(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106919603 B
(45)授权公告日 2020.12.04

(21)申请号 201510997477.6
(22)申请日 2015.12.25
(65)同一申请的已公布的文献号
申请公布号 CN 106919603 A
(43)申请公布日 2017.07.04
(73)专利权人 北京奇虎科技有限公司
地址 100088 北京市西城区新街口外大街28号D座112室(德胜园区)
专利权人 奇智软件(北京)有限公司
(72)发明人 陈进平
(74)专利代理机构 北京律诚同业知识产权代理有限公司 11006
代理人 王玉双
(51)Int.Cl.
G06F 16/953(2019.01)
G06F 16/955(2019.01)

(56)对比文件
CN 104778159 A,2015.07.15
CN 104361115 A,2015.02.18
CN 104778233 A,2015.07.15
CN 103838744 A,2014.06.04
CN 101464898 A,2009.06.24
CN 103218364 A,2013.07.24
US 2013246407 A1,2013.09.19

审查员 夏雪

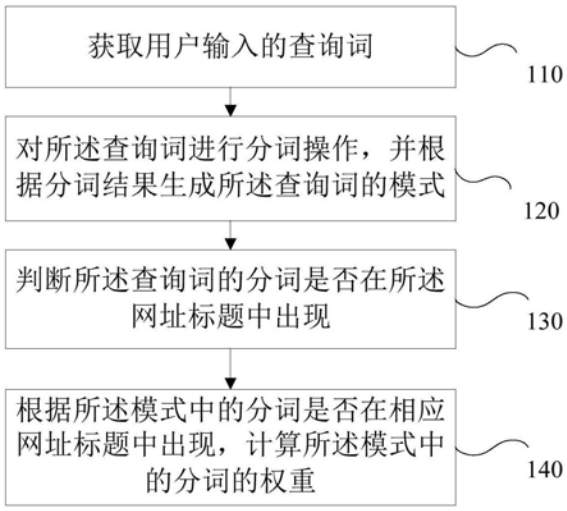
权利要求书2页 说明书9页 附图2页

(54)发明名称

计算查询词模式中分词权重的方法和装置

(57)摘要

本发明提供一种计算查询词模式中分词权重的方法和装置，方法包括：获取用户输入的查询词，以及查询词对应的搜索结果中用户点击的网址标题；对查询词进行分词操作，并根据分词结果生成查询词的模式；判断查询词的分词是否在网址标题中出现；根据模式中的分词是否在相应网址标题中出现，计算模式中的分词的权重。根据本发明，计算得到的查询词模式的分词权重，能够为用户推送符合用户需求的搜索结果。



CN 106919603 B

1.一种计算查询词模式中分词权重的方法,其特征在于,包括:

获取用户输入的查询词,以及所述查询词对应的搜索结果中所述用户点击的网址标题;

对所述查询词进行分词操作,并根据分词结果生成所述查询词的模式;

判断所述查询词的分词是否在所述网址标题中出现;

根据所述模式中的分词是否在相应网址标题中出现,计算所述模式中的分词的权重;

根据所述模式中可替换分词的位置和个数,将所述模式中包含的分词组合划分为多组,计算所有满足所述模式的查询词中,同时包含模式的分词在网址标题里的出现概率,这个概率可以作为分词权重的重要特征。

2.根据权利要求1所述的方法,其特征在于,根据所述模式中的分词是否在相应网址标题中出现,计算所述模式中的分词的权重,具体包括:

根据所述模式中可替换分词的位置和个数,将所述模式中包含的分词组合划分为多组,分别计算多组分组组合中分词的权重。

3.根据权利要求2所述的方法,其特征在于,根据所述模式中的分词是否在相应网址标题中出现,计算所述模式中的分词的权重,还包括:

对所述多组分组组合中分词的权重进行合并,得到所述模式中分词的权重。

4.根据权利要求1-3任一项所述的方法,其特征在于,还包括:

获取多个模式中查找相同的模式,对所述相同模式的权重进行合并。

5.根据权利要求4所述的方法,其特征在于,还包括:

检测所述模式在已知多个查询词中是否出现,根据检测结果判断是否保留所述模式。

6.一种计算查询词模式中分词权重的装置,其特征在于,包括:

获取模块,用于获取用户输入的查询词,以及所述查询词对应的搜索结果中所述用户点击的网址标题;

模式生成模块,用于对所述查询词进行分词操作,并根据分词结果生成所述查询词的模式;

分词判断模块,用于判断所述查询词的分词是否在所述网址标题中出现;

