、英文名称、拼音所系、公司简称、公司全称等。进一步地，在资讯索引阶段还可以提取其他特征，例如，基础特征和排序特征。其中基础特征是用于标识文档的基础信息，例如文章标题、文章标识、文章媒体源、文章类型、文章发布时间等。排序特征，例如Tittle2vec，文章质量等。其中排序特征用于影响最后查找结果的排序。

[0121] 股票实体名识别算法可以通过深度学习算法来实现，该算法主要包括识别处理和消歧处理。其中识别处理可以通过对每篇文档按照预先收集的具有标识股票功能的属性文本进行匹配，来发现潜在股票实体。消歧处理是对潜在股票实体按照上下文信息进行分词处理后，通过多个分类器算法进行分类处理。这里的分类器例如可以是多层感知机算法(MLP,Multilayer Perceptron)，xgboost算法(xgboost,Extreme Gradient Boosting)，BERT算法等。然后对多个分类器的分类结果进行投票，以最终确定实体名是否为股票实体。例如，第一查询词为，吃水果，第二查询词为用水果。假设第一查询词中“水果”，经过三个分类器分类处理后判定结果为【非股票、非股票、非股票】那候选的“水果”最终结果就是非股票；假设第二查询词，三个分类器分类处理后的判定结果为【非股票、股票、股票】，那候选的“水果”最终判定结果就是股票实体。

[0122] 将经过股票命名实体识别算法处理过的文档和股票代码之间建立关联关系，如图5所示，构建索引数据库。可以通过ES(ElasticSearch)索引数据库来实现，ElasticSearch数据库是一款分布式全文检索框架，其使用JSON(JavaScript Object Notation)格式存储数据，采用倒排索引。采用ES索引数据库可以大幅地提升数据搜索的速度，节省处理时间。对文本进行分词处理，记录单词，词频，文本标识等信息，搜索时基于内容(根据单词和词频词向量等来计算评分)来找文本标识。

[0123] 资讯召回阶段，接收查询词，对查询词按照与资讯索引阶段相同的股票命名实体识别算法进行处理，以识别出查询词中是否包含股票实体。在咨询召回阶段还可以提取其他特征，例如排序特征。

[0124] 在咨询召回阶段基于股票命名实体识别算法得到的召回特征，在索引数据库查找与召回特征相关的文档，得到召回列表。基于召回列表提取文章特征，基于召回特征提取用户查询特征，通过这些特征中影响排序的排序特征输入到重排模型，调整重排规则，按照重排规则和预先建立的重排模型得到召回列表排序后的结果。其中，重排模型可以是基于机器学习算法实现的，例如学习排序算法(LTR,Learning to rank)，梯度提升树(GBDT,Gradient Boosting Decision Tree)算法等。

[0125] 下面参考图9，图9示出了适于用来实现本申请实施例的计算机设备的计算机系统800的结构示意图。

[0126] 如图9所示，计算机系统800包括中央处理单元(CPU)801，其可以根据存储在只读存储器(ROM)802中的程序或者从存储部分808加载到随机访问存储器(RAM)803中的程序而执行各种适当的动作和处理。在RAM803中，还存储有系统800操作所需的各种程序和数据。CPU 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。

[0127] 以下部件连接至I/O接口805：包括键盘、鼠标等的输入部分806；包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分807；包括硬盘等的存储部分808；以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分809。通信部分509经由诸如因特网的网络执行通信处理。驱动器810也根据需要连接至I/O接口805。可拆卸介质811，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器810上，以便于从其上读出的计算机程序根据需要被安装入存储部分808。

[0128] 特别地，根据本公开的实施例，上文参考流程图图2-4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，其包括承载在机器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部分809从网络上被下载和安装，和/或从可拆卸介质811被安装。在该计算机程序被中央处理单元(CPU)801执行时，执行本申请的系统中限定的上述功能。

[0129] 需要说明的是，本公开所示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。而在本公开中，计算机可读的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。计算机可读的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：无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0130] 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，前述模块、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。

