



(12) 发明专利申请

(10) 申请公布号 CN 104572612 A

(43) 申请公布日 2015. 04. 29

(21) 申请号 201310489328. X

(22) 申请日 2013. 10. 18

(71) 申请人 腾讯科技(深圳)有限公司

地址 518044 广东省深圳市福田区振兴路赛
格科技园 2 栋东 403 室

(72) 发明人 程刚

(74) 专利代理机构 北京德琦知识产权代理有限
公司 11018

代理人 杨春香 宋志强

(51) Int. Cl.

G06F 17/27(2006. 01)

G06F 17/30(2006. 01)

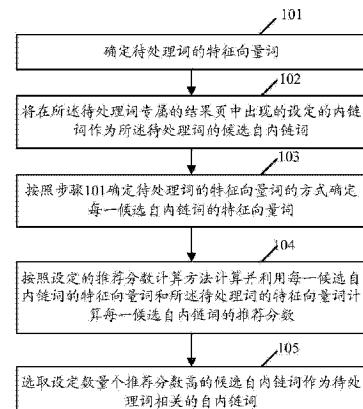
权利要求书4页 说明书11页 附图4页

(54) 发明名称

数据处理方法和装置

(57) 摘要

本申请提供了数据处理方法和装置。其中，该方法包括：确定待处理词的特征向量词；将在所述待处理词专属的结果页中出现的设定的内链词作为所述待处理词的候选自内链词；按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数；选取设定数量个推荐分数高的候选自内链词作为所述待处理词相关的自内链词。采用本发明，能够实现在处理某一词时，自动挖掘出该词的自内链词。



1. 一种数据处理方法,其特征在于,该方法包括:

确定待处理词的特征向量词;

将在所述待处理词专属的结果页中出现的设定的内链词作为所述待处理词的候选自内链词;

按照确定待处理词的特征向量词的方式确定每一候选自内链词的特征向量词;

按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数;

选取设定数量个推荐分数高的候选自内链词作为所述待处理词相关的自内链词。

2. 根据权利要求 1 所述的方法,其特征在于,所述确定待处理词的特征向量词包括:

确定所述待处理词专属结果页的文档;

确定设定阈值个与所述文档具有高相关度的词;

将确定的词确定为所述待处理词的特征向量词。

3. 根据权利要求 2 所述的方法,其特征在于,所述确定设定阈值个与所述文档具有高相关度的词包括:

对所述待处理词进行分词处理和去噪声干扰,得到对应的处理结果;

从所述处理结果中提取满足设定规定的词作为主题词;

计算每一主题词与所述文档的相关度;

选取设定阈值个与所述文档具有高相关度的主题词。

4. 根据权利要求 3 所述的方法,其特征在于,所述每一主题词与所述文档的相关度通过以下公式计算:

$$score(w, d) = IDF(w) * \frac{f(w, d) * (k_1 + 1)}{f(w, d) + k_1(1 - b + b \frac{|D|}{avgDL})}$$

其中, w 表示任一主题词, d 表示所述文档, score(w, d) 代表主题词 w 与所述文档之间的相关度, f(w, d) 表示主题词 w 在所述待处理词专属结果页的出现次数,D 为所述文档的文档长度, avgDL 为所述待处理词所属词类中所有词专属的所有结果页的文档平均长度, k₁、b

为计算相关度的设定参数, IDF(w) 通过以下公式确定: $IDF(w) = \log \frac{N - n(w) + 0.5}{n(w) + 0.5}$; N 表示

所述待处理词所属词类中所有词专属的所有结果页的总页数, n(w) 为所述待处理词所属词类里所有词专属的所有结果页中包含该主题词 w 的结果页数。

5. 根据权利要求 1 所述的方法,其特征在于,所述确定待处理词的特征向量词包括:

从所述待处理词专属的结果页中找到在预先设置的知识库中具有专属结果页的其他词;

从找到的词中选取设定阈值个词作定为所述待处理词的特征向量词。

6. 根据权利要求 1 所述的方法,其特征在于,所述按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数包括:

针对每一候选自内链词,按照设定的相关度计算方法计算该候选自内链词的所有特征向量词和所述待处理词的所有特征向量词之间的相关度,将计算出的相关度作为该候选自

内链词的推荐分数。

7. 根据权利要求 6 所述的方法, 其特征在于, 所述候选自内链词的所有特征向量词和所述待处理词的所有特征向量词之间的相关度通过以下公式计算 :

$$\text{recommend_score}(x, y) = \frac{\sum_{i=1}^n \sum_{j=1}^m \text{match}(v_i, w_j)}{\sqrt{\text{length}(v) * \text{length}(w)}};$$

其中, x 为所述待处理词, y 为任一候选自内链词, $\text{recommend_score}(x, y)$ 为所述待处理词 x 的所有特征向量词与任一候选自内链词 y 的所有特征向量词之间的相关度, n 为所述待处理词 x 的所有特征向量词的数量, m 为任一候选自内链词 y 的所有特征向量词的数量, $\text{length}(v)$ 为所述待处理词 x 的所有特征向量词的总长度, $\text{length}(w)$ 任一候选自内链词 y 的所有特征向量词的总长度, $\text{match}(v_i, w_j)$ 表示所述待处理词 x 的各个特征向量词与任一候选自内链词 y 的各个特征向量词之间的匹配度, 当 v_i 等于 w_j 时, $\text{match}(v_i, w_j)=1$, 当 v_i 不等于 w_j 时, $\text{match}(v_i, w_j)=0$ 。

8. 一种数据处理方法, 其特征在于, 该方法包括 :

将预先设置的知识库中除待处理词之外的其他词作为所述待处理词的候选自内链词 ;

获取每一候选自内链词在设定时间内被用户访问的次数 ;

计算所述知识库中所有词在所述设定时间内被用户访问的次数之和 ;

按照设定的推荐分数计算方法并利用每一候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算所述每一候选自内链词的推荐分数 ;

选取设定数量个推荐分数高的候选自内链词作为待处理词的相关的自内链词。

9. 根据权利要求 8 所述的方法, 其特征在于, 所述按照设定的推荐分数计算方法并利用该候选自内链词在设定时间内被用户访问的次数和计算的所述知识库中所有词在所述设定时间内被用户访问的次数之和计算该候选自内链词的推荐分数包括 :