权重计算模块,用于根据所述模式中的分词是否在相应网址标题中出现,计算所述模式中的分词的权重;

根据所述模式中可替换分词的位置和个数,将所述模式中包含的分词组合划分为多组,计算所有满足所述模式的查询词中,同时包含模式的分词在网址标题里的出现概率,这个概率可以作为分词权重的重要特征。

7.根据权利要求6所述的装置,其特征在于,

所述权重计算模块根据所述模式中可替换分词的位置和个数,将所述模式中包含的分词组合划分为多组,分别计算多组分组组合中分词的权重。

8.根据权利要求7所述的装置,其特征在于,还包括:

所述权重计算模块对所述多组分组组合中分词的权重进行合并,得到所述模式中分词的权重。

9.根据权利要求6-8任一项所述的装置,其特征在于,

所述权重计算模块获取多个模式中查找相同的模式,对所述相同模式的权重进行合

并。

10.根据权利要求9所述的装置,其特征在于,还包括:

过滤模块,用于检测所述模式在已知多个查询词中是否出现,根据检测结果判断是否保留所述模式。

计算查询词模式中分词权重的方法和装置

技术领域

[0001] 本发明涉及计算机技术领域，具体而言，涉及一种计算查询词模式中分词权重的方法和装置。

背景技术

[0002] 查询词是用户通过浏览器提交给搜索引擎的请求，通常是一串表达用户需求的字符串。搜索引擎在根据查询词进行搜索时，需要对查询词进行分词操作，并分析分词结果的权重，以按照得到分词的权重提供搜索结果；分词权重是查询词分析中非常重要的目标，对搜索引擎的能否满足用户的搜索需求起着决定性的作用。

[0003] 目前，对于查询词的分词权重的计算存在很多的方法，例如下面的一些技术：1、基于共同点击的分词权重计算方法；2、基于分词词性的分词权重计算方法；3、基于命名实体的分词权重计算方法。但是以上的这些技术，所计算得到的分词权重的方案都存在相应缺陷，因此需要提出一种新的用于计算分词权重的方案。

发明内容

[0004] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的计算查询词模式中分词权重的方法和装置。

[0005] 依据本发明的一种计算查询词模式中分词权重的方法，包括：获取用户输入的查询词，以及所述查询词对应的搜索结果中所述用户点击的网址标题；对所述查询词进行分词操作，并根据分词结果生成所述查询词的模式；判断所述查询词的分词是否在所述网址标题中出现；根据所述模式中的分词是否在相应网址标题中出现，计算所述模式中的分词的权重。

[0006] 可选地，前述的方法，根据所述模式中的分词是否在相应网址标题中出现，计算所述模式中的分词的权重，具体包括：根据所述模式中可替换分词的位置和个数，将所述模式中包含的分词组合划分为多组，分别计算多组分组组合中分词的权重。

[0007] 可选地，前述的方法，根据所述模式中的分词是否在相应网址标题中出现，计算所述模式中的分词的权重，还包括：对所述多个分组组合中分词的权重进行合并，得到所述模式中分词的权重。

[0008] 可选地，前述的方法，还包括：获取多个模式中查找相同的模式，对所述相同模式的权重进行合并。

[0009] 可选地，前述的方法，还包括：检测所述模式在已知多个查询词中是否出现，根据检测结果判断是否保留所述模式。

[0010] 依据本发明的一种计算查询词模式中分词权重的装置，包括：获取模块，用于获取用户输入的查询词，以及所述查询词对应的搜索结果中所述用户点击的网址标题；模式生成模块，用于对所述查询词进行分词操作，并根据分词结果生成所述查询词的模式；分词判断模块，用于判断所述查询词的分词是否在所述网址标题中出现；权重计算模块，用于根据

所述模式中的分词是否在相应网址标题中出现，计算所述模式中的分词的权重。

[0011] 可选地，前述的装置，所述权重计算模块根据所述模式中可替换分词的位置和个数，将所述模式中包含的分词组合划分为多组，分别计算多组分组组合中分词的权重。

[0012] 可选地，前述的装置，还包括：所述权重计算模块对所述多个分组组合中分词的权重进行合并，得到所述模式中分词的权重。

[0013] 可选地，前述的装置，所述权重计算模块获取多个模式中查找相同的模式，对所述相同模式的权重进行合并。

[0014] 可选地，前述的装置，还包括：过滤模块，用于检测所述模式在已知多个查询词中是否出现，根据检测结果判断是否保留所述模式。

[0015] 根据以上技术方案，本发明的计算查询词模式中分词权重的方法和装置至少具有以下优点：

[0016] 在本发明的技术方案中，用户输入查询词后，在搜索结果中点击的网址标题反映了用户输入的查询词的需求，因此基于用户所点击的网址标题，对查询词拆分模式并分析模式分词的权重，得到模式中的分词权重值能够体现该分词对于用户的重要程度；基于本发明计算得到的查询词模式的分词权重，能够为用户推送符合用户需求的搜索结果。