[0131] 描述于本申请实施例中所涉及到的单元或模块可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元或模块也可以设置在处理器中，例如，可以描述为：一种处理器包括接收单元、识别单元以及召回单元。其中，这些单元或模块的名称在某种情况下并不构成对该单元或模块本身的限定，例如，接收单元还可以被描述为“用于接收输入的查询词的单元”。

[0132] 作为另一方面，本申请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可以是上述实施例中描述的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的。上述计算机可读存储介质存储有一个或者多个程序，当上述前述程序被一个或者一个以上的处理器用来执行描述于本申请的搜索召回方法。

[0133] 以上描述仅为本申请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，本申请中所涉及的公开范围，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前述公开构思的情况下，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。例如上述特征与本申请中公开的(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。

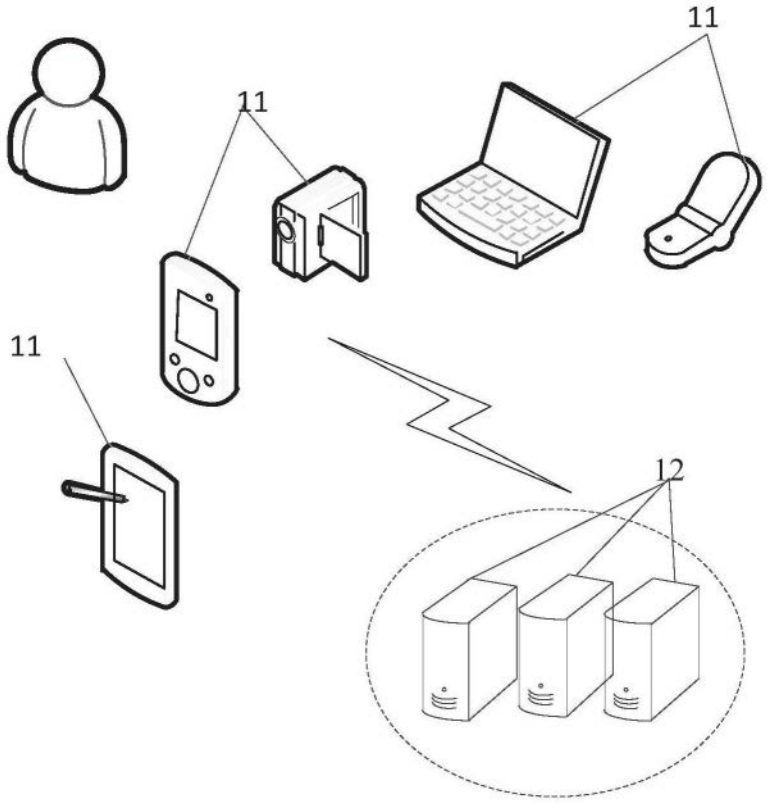

11
11
11
12

图1

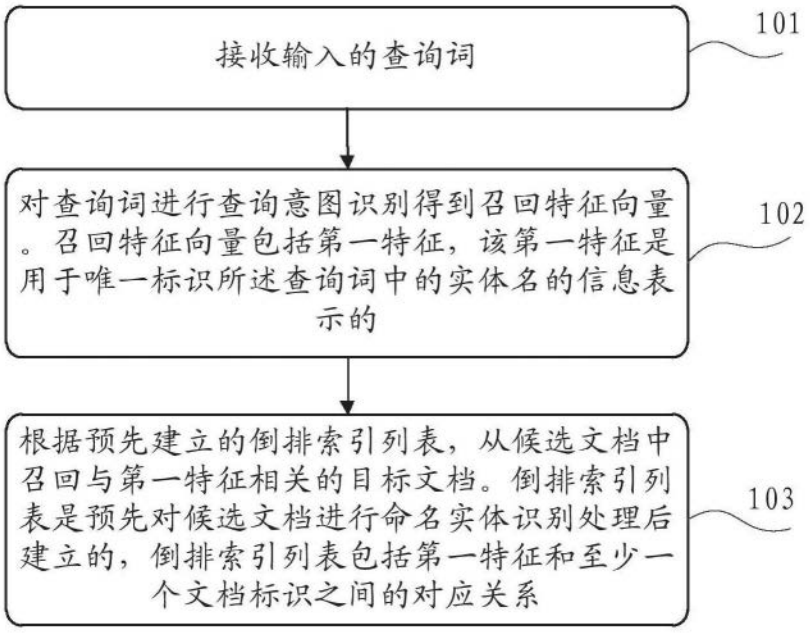

接收输入的查询词
101
对查询词进行查询意图识别得到召回特征向量。召回特征向量包括第一特征，该第一特征是用于唯一标识所述查询词中的实体名的信息表示的
102
根据预先建立的倒排索引列表，从候选文档中召回与第一特征相关的目标文档。倒排索引列表是预先对候选文档进行命名实体识别处理后建立的，倒排索引列表包括第一特征和至少一个文档标识之间的对应关系
103

图2

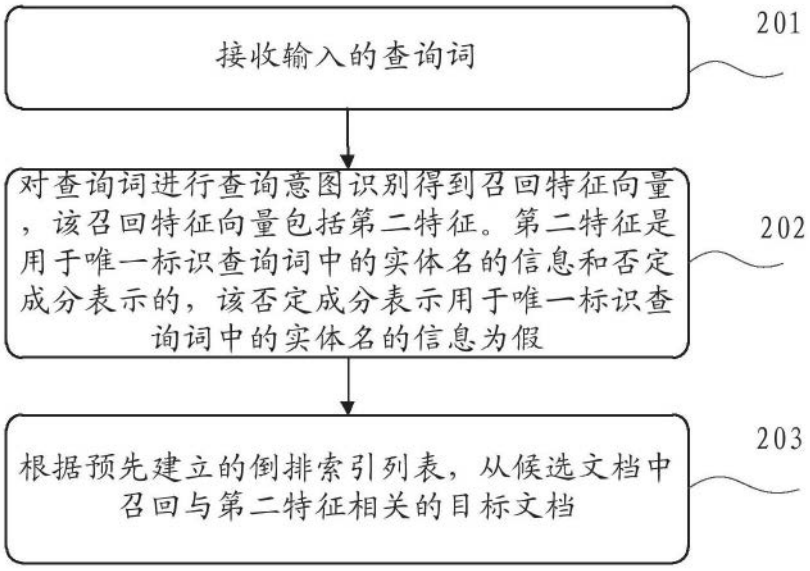

接收输入的查询词
201
对查询词进行查询意图识别得到召回特征向量，该召回特征向量包括第二特征。第二特征是用于唯一标识查询词中的实体名的信息和否定成分表示的，该否定成分表示用于唯一标识查询词中的实体名的信息为假
202
根据预先建立的倒排索引列表，从候选文档中召回与第二特征相关的目标文档
203