针对每一候选自内链词, 按照设定的热度计算方法并利用该候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算该候选自内链词的热度, 将计算出的热度作为该候选自内链词的推荐分数。

10. 根据权利要求 9 所述的方法, 其特征在于, 所述候选自内链词的热度通过以下公式计算 :

$$\text{hot_score}(v) = \frac{\sum_{k=0}^n p(v)}{\sum_{k=0}^n p(u)}$$

其中, v 表示任一候选自内链词, $\text{hot_score}(v)$ 表示候选自内链词 v 的热度, $\sum_{k=0}^n p(v)$ 表

示候选自内链词 v 在设定时间内被用户访问的次数, $\sum_{k=0}^n p(u)$ 表示知识库中所有词在所述设定时间内被用户访问的次数之和。

11. 一种数据处理装置，其特征在于，该装置包括：

第一确定单元，用于确定待处理词的特征向量词；

第二确定单元，用于将在所述待处理词专属的结果页中出现的设定的内链词作为所述待处理词的候选自内链词；

第三确定单元，用于按照第一确定单元确定待处理词的特征向量词的方式确定每一候选自内链词的特征向量词；

计算单元，用于按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数；

选取单元，用于选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。

12. 根据权利要求 11 所述的装置，其特征在于，所述第一确定单元确定待处理词的特征向量词时先确定所述待处理词专属结果页的文档，确定设定阈值个与所述文档具有高相关度的词，将确定的词确定为所述待处理词的特征向量词。

13. 根据权利要求 12 所述的装置，其特征在于，所述第一确定单元在确定设定阈值个与所述文档具有高相关度的词时，对所述待处理词进行分词处理和去噪声干扰，得到对应的处理结果，从所述处理结果中提取满足设定规定的词作为主题词，计算每一主题词与所述文档的相关度，选取设定阈值个与所述文档具有高相关度的主题词作为所述待处理词的特征向量词。

14. 根据权利要求 11 所述的装置，其特征在于，所述第一确定单元在确定待处理词的特征向量词时从所述待处理词专属的结果页中找到在预先设置的知识库中具有专属结果页的其他词，从找到的词中选取设定阈值个词作定为所述待处理词的特征向量词。

15. 根据权利要求 11 所述的装置，其特征在于，所述计算单元在计算每一候选自内链词的推荐分数时，先针对每一候选自内链词，按照设定的相关度计算方法计算待处理词的所有特征向量词和该候选自内链词的所有特征向量词之间的相关度，将计算出的相关度作为该候选自内链词的推荐分数。

16. 一种数据处理装置，其特征在于，该装置包括：

确定单元，用于将预先设置的知识库中除待处理词之外的其他词作为所述待处理词的候选自内链词；

获取单元，用于获取每一候选自内链词在设定时间内被用户访问的次数；

第一计算单元，用于计算所述知识库中所有词在所述设定时间内被用户访问的次数之和；

第二计算单元，用于按照设定的推荐分数计算方法并利用每一候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算所述每一候选自内链词的推荐分数；

选取单元，用于选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。

17. 根据权利要求 16 所述的装置，其特征在于，所述第二计算单元在计算每一候选自内链词的推荐分数时，先针对每一候选自内链词，按照设定的热度计算方法并利用该候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用

户访问的次数之和计算该候选自内链词的热度,将计算出的热度作为该候选自内链词的推荐分数。

数据处理方法和装置

技术领域

[0001] 本申请涉及互联网技术,特别涉及数据处理方法和装置。

背景技术

[0002] 为使本申请容易理解,下面先对本申请涉及的技术术语进行描述:

[0003] 分词:是将一个序列切分成一个一个单独的词。该序列可以为中文汉字序列,也可以为中文汉字和专有英文词组成的序列。

[0004] 知识库:其是多个语义树的集合。而一个语义树是由语义相同或者相近的一组词的集合组成的。

[0005] 特征向量词:用来表示某一个文档的特征的词,其包括至少一个词。

[0006] 内链词:是在问答社区的正文中出现的,用户可以点击并且跳转到其它页面上的链接及描述文字。其可以作为一个文档的特征向量词。

[0007] 自内链词:属于内链词的一种,是知识库中某一类词条中用于指向同一类词条中其他词条的链接及描述文字。

[0008] 以上对本申请涉及的技术术语进行了描述。

[0009] 在现有技术中,当对知识库中的词(称为待处理词)进行一些数据处理时,如果能够自动推荐出该待处理词相关的自内链词,使用户从推荐的自内链词中找到自己感兴趣的词,无需用户主动重新获取,这一方面提高知识库的词访问效率,另一方面也能节省因为用户频繁访问知识库所浪费的资源。然而,现有技术中尚没有一种方式能够挖掘并推荐待处理词相关的自内链词。因此,一种用于挖掘待处理词相关的自内链词的数据处理方法是当前亟待解决的技术问题。

发明内容

[0010] 本申请提供了数据处理方法和装置,以实现在处理知识库中某一词时,自动挖掘出该词相关的自内链词。

[0011] 本申请提供的技术方案包括:

[0012] 一种数据处理方法,包括:

[0013] 确定待处理词的特征向量词;

[0014] 将在所述待处理词专属的结果页中出现的设定的内链词作为所述待处理词的候选自内链词;

[0015] 按照确定待处理词的特征向量词的方式确定每一候选自内链词的特征向量词;

[0016] 按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数;

[0017] 选取设定数量个推荐分数高的候选自内链词作为所述待处理词相关的自内链词。

[0018] 一种数据处理方法,该方法包括:

[0019] 将预先设置的知识库中除待处理词之外的其他词作为所述待处理词的候选自内

链词；

- [0020] 获取每一候选自内链词在设定时间内被用户访问的次数；
- [0021] 计算所述知识库中所有词在所述设定时间内被用户访问的次数之和；
- [0022] 按照设定的推荐分数计算方法并利用每一候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算所述每一候选自内链词的推荐分数；
- [0023] 选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。
- [0024] 一种数据处理装置，该装置包括：
 - [0025] 第一确定单元，用于确定待处理词的特征向量词；
 - [0026] 第二确定单元，用于将在所述待处理词专属的结果页中出现的设定的内链词作为所述待处理词的候选自内链词；
 - [0027] 第三确定单元，用于按照第一确定单元确定待处理词的特征向量词的方式确定每一候选自内链词的特征向量词；
 - [0028] 计算单元，用于按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数；
 - [0029] 选取单元，用于选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。
- [0030] 一种数据处理装置，该装置包括：
 - [0031] 确定单元，用于将预先设置的知识库中除待处理词之外的其他词作为所述待处理词的候选自内链词；
 - [0032] 获取单元，用于获取每一候选自内链词在设定时间内被用户访问的次数；
 - [0033] 第一计算单元，用于计算所述知识库中所有词在所述设定时间内被用户访问的次数之和；
 - [0034] 第二计算单元，用于按照设定的推荐分数计算方法并利用每一候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算所述每一候选自内链词的推荐分数；
 - [0035] 选取单元，用于选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。
- [0036] 由以上技术方案可以看出，本发明中，通过确定待处理词的特征向量词和候选自内链词，利用所述待处理词的特征向量词和每一候选自内链词的特征向量词计算每一候选自内链词的推荐分数，选取设定数量个推荐分数高的候选自内链词作为所述待处理词相关的自内链词，能够实现在处理某一词时，自动挖掘出该词的自内链词的目的。
- [0037] 进一步地，本发明中，由于在处理某一词时能够自动推荐出该词相关的自内链词，使用户从推荐的自内链词中找到自己感兴趣的词，无需用户主动重新获取，这一方面提高知识库的词访问效率，另一方面也能节省因为用户频繁访问知识库所浪费的资源。

附图说明

- [0038] 图 1 为本发明实施例 1 提供的方法流程图；
- [0039] 图 2 为本发明实施例 1 提供的特征向量词确定流程图；

- [0040] 图 3 为本发明实施例 2 提供的相关度确定流程图；
- [0041] 图 4 为本发明实施例 1 提供的特征向量词另一确定流程图；
- [0042] 图 5 为本发明实施例 2 提供的方法流程图；
- [0043] 图 6 为本发明实施例提供的装置结构图；
- [0044] 图 7 为本发明实施例提供的另一装置结构图。

具体实施方式

[0045] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0046] 本发明提供的方法能够在处理某一词时，能够自动挖掘出该词相关的自内链词，实现在处理某一词时，自动挖掘出该词的自内链词的目的。

- [0047] 下面通过两个实施例对本发明提供的方法进行描述：

- [0048] 实施例 1：

[0049] 参见图 1，图 1 为本发明实施例 1 提供的方法流程图。如图 1 所示，该方法包括以下步骤：

- [0050] 步骤 101，确定待处理词的特征向量词。

- [0051] 本发明中，所述待处理词可包括至少一个词。

- [0052] 下文重点描述了如何确定待处理词的特征向量词的方法，本步骤 101 暂不赘述。

[0053] 步骤 102，将在所述待处理词专属的结果页中出现的设定的内链词作为所述待处理词的候选自内链词。

[0054] 本发明中，待处理词为预先设置的知识库中的词，其中，在设置知识库时，本发明可针对知识库中的每一词都专门设定一个专属的结果页，用于解释或者描述该词。

[0055] 基于此，本步骤 102 中，就基于知识库的设置，从知识库中找到所述待处理词专属的结果页。其中，该结果页中可包括一些在知识库中有专属结果页的词，针对这些词，其在接收到用户触发比如点击时会自动跳转到其专属结果页，因此可称为内链词。当本步骤 102 发现所述待处理词专属的结果页中出现一些如前所述的内链词时，本步骤 102 就将该发现的内链词作为所述待处理词的候选自内链词，以便后续从所述待处理词的候选自内链词中挖掘出优先级比较高的词作为待处理词相关的自内链词并推荐给用户。

[0056] 步骤 103，按照步骤 101 确定待处理词的特征向量词的方式确定每一候选自内链词的特征向量词。

[0057] 步骤 104，按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数。

[0058] 优选地，在上述步骤 103 中，之所以按照相同方式确定候选自内链词与待处理词的特征向量词，目的是方便本步骤 104 计算推荐分数，避免因为不同方式确定的特征向量词无法进行推荐分数计算。

[0059] 另外，至于本步骤 104 中设定的推荐分数计算方法，其可根据实际情况设置，比如，可设置为相关度计算方法，或者其他方式，本发明并不具体限定。

[0060] 步骤 105，选取设定数量个推荐分数高的候选自内链词作为待处理词相关的自内链词。

- [0061] 至此,通过上述步骤 101 至步骤 105 即可自动挖掘出待处理词相关的自内链词。
- [0062] 下面对图 1 所示流程中步骤 101 确定待处理词的特征向量词的方式进行描述:
- [0063] 优选地,本发明中可采用以下两种方式确定待处理词的特征向量词:
- [0064] 方式 1:
- [0065] 本方式 1 下,步骤 101 确定待处理词的特征向量词的方法可包括图 2 所示的以下步骤:
- [0066] 步骤 201,确定所述待处理词专属结果页的文档。
- [0067] 基于上文描述的,在知识库中的每一词都有一个专属的结果页,所述待处理词条作为知识库的词,其肯定有一个专属的结果页。当进入所述待处理词专属的结果页时,按照现有文档规定很容易确定所述待处理词专属的结果页对应的文档,即称为所述待处理词专属结果页的文档。
- [0068] 步骤 202,确定设定阈值个与所述文档具有高相关度的词,将确定的词确定为所述待处理词的特征向量词。
- [0069] 优选地,本方式 1 下,步骤 202 具体实现可包括如图 3 所示流程:
- [0070] 步骤 301,对所述待处理词进行分词处理和去噪声干扰,得到对应的处理结果。
- [0071] 本步骤 301 中,待处理词并非一个单独的中文汉字,其可为两个以上的中文汉字和 / 或专有英文词组成,针对这种情况,本步骤 301 可按照现有分词方式和去噪声干扰方式对所述待处理词进行处理,得到对应的处理结果。
- [0072] 步骤 302,从所述处理结果中提取满足设定规定的词作为主题词。
- [0073] 也即,当经过步骤 301 得到处理结果后,本步骤 302 就可以从处理结果中提取满足设定规定的词作为主题词。这里,设定规定可根据实际情况设置,比如从处理结果中提取作为动词和 / 或名词的词作为主题词。
- [0074] 步骤 303,计算每一主题词与所述文档的相关度,选取设定阈值个与所述文档具有高相关度的主题词作为所述待处理词的特征向量词。
- [0075] 优选地,作为一个实施例,本发明中可采用相关性算法模型 BM25 算法计算主题词与所述文档的相关度。需要说明的是,该 BM25 算法只是为便于本申请容易理解所举的实施例,并非用于限定本发明。
- [0076] 以 w 表示任一主题词, d 表示所述待处理词专属结果页的文档为例,则 BM25 算法可通过以下公式 1 表示:

$$[0077] score(w,d) = IDF(w) * \frac{f(w,d)(k_1+1)}{f(w,d)+k_1(1-b+b\frac{|D|}{avgDL})} ; \quad (\text{公式 1})$$

[0078] 其中, $score(w,d)$ 代表主题词 w 与所述文档之间的相关度, $f(w,d)$ 表示主题词 w 在所述待处理词专属结果页的出现次数, D 为所述文档的文档长度, $avgDL$ 为所述待处理词所属词类中所有词专属的所有结果页的文档平均长度, k_1 、 b 为计算相关度的设定参数, $IDF(w)$ 通过以下公式 2 确定:

$$[0079] IDF(w) = \log \frac{N - n(w) + 0.5}{n(w) + 0.5} ; \quad (\text{公式 2})$$

[0080] 其中, N 表示所述待处理词所属词类中所有词专属的所有结果页的总页数, $n(w)$

为所述待处理词所属词类里所有词专属的所有结果页中包含该主题词 w 的结果页数量。

[0081] 下面以待处理词所属词类为情感类为例, 对图 3 所示流程计算相关度的流程进行描述:

[0082] 假如待处理词专属的结果页标注为 d, 该结果页的文档长度为 12, 知识库中属于情感类的所有词所专属的结果页总数 N=99000, 该属于情感类的所有词所专属的结果页的文档平均长度 avgDL 为 10, 则, 基于图 3 所示流程, 首先, 本发明对待处理词进行分词和去除噪音干扰处理; 之后从处理结果中提取出满足设定规定的词作为主题词。以提取出的主题词为“蜜月”、“换位思考”“爱情”为例, 假如情感类里所有词专属的所有结果页中分别包含“蜜月”、“换位思考”“爱情”三个主题词的结果页数、以及提取出的“蜜月”、“换位思考”“爱情”三个主题词分别在待处理词专属结果页的出现次数如表 1 所示:

[0083] 表 1

主题词	蜜月	换位思考	爱情
出现主题词的结果页数	54000	2000	90000
主题词在待处理词专属结果页中出现的次数	5	2	6

[0085] 则, 依据公式 2 得到 IDF(蜜月)、IDF(换位思考)、IDF(爱情)如下:

$$\text{IDF}(\text{蜜月}) = \log((99000 - 54000 + 0.5) / 54000 + 0.5) = 0.83;$$

$$\text{IDF}(\text{换位思考}) = \log((99000 - 2000 + 0.5) / 2000 + 0.5) = 48.49;$$

$$\text{IDF}(\text{爱情}) = \log((99000 - 90000 + 0.5) / 90000 + 0.5) = 0.10;$$

[0089] 假设上述公式 1 中的 k_1 取值 1.5, b 取值为 0.75, 依据公式 1 和上面计算的 IDF(蜜月)、IDF(换位思考)、IDF(爱情)可以得到“蜜月”、“换位思考”“爱情”三个主题词分别与文档 d 的相关度为:

[0090]

$$\text{Score}(\text{蜜月}, d) = 0.83 * \frac{5 * (1.5 + 1)}{5 + 1.5 * (1 - 0.75 + 0.75 * \frac{12}{10})} = 1.54;$$

[0091]

$$\text{Score}(\text{换位思考}, d) = 48.49 * \frac{2 * (1.5 + 1)}{2 + 1.5 * (1 - 0.75 + 0.75 * \frac{12}{10})} = 65.09;$$

[0092]

$$\text{Score}(\text{爱情}, d) = 0.10 * \frac{6 * (1.5 + 1)}{6 + 1.5 * (1 - 0.75 + 0.75 * \frac{12}{10})} = 0.19;$$

[0093] 通过以上计算可以得出, “换位思考”与文档 d 的相关度最高, “蜜月”次之, “爱情”最后。假如按照规定仅为待处理词选取 2 个特征向量词, 则就将“换位思考”、“蜜月”作为待处理词的特征向量词。

[0094] 至此,即可完成方式 1 下确定待处理词的特征向量词操作。

[0095] 下面对方式 2 进行描述:

[0096] 方式 2:

[0097] 本发明中,本方式 2 相对于方式 1,在确定待处理词的特征向量词时不需要进行相关度计算,比较简单。如图 4 所示,本方式 2 下,步骤 101 中确定待处理词的特征向量词可包括:

[0098] 步骤 401,从所述待处理词专属的结果页中找到在预先设置的知识库中具有专属结果页的其他词。

[0099] 如上所述,在知识库中的每一词都有一个专属的结果页,基于此,当进入所述待处理词专属的结果页时,就基于知识库逐次分析待处理词专属结果页中的词,以找到在知识库中具有专属结果页的词。

[0100] 步骤 402,从找到的词中选取设定阈值个词作定为所述待处理词的特征向量词。

[0101] 至此,完成图 4 所示的流程。

[0102] 以待处理词为“张蔚冉”为例;则基于图 4 所示流程,就需要逐次分析待处理词“张蔚冉”专属结果页的词,从中找到预先设置的知识库中具有专属结果页的其他词,假如该找到的词有“不离不弃”,“蓝色妖姬”,“陈冠希”,“快乐男声”,“莫斯科大学”,“朝鲜语”等,则就从该找到的词中任意选取设定阈值个词,比如选取两个词“不离不弃”,“蓝色妖姬”,将该选取的词作定为所述待处理词的特征向量词。