[0017] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0018] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0019] 图1示出了根据本发明的一个实施例的一种计算查询词模式中分词权重的方法的流程图；

[0020] 图2示出了根据本发明的一个实施例的一种计算查询词模式中分词权重的装置的框图；

[0021] 图3示出了根据本发明的一个实施例的一种计算查询词模式中分词权重的装置的框图。

具体实施方式

[0022] 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0023] 在描述本发明的实施例前，需要对以下概念进行说明：

[0024] 查询词(query)是指，用户通过浏览器提交给搜索引擎的请求，通常是一串表达用户需求的字符串。

[0025] 查询词的模式(pattern)是指：模式是指不同的查询词都能按某种方式来表示，例

如用正则表达式;例如下面的这几个查询词:

[0026] 查询词1:但字怎么造句

[0027] 查询词2:即字怎么造句

[0028] 这两个查询词表达了不同的事情(但和即的造句),但是有相同的说法,根据这两个查询词可以得到如下的模式:*字怎么造句,这里的“*”为通配符,表示无或任意的汉字。又比如,对于查询词:混合性皮肤适合用的化妆品,可以得到如下的模式:混合*皮肤*化妆*品*。

[0029] 分词(term)权重:分词是指对查询词进行分词操作后的基本单位,分词权重就是指计算查询词分词后得到的每个分词在这个查询词里的相对权重,分词权重是查询词分析中非常重要的目标,对搜索引擎的能否满足用户的搜索需求起着决定性的作用。

[0030] 如图1所示,本发明的一个实施例中提供一种计算查询词模式中分词权重的方法,包括:

[0031] 步骤110,获取用户输入的查询词,以及所述查询词对应的搜索结果中所述用户点击的网址标题。在本实施例中,将将用户提交给搜索引擎的查询词以及查询词点击的网址(url)标题作为输入。用户输入查询词后,在搜索结果中点击的网址标题反映了用户输入的查询词的需求。

[0032] 步骤120,对所述查询词进行分词操作,并根据分词结果生成所述查询词的模式。在本实施例中,对每一个<查询词,标题>的组合,首先对查询词进行分词操作,在查询词的分词结果中任意选取一个词、两个词、三个词、四个词的所有组合,按照在查询词中的顺序组装为模式。例如:某个查询词为ABCDE,假设每个字母表示分词后的分词,则可以得到如下的模式:

[0033] 1、一个词,A*,*B*,*C*,*D*,*E,这里用“*”表示通配符;

[0034] 2、两个词,A*B*,A*C*,A*D*,A*E……

[0035] 3、三个词,A*B*C*,A*B*D*,A*B*E……

[0036] 4、四个词,A*B*C*D*,A*B*C*E,*B*C*D*E……

[0037] 步骤130,判断所述查询词的分词是否在所述网址标题中出现。在本实施例中,需要计算查询词中的分词是否在标题中出现,出现记录为1,否则为0:假设ABCDE这5个词在标题中的出现情况为1、0、1、1、0,即A、C、D在标题里出现,B、E在标题中没有出现。

[0038] 步骤140,根据所述模式中的分词是否在相应网址标题中出现,计算所述模式中的分词的权重。根据本实施例,可以将分词在标题中的出现情况作为模式的权重值输出。由于在搜索结果中点击的网址标题反映了用户输入的查询词的需求,因此基于用户所点击的网址标题,对查询词拆分模式并分析模式分词的权重,得到模式中的分词权重值能够体现该分词对于用户的重要程度;基于本发明计算得到的查询词模式的分词权重,能够为用户推送符合用户需求的搜索结果。

[0039] 本发明的一个实施例中提供一种计算查询词模式中分词权重的方法,相比于前述的实施例,本实施例的计算查询词模式中分词权重的方法,步骤140,具体包括:

[0040] 根据所述模式中可替换分词的位置和个数,将所述模式中包含的分词组合划分为多组,分别计算多组分组组合中分词的权重。在本实施例中,计算权重值时按照可替换的分词的位置和个数进行分组,例如:对于模式:*B*C*D*E,*通配符代表了可替换分词,则在计