图3

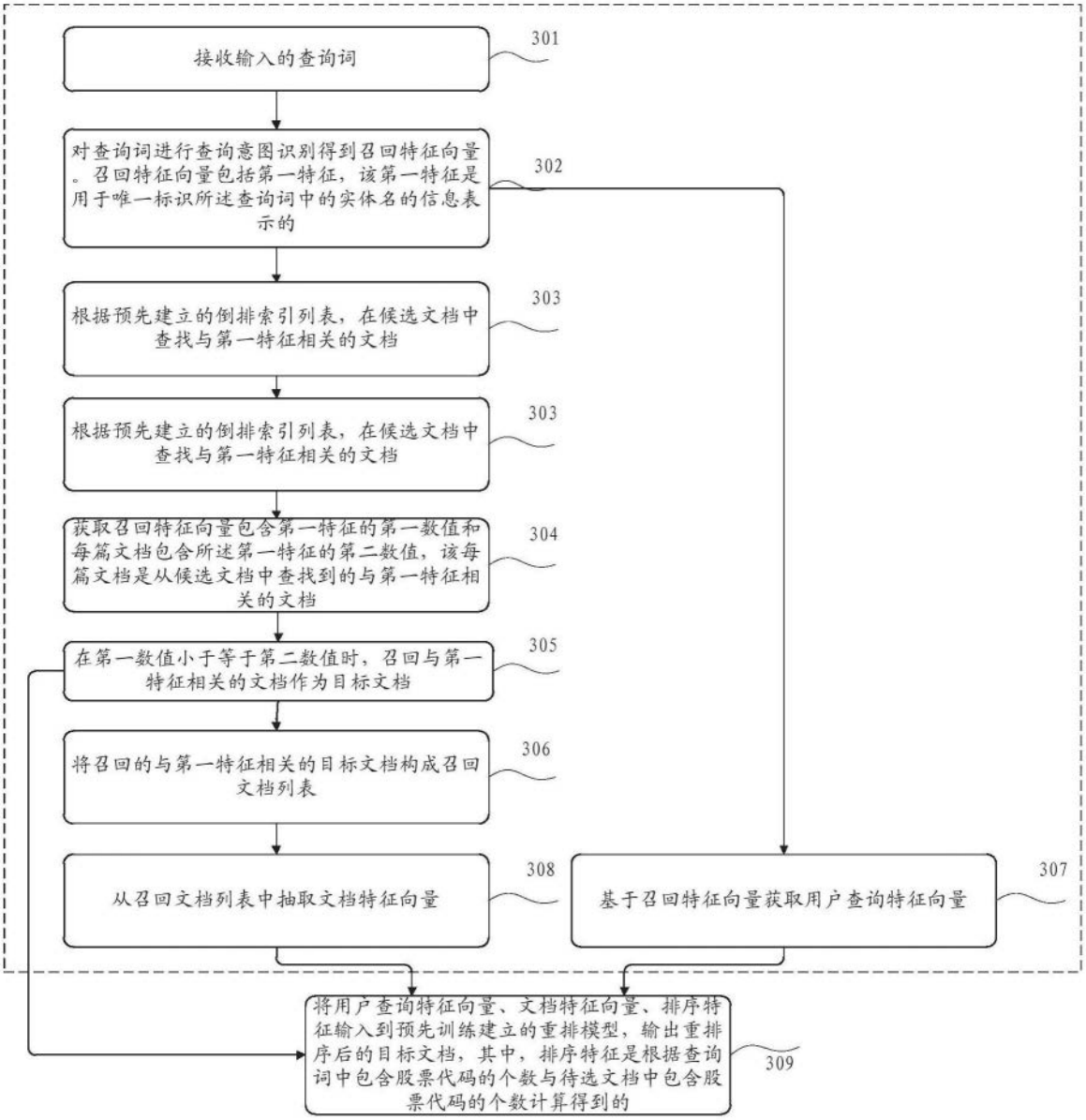

接收输入的查询词
301
对查询词进行查询意图识别得到召回特征向量。召回特征向量包括第一特征，该第一特征是用于唯一标识所述查询词中的实体名的信息表示的
302
根据预先建立的倒排索引列表，在候选文档中查找与第一特征相关的文档
303
根据预先建立的倒排索引列表，在候选文档中查找与第一特征相关的文档
303
获取召回特征向量包含第一特征的第一数值和每篇文档包含所述第一特征的第二数值，该每篇文档是从候选文档中查找到的与第一特征相关的文档
304
在第一数值小于等于第二数值时，召回与第一特征相关的文档作为目标文档
305
将召回的与第一特征相关的目标文档构成召回文档列表
306
从召回文档列表中抽取文档特征向量
308
基于召回特征向量获取用户查询特征向量
307
将用户查询特征向量、文档特征向量、排序特征输入到预先训练建立的重排模型，输出重排序后的目标文档，其中，排序特征是根据查询词中包含股票代码的个数与待选文档中包含股票代码的个数计算得到的
309

图4

标识 401	标识 402	标识 403
股票代码 1	文档 1，文档 2，…，文档 N	
股票代码 2	文档 3，文档 5，…，文档 X	
股票代码 3	文档 1，文档 4，文档 N	
股票代码 4	文档 2，文档 3	NO