[0103] 以上对方式 2 进行了描述。

[0104] 通过以上方式 1 或方式 2 就能实现图 1 所示流程中步骤 101 确定待处理词的特征向量词。

[0105] 至此,完成图 1 所示流程步骤 101 的详细描述。

[0106] 由于图 1 所示流程中步骤 103 中候选自内链词的特征向量词确定方式与待处理词的特征向量词确定方式一样,则当步骤 101 采用方式 1 确定待处理词的特征向量词时,步骤 103 采用类似方式 1 确定待处理词的特征向量词的方式确定候选自内链词的特征向量词;而当步骤 101 采用方式 2 确定待处理词的特征向量词时,步骤 103 采用类似方式 2 确定待处理词的特征向量词的方式确定候选自内链词的特征向量词。

[0107] 下面对图 1 所示流程步骤 104 进行描述:

[0108] 以设定的推荐分数计算方法为相关度计算方法为例,则步骤 104 中,按照设定的推荐分数计算方法计算并利用每一候选自内链词的特征向量词和所述待处理词的特征向量词计算每一候选自内链词的推荐分数可包括:

[0109] 针对每一候选自内链词,按照设定的相关度计算方法计算该候选自内链词的所有特征向量词和所述待处理词的所有特征向量词之间的相关度,将计算出的相关度作为该候选自内链词的推荐分数。

[0110] 优选地,本发明中,待处理词的所有特征向量词、以及候选自内链词的所有特征向量词通常都是通过特征向量矩阵表示。基于此,作为本发明的一个实施例,所述候选自内链词的所有特征向量词和所述待处理词的所有特征向量词之间的相关度可通过以下公式 3 表示:

$$[0111] \quad recommend_score(y) = \frac{\sum_{i=1}^n \sum_{j=1}^m match(v_i, w_j)}{\sqrt{length(v) * length(w)}} ; \quad (\text{公式 } 3)$$

[0112] 其中, x 为所述待处理词, y 为任一候选自内链词, $recommend_score(x, y)$ 为所述待处理词 x 的所有特征向量词与任一候选自内链词 y 的所有特征向量词之间的相关度, n 为所述待处理词 x 的所有特征向量词的数量, m 为任一候选自内链词 y 的所有特征向量词的数量, $length(v)$ 为所述待处理词 x 的所有特征向量词的总长度, $length(w)$ 任一候选自内链词 y 的所有特征向量词的总长度, $match(v_i, w_j)$ 表示所述待处理词 x 的各个特征向量词与任一候选自内链词 y 的各个特征向量词之间的匹配度, 当 v_i 等于 w_j 时, $match(v_i, w_j)=1$, 当 v_i 不等于 w_j 时, $match(v_i, w_j)=0$ 。

[0113] 基于公式 3 可以看出, 当待处理词与候选自内链词相同时, 该待处理词与候选自内链词之间的相关度为 1, 反之, 当待处理词与候选自内链词没有一个词相同时, 待处理词与候选自内链词之间的相关度为 0。

[0114] 需要说明的是, 上述公式 3 只是在设定的推荐分数计算方法为相关度计算方法时针对相关度计算的一种举例, 并非用于限定本发明。本领域技术人员还可以采用其他方式计算相关度。还有, 本发明中, 设定的推荐分数计算方法并不局限相关度计算方法, 其还可为其他方法, 具体可根据实际需求设置。

[0115] 至此, 完成图 1 所示流程步骤 104 的详细描述。

[0116] 以上对本发明提供的实施例 1 进行了描述。

[0117] 下面对实施例 2 进行描述:

[0118] 参见图 5, 图 5 为本发明实施例 2 提供的方法流程图。本实施例 2 相比于上述的实施例 1, 不需要针对待处理词计算特征向量词, 而是依赖于预先设置的知识库中除待处理词之外的其他词被用户访问的频率确定待处理词的相关自内链词, 比实施例 1 简单, 下面进行详细描述:

[0119] 如图 5 所示, 该流程可包括以下步骤:

[0120] 步骤 501, 将预先设置的知识库中除待处理词之外的其他词作为所述待处理词的候选自内链词。

[0121] 步骤 502, 获取每一候选自内链词在设定时间内被用户访问的次数。

[0122] 通常, 当知识库中任一词被访问时, 知识库自身都有一种记录功能, 用于记录其中的此被访问的时间、次数。基于此, 本步骤 502 依赖于知识库自身的记录功能, 很容易获取每一候选自内链词在设定时间内被用户访问的次数。这里, 设定时间的单位可以是小时、天数、分钟等, 本发明并不具体限定。

[0123] 步骤 503, 计算所述知识库中所有词在所述设定时间内被用户访问的次数之和。

[0124] 这里, 所有词, 也即包括上述的候选自内链词和待处理词。同样, 基于知识库自身的记录功能, 本步骤 503 也会很容易获取待处理词、以及每一候选自内链词在设定时间内被用户访问的次数, 将获取的待处理词、以及每一候选自内链词在设定时间内被用户访问的次数相加, 即为所述知识库中所有词在所述设定时间内被用户访问的次数之和。

[0125] 步骤 504, 按照设定的推荐分数计算方法并利用每一候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算

所述每一候选自内链词的推荐分数。

[0126] 本步骤 504 中设定的推荐分数计算方法，其可根据实际情况设置，比如，可设置为热度计算方法，或者其他方式，本发明并不具体限定。

[0127] 以设定的推荐分数计算方法为热度计算方法为例，则步骤 504 中，按照设定的推荐分数计算方法并利用每一候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算所述每一候选自内链词的推荐分数可包括：

[0128] 针对每一候选自内链词，按照设定的热度计算方法并利用该候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算该候选自内链词的热度，将计算出的热度作为该候选自内链词的推荐分数。

[0129] 优选地，作为本发明一个实施例，候选自内链词的热度通过以下公式实现：

$$[0130] \quad hot_score(v) = \frac{\sum_{k=0}^n p(v)}{\sum_{k=0}^n p(u)} \quad (\text{公式 4})$$