算流程里会如下计算权重值：

[0041] 1、计算所有满足这个模式的查询词中，B、C、D、E这四个分词在标题中的出现概率；

[0042] 2、对于出现在B之前的可替换分词的情况，按照个数进行分组，例如，针对在B之前只有一个分词的、有2个分词的、有3个分词的、有4个分词的分词情况，分别统计这4种情况下形成的分词组合中每个分词在标题中出现的概率；

[0043] 3、同样地，对于出现在B和C之间的可替换分词的情况、C和D之间的分词情况、D和E之间的分词情况、E后面的分词情况，也按照分词的个数进行分组得到多个分词组合，为每个分词组合计算得到在标题中的出现概率。

[0044] 在上面的例子上，假定A、C、D在标题中出现，那么对于*B*E*这个模式的，其中一个分词组合的分词权重值如下：

[0045] *B*E*:1,0,11,0

[0046] 第一个1表示B前面有一个分词，并且出现在标题；

[0047] 第二个0表示B没有出现在标题；

[0048] 第三个11表示B和E中间有两个分词，并且都在标题出现；

[0049] 第四个0表示E没有出现在标题。

[0050] 在本实施例中，基于可替换分词的个数和位置，对模式进行了细分，以利于更准确地计算每个分词的权重。

[0051] 本发明的一个实施例中提供一种计算查询词模式中分词权重的方法，相比于前述的实施例，本实施例的计算查询词模式中分词权重的方法，步骤140，还包括：

[0052] 对所述多个分组组合中分词的权重进行合并，得到所述模式中分词的权重。在本实施例中，多个分词组合合并后输出权重值的格式举例：

[0053] *B*E*:x|xx|xxx|xxxx,x,|x|xx|xxx|xxx|xxxx,x,x|xx|xxx|xxxx

[0054] 上面这个例子中每一个x表示一个实际的数，可能是0或者1，表示当前<查询词，标题>对中某个分词是否出现在标题中的统计。

[0055] 用“|”分隔的表示某个区间里1个、2个、3个、4个分词在标题出现的情况，例如一开始的3个“|”分别记录B前面只有一个分词时这个分词是否在标题中出现、有2个分词时这2个分词的出现情况等等，用逗号隔开了表示在模式B、E之间可替换的分词在标题里的出现情况，以及B和E在标题中的出现情况；在本实施例中，综合了多个分词组合的分词权重得到模式中分词的权重，数据量减少更加适于存储和使用。

[0056] 本发明的一个实施例中提供一种计算查询词模式中分词权重的方法，相比于前述的实施例，本实施例的计算查询词模式中分词权重的方法，还包括：

[0057] 获取多个模式中查找相同的模式，对所述相同模式的权重进行合并。

[0058] 在本实施例中，在每个<查询词，标题>中，能够得到模式的一个值；最后把相同模式的不同值进行合并，主要是处理不同分词的情况，例如：

[0059] *B*E*:1,0,11,0

[0060] *B*E*:11,1,1,0,1

[0061] 合并后为

[0062] *B*E:1|11,0.5,1|11,0,1

[0063] 第一个1|11，表示B前面存在一个分词和2个分词这两种情况，且他们都在标题里

出现；

[0064] 第二个0.5，表示B在标题中出现的概率是0.5；

[0065] 第三个1|11表示B和E之间存在一个分词和2个分词这两种情况，且他们都在标题出现；

[0066] 第四个0表示E没有在标题出现；

[0067] 第五个1表示E后面有一个分词，并且在标题出现。

[0068] 在本实施例中，用户可能多次输入同一个查询词而点击了不同的搜索结果，则根据查询词和单次点击的搜索结果的网址标题计算模式的分词权重可能存在不准确的情况；而本实施例中对相同模式的分词权重组合，相当于综合了用户点击同一查询词以及用多次点击的搜索结果的网址标题来计算查询词模式的分词权重，所以计算结果更加准确。

[0069] 本发明的一个实施例中提供一种计算查询词模式中分词权重的方法，相比于前述的实施例，本实施例的计算查询词模式中分词权重的方法，还包括：

[0070] 检测所述模式在已知多个查询词中是否出现，根据检测结果判断是否保留所述模式。

[0071] 在本实施例中，通过模式在所有<查询词，标题>的出现次数进行过滤，最后得到大概1亿个模式，清除了重复的数据。

[0072] 综合以上实施例，可以大规模地挖掘查询词的模式，并且同时包含模式的分词在网址标题里的出现概率，这个概率可以作为分词权重的重要特征，例如：

[0073] 查询词：但怎么造句，可以匹配如下模式：

[0074] *怎么*造句*：0.79|0.72 0.73|0.64 0.65 0.65|0.67 0.61 0.62 0.63，0.29…

[0075] 通过这个模式，我们能够发现“但”这个单字，并且是停用词的单字，在这个查询词里有重要的作用，因为当“怎么”前面只有一个分词时，这个分词在标题中的出现概率是0.79；利用这个信息来改进分词的权重值，有利于节省对查询词的分析，搜索结果的质量能够取得明显改进。