图5

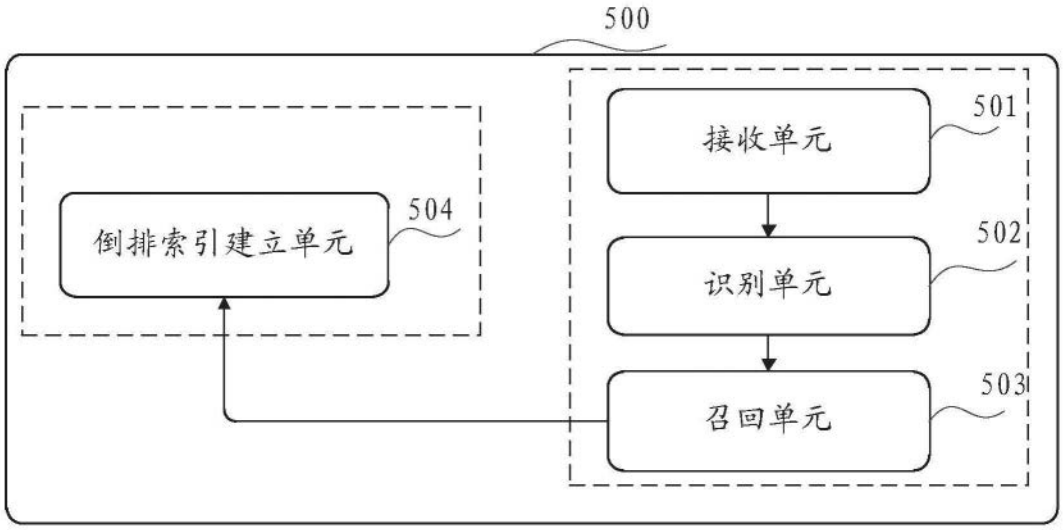


图6

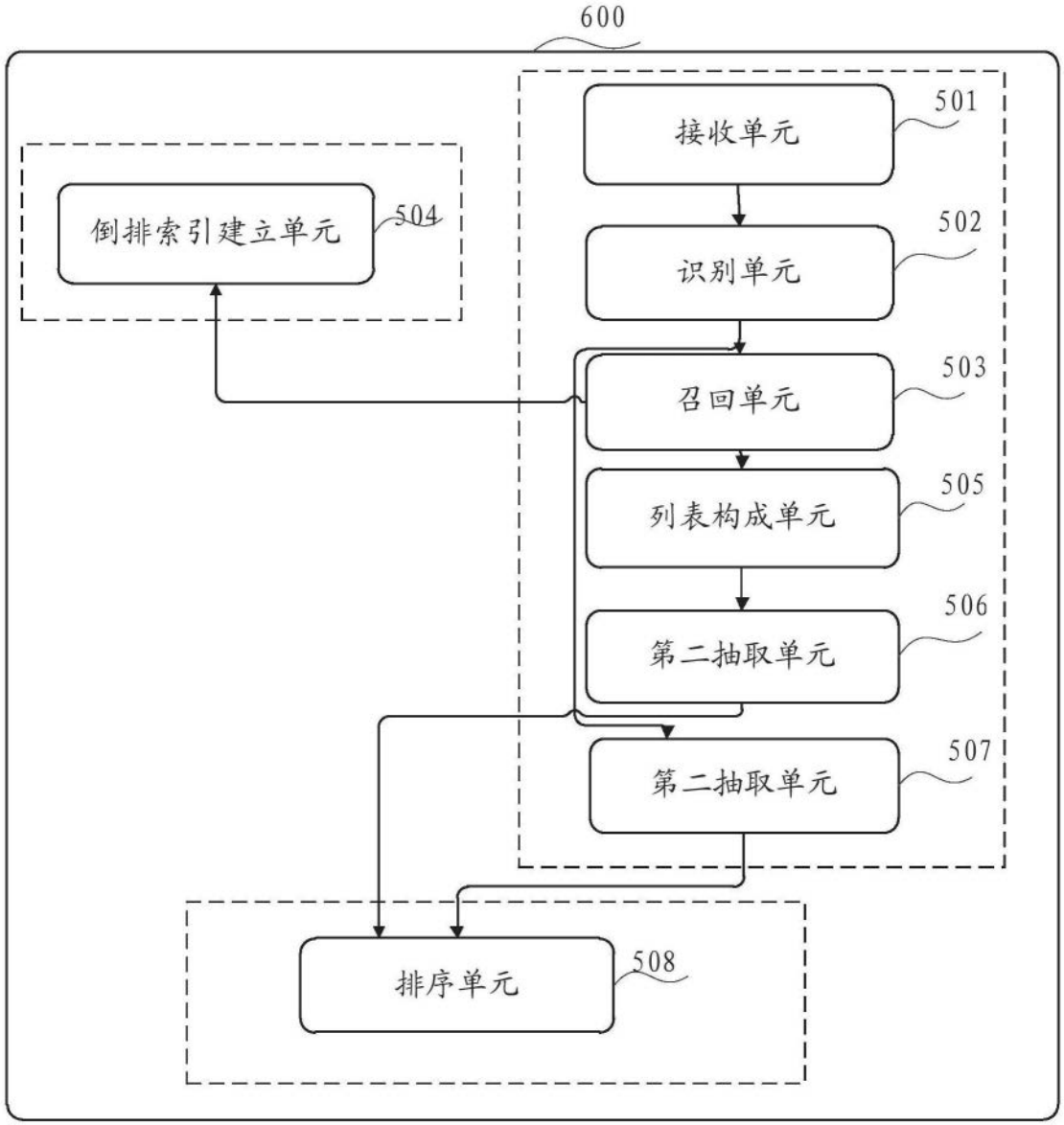

600
接收单元
501
识别单元
502
召回单元
503
列表构成单元
505
第二抽取单元
506
第二抽取单元
507