[0131] 其中，v 表示任一候选自内链词，hot_score(v) 表示候选自内链词 v 的热度， $\sum_{k=0}^n p(v)$ 表示候选自内链词 v 在设定时间内被用户访问的次数， $\sum_{k=0}^n p(u)$ 表示知识库中所有词在所述设定时间内被用户访问的次数之和。

[0132] 需要说明的是，上述公式 4 只是在设定的推荐分数计算方法为热度计算方法时针对热度计算的一种举例，并非用于限定本发明。本领域技术人员还可以采用其他方式计算相关度。还有，本发明中，设定的推荐分数计算方法并不局限热度计算方法，其还可为其他方法，具体可根据实际需求设置。

[0133] 步骤 505，选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。

[0134] 至此，完成图 5 所示流程。

[0135] 通过图 5 所示流程，即可实现自动挖掘出待处理词相关的自内链词。并且，优选地，在设定的推荐分数计算方法为热度计算方法时，该挖掘出的待处理词相关的自内链词也是一些被经常访问的热词。

[0136] 至此，完成实施例 2 的描述。

[0137] 可以看出，本发明通过上述实施例 1 或者实施例 2，能够实现在处理某一词时，自动挖掘出该词的自内链词的目的。

[0138] 以上对本发明提供的方法进行了描述。

[0139] 下面对本发明提供的装置进行描述：

[0140] 参见图 6，图 6 为本发明实施例提供的第一种装置结构图。该装置应用于上述的实施例，如图 6 所示，该装置包括：

[0141] 第一确定单元，用于确定待处理词的特征向量词；

[0142] 第二确定单元，用于将在所述待处理词专属的结果页中出现的设定的内链词作为所述待处理词的候选自内链词；

[0143] 第三确定单元,用于按照第一确定单元确定待处理词的特征向量词的方式确定每一候选自内链词的特征向量词;

[0144] 计算单元,用于利用所述待处理词的特征向量词和每一候选自内链词的特征向量词并按照设定的推荐分数计算方法计算所述每一候选自内链词的推荐分数;

[0145] 选取单元,用于选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。

[0146] 优选地,所述第一确定单元确定所述待处理词专属结果页的文档,确定设定阈值个与所述文档具有高相关度的词,将确定的词确定为所述待处理词的特征向量词。

[0147] 优选地,所述第一确定单元在确定设定阈值个与所述文档具有高相关度的词时,对所述待处理词进行分词处理和去噪声干扰,得到对应的处理结果,从所述处理结果中提取满足设定规定的词作为主题词,计算每一主题词与所述文档的相关度,选取设定阈值个与所述文档具有高相关度的主题词作为所述待处理词的特征向量词。

[0148] 优选地,本发明中,所述第一确定单元通过以下公式计算所述每一主题词与所述文档的相关度:

[0149]

$$score(w, d) = IDF(w) * \frac{f(w, d) * (k_1 + 1)}{f(w, d) + k_1(1 - b + b \frac{|D|}{avgDL})}$$

[0150] 其中, w 表示任一主题词, d 表示所述文档, score(w, d) 代表主题词 w 与所述文档之间的相关度, f(w, d) 表示主题词 w 在所述待处理词专属结果页的出现次数, D 为所述文档的文档长度, avgDL 为所述待处理词所属词类中所有词专属的所有结果页的文档平均长度,

k₁、b 为计算相关度的设定参数, IDF(w) 通过以下公式确定: $IDF(w) = \log \frac{N - n(w) + 0.5}{n(w) + 0.5}$; N 表示所述待处理词所属词类中所有词专属的所有结果页的总页数, n(w) 为所述待处理词所属词类里所有词专属的所有结果页中包含该主题词 w 的结果页数。

[0151] 优选地,所述第一确定单元确定待处理词的特征向量词时先从所述待处理词专属的结果页中找到在预先设置的知识库中具有专属结果页的其他词,从找到的词中选取设定阈值个词作定为所述待处理词的特征向量词。

[0152] 优选地,所述计算单元针对每一候选自内链词,按照设定的相关度计算方法计算待处理词的所有特征向量词和该候选自内链词的所有特征向量词之间的相关度,将计算出的相关度作为该候选自内链词的推荐分数。

[0153] 优选地,所述计算单元通过以下公式计算候选自内链词的所有特征向量词和所述待处理词的所有特征向量词之间的相关度:

$$[0154] recommend_score(x, y) = \frac{\sum_{i=1}^n \sum_{j=1}^m match(v_i, w_j)}{\sqrt{length(v) * length(w)}};$$

[0155] 其中, x 为所述待处理词, y 为任一候选自内链词, recommend_score(x, y) 为所述待处理词 x 的所有特征向量词与任一候选自内链词 y 的所有特征向量词之间的相关度, n 为所述待处理词 x 的所有特征向量词的数量, m 为任一候选自内链词 y 的所有特征向量词的

数量, $\text{length}(v)$ 为所述待处理词 x 的所有特征向量词的总长度, $\text{length}(w)$ 任一候选自内链词 y 的所有特征向量词的总长度, $\text{match}(v_i, w_j)$ 表示所述待处理词 x 的各个特征向量词与任一候选自内链词 y 的各个特征向量词之间的匹配度, 当 v_i 等于 w_j 时, $\text{match}(v_i, w_j)=1$, 当 v_i 不等于 w_j 时, $\text{match}(v_i, w_j)=0$ 。

[0156] 至此,完成图 6 所示装置的结构描述。

[0157] 作为本发明另一实施例,本发明还提供了独立于图 6 所示装置的另一种装置。参见图 7,图 7 为本发明实施例提供的另一种装置结构图。如图 7 所示,该装置可包括:

[0158] 确定单元,用于将预先设置的知识库中除待处理词之外的其他词作为所述待处理词的候选自内链词;

[0159] 获取单元,用于获取每一候选自内链词在设定时间内被用户访问的次数;

[0160] 第一计算单元,用于计算所述知识库中所有词在所述设定时间内被用户访问的次数之和;

[0161] 第二计算单元,用于按照设定的推荐分数计算方法并利用每一候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算所述每一候选自内链词的推荐分数;

[0162] 选取单元,用于选取设定数量个推荐分数高的候选自内链词作为待处理词的相关自内链词。

[0163] 优选地,所述第二计算单元在计算每一候选自内链词的推荐分数时,先针对每一候选自内链词,按照设定的热度计算方法并利用该候选自内链词在设定时间内被用户访问的次数和所述知识库中所有词在所述设定时间内被用户访问的次数之和计算该候选自内链词的热度,将计算出的热度作为该候选自内链词的推荐分数。

[0164] 优选地,所述第二计算单元通过以下公式计算候选自内链词的热度:

$$[0165] \quad \text{hot_score}(v) = \frac{\sum_{k=0}^n p(v)}{\sum_{k=0}^n p(u)}$$

[0166] 其中, v 表示任一候选自内链词, $\text{hot_score}(v)$ 表示候选自内链词 v 的热度, $\sum_{k=0}^n p(v)$ 表示候选自内链词 v 在设定时间内被用户访问的次数, $\sum_{k=0}^n p(u)$ 表示知识库中所有词在所述设定时间内被用户访问的次数之和。

[0167] 至此,完成图 7 所示装置的结构描述。

[0168] 由以上技术方案可以看出,本发明中,通过确定待处理词的特征向量词和候选自内链词,利用所述待处理词的特征向量词和每一候选自内链词的特征向量词计算每一候选自内链词的推荐分数,选取设定数量个推荐分数高的候选自内链词作为所述待处理词相关的自内链词,能够实现在处理某一词时,自动挖掘出该词的自内链词的目的。

[0169] 进一步地,本发明中,由于在处理某一词时能够自动推荐出该词相关的自内链词,使用户从推荐的自内链词中找到自己感兴趣的词,无需用户主动重新获取,这一方面提高知识库的词访问效率,另一方面也能节省因为用户频繁访问知识库所浪费的资源。