倒排索引建立单元
504
排序单元
508

图7

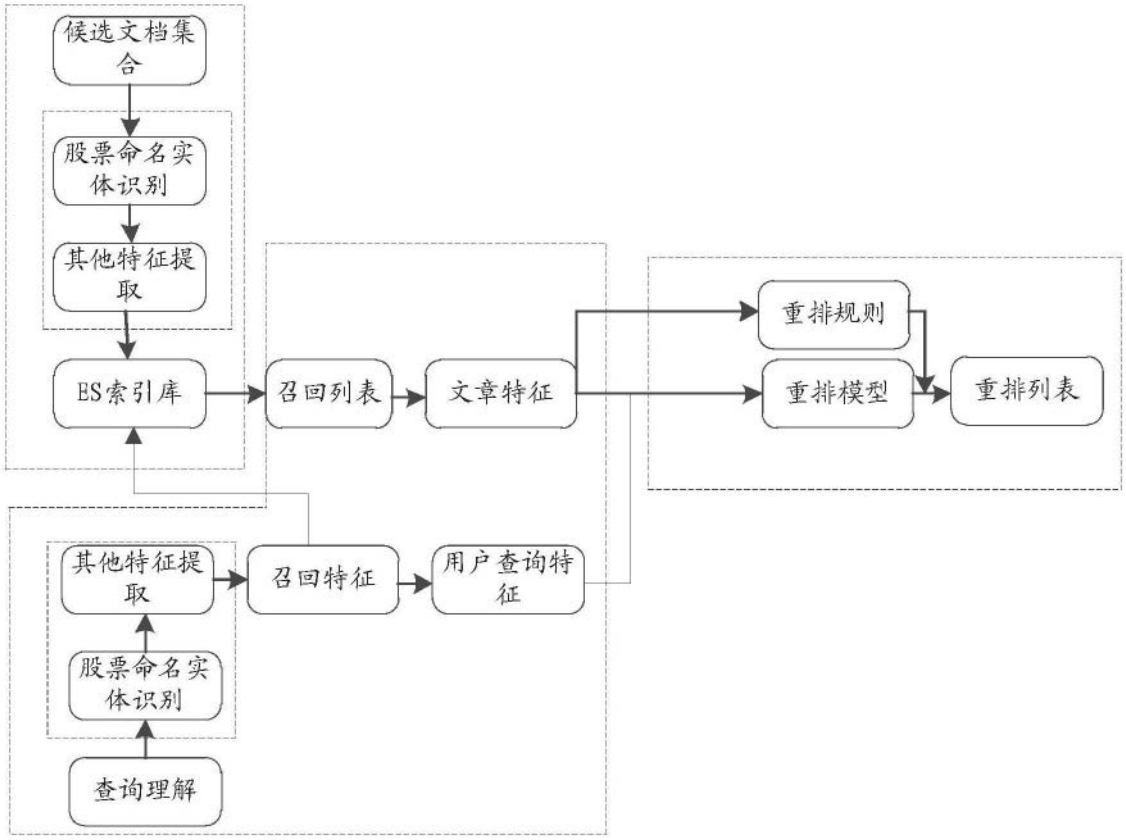

候选文档集合
股票命名实体识别
其他特征提取
ES索引库
召回列表
文章特征
重排规则
重排模型
重排列表
其他特征提取
召回特征
用户查询特征
股票命名实体识别
查询理解