[0170] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精

神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

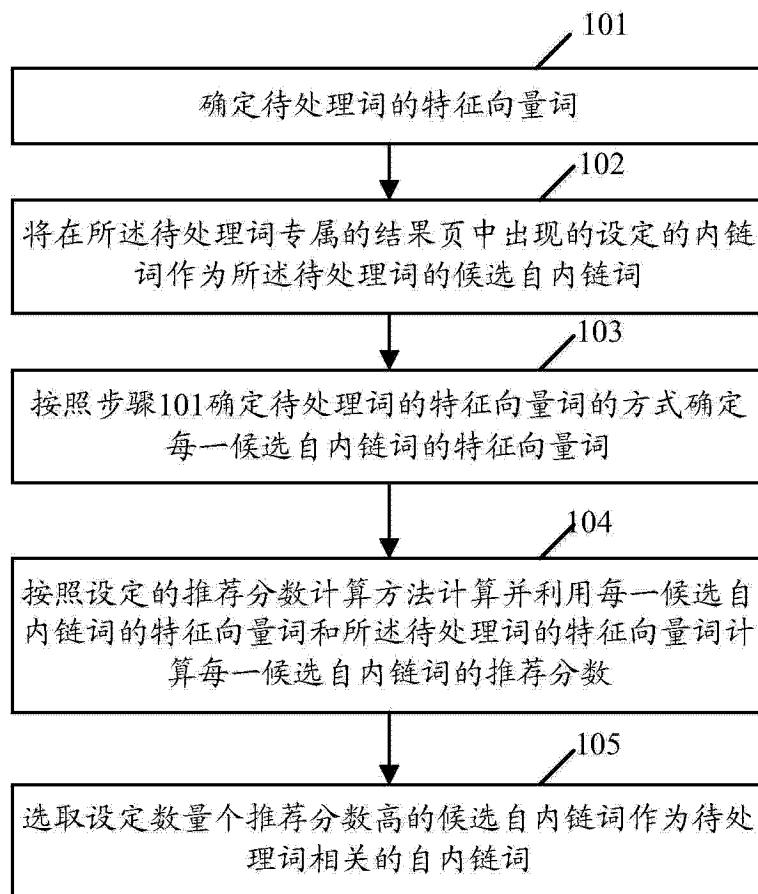


图 1

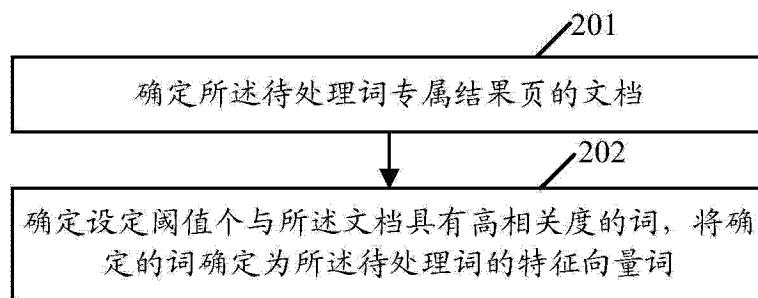


图 2

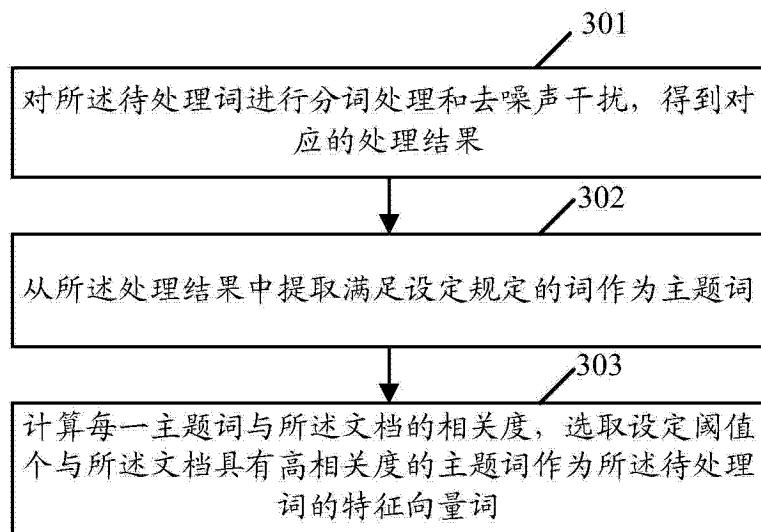


图 3

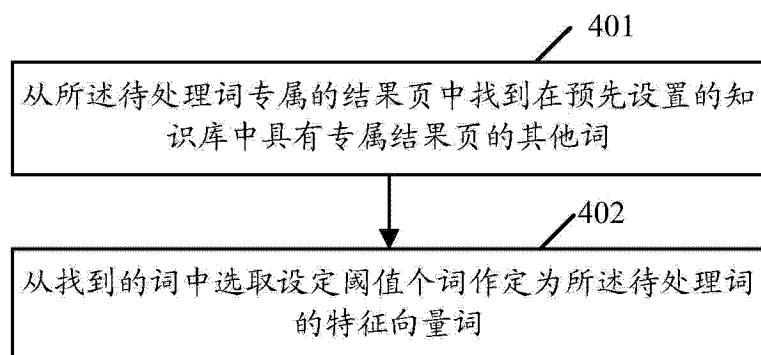


图 4

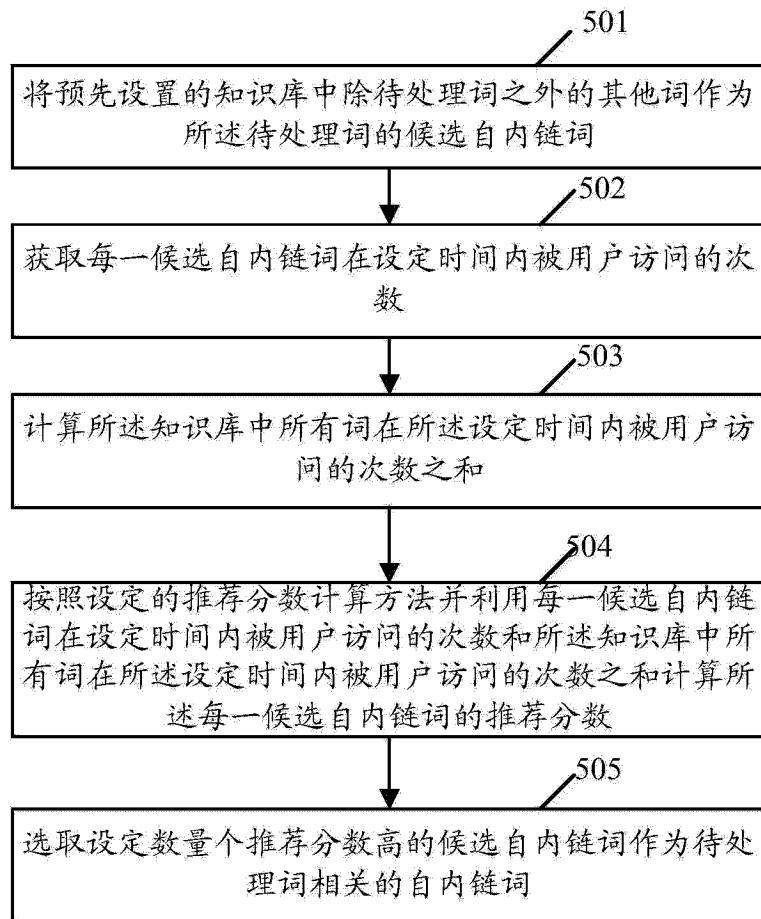


图 5

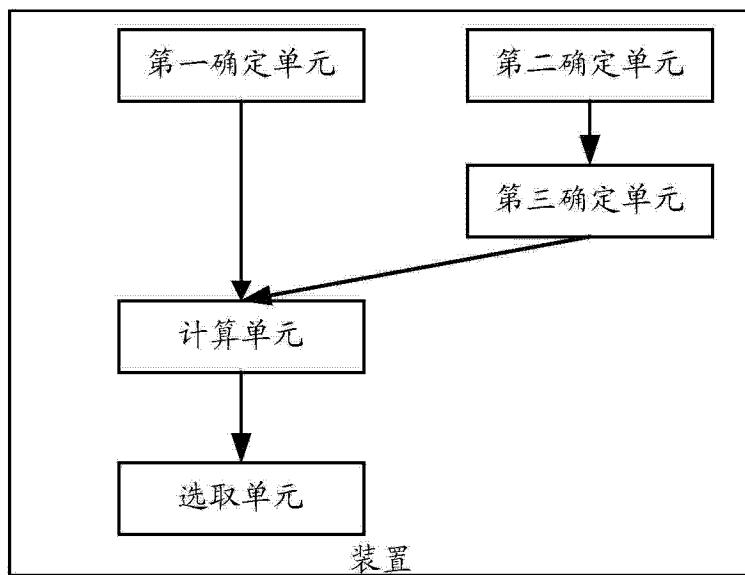


图 6

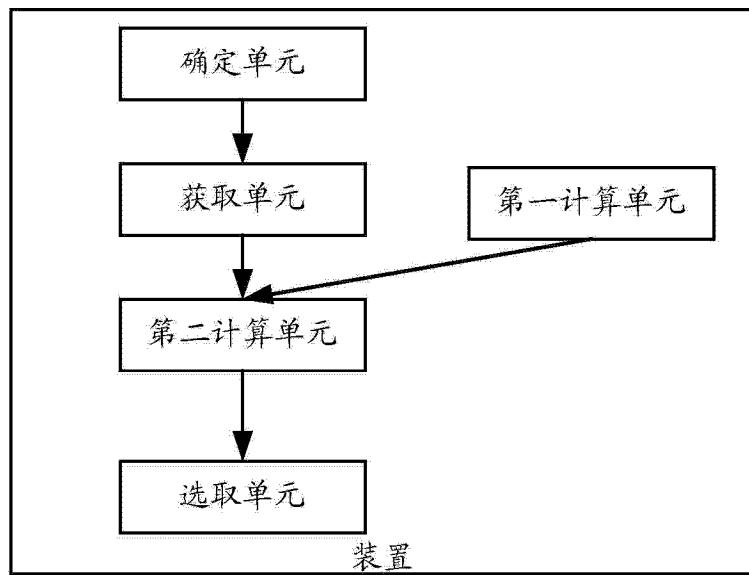


图 7