图8

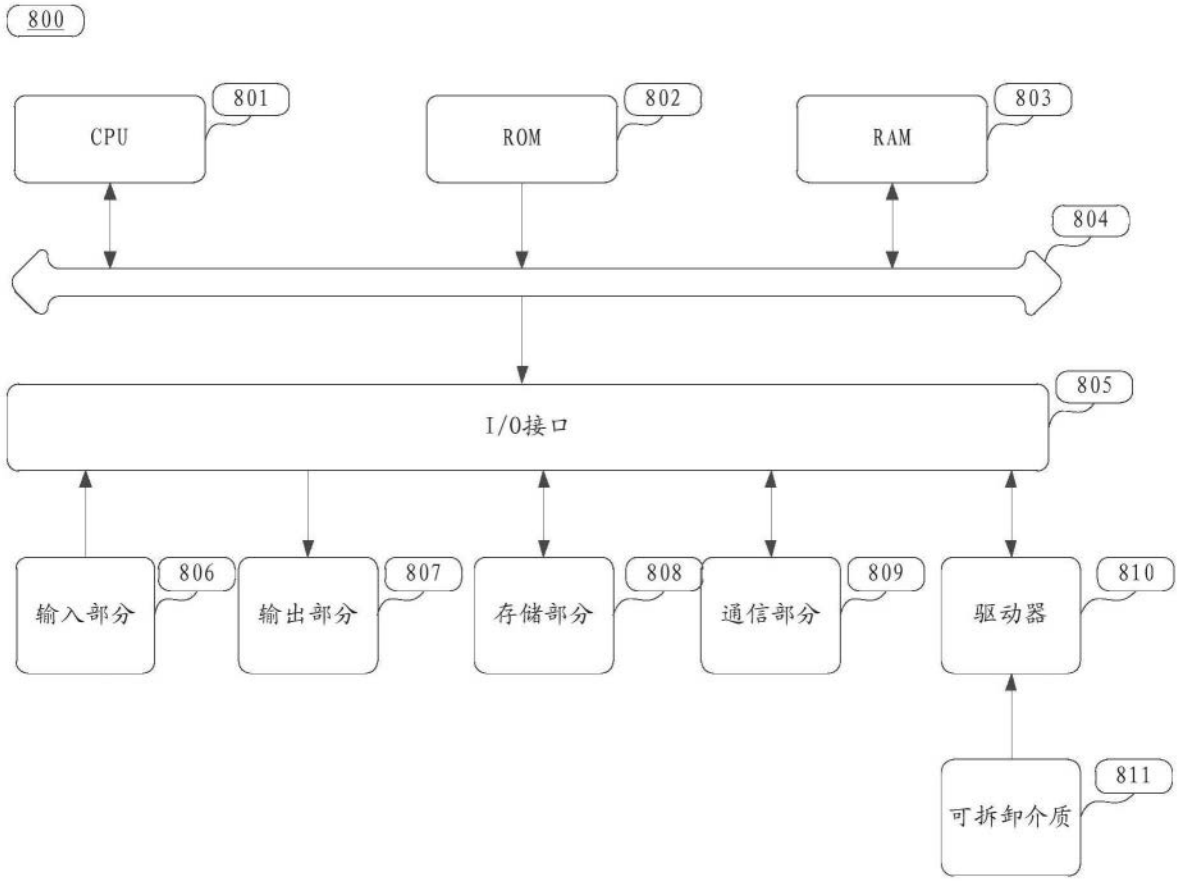

800
CPU
801
ROM
802
RAM
803
804
I/O接口
805
输入部分
806
输出部分
807
存储部分
808
通信部分
809
驱动器
810
可拆卸介质
811

图9