[0076] 如图2所示，本发明的一个实施例中提供一种计算查询词模式中分词权重的装置，包括：

[0077] 获取模块210，获取用户输入的查询词，以及所述查询词对应的搜索结果中所述用户点击的网址标题。在本实施例中，将将用户提交给搜索引擎的查询词以及查询词点击的网址(url)标题作为输入。用户输入查询词后，在搜索结果中点击的网址标题反映了用户输入的查询词的需求。

[0078] 模式生成模块220，对所述查询词进行分词操作，并根据分词结果生成所述查询词的模式。在本实施例中，对每一个<查询词，标题>的组合，首先对查询词进行分词操作，在查询词的分词结果中任意选取一个词、两个词、三个词、四个词的所有组合，按照在查询词中的顺序组装为模式。例如：某个查询词为ABCDE，假设每个字母表示分词后的分词，则可以得到如下的模式：

[0079] 1、一个词，A*，*B*，*C*，*D*，*E，这里用“*”表示通配符；

[0080] 2、两个词，A*B*，A*C*，A*D*，A*E……

[0081] 3、三个词，A*B*C*，A*B*D*，A*B*E……

[0082] 4、四个词，A*B*C*D*，A*B*C*E，*B*C*D*E……

[0083] 分词判断模块230，判断所述查询词的分词是否在所述网址标题中出现。在本实施例中，需要计算查询词中的分词是否在标题中出现，出现记录为1，否则为0：假设ABCDE这5个词在标题中的出现情况为1、0、1、1、0，即A、C、D在标题里出现，B、E在标题中没有出现。

[0084] 权重计算模块240，根据所述模式中的分词是否在相应网址标题中出现，计算所述模式中的分词的权重。根据本实施例，可以将分词在标题中的出现情况作为模式的权重值输出。由于在搜索结果中点击的网址标题反映了用户输入的查询词的需求，因此基于用户所点击的网址标题，对查询词拆分模式并分析模式分词的权重，得到模式中的分词权重值能够体现该分词对于用户的重要程度；基于本发明计算得到的查询词模式的分词权重，能够为用户推送符合用户需求的搜索结果。

[0085] 本发明的一个实施例中提供一种计算查询词模式中分词权重的装置，相比于前述的实施例，本实施例的计算查询词模式中分词权重的装置，

[0086] 权重计算模块240根据所述模式中可替换分词的位置和个数，将所述模式中包含的分词组合划分为多组，分别计算多组分组组合中分词的权重。在本实施例中，计算权重值时按照可替换的分词的位置和个数进行分组，例如：对于模式：*B*C*D*E，*通配符代表了可替换分词，则在计算流程里会如下计算权重值：

[0087] 1、计算所有满足这个模式的查询词中，B、C、D、E这四个分词在标题中的出现概率；

[0088] 2、对于出现在B之前的可替换分词的情况，按照个数进行分组，例如，针对在B之前只有一个分词的、有2个分词的、有3个分词的、有4个分词的分词情况，分别统计这4种情况下形成的分词组合中每个分词在标题中出现的概率；

[0089] 3、同样地，对于出现在B和C之间的可替换分词的情况、C和D之间的分词情况、D和E之间的分词情况、E后面的分词情况，也按照分词的个数进行分组得到多个分词组合，为每个分词组合计算得到在标题中的出现概率。

[0090] 在上面的例子上，假定A、C、D在标题中出现，那么对于*B*E*这个模式的，其中一个分词组合的分词权重值如下：

[0091] *B*E*：1，0，11，0

[0092] 第一个1表示B前面有一个分词，并且出现在标题；

[0093] 第二个0表示B没有出现在标题；

[0094] 第三个11表示B和E中间有两个分词，并且都在标题出现；

[0095] 第四个0表示E没有出现在标题。

[0096] 在本实施例中，基于可替换分词的个数和位置，对模式进行了细分，以利于更准确地计算每个分词的权重。

[0097] 本发明的一个实施例中提供一种计算查询词模式中分词权重的装置，相比于前述的实施例，本实施例的计算查询词模式中分词权重的装置，

[0098] 权重计算模块240对所述多个分组组合中分词的权重进行合并，得到所述模式中分词的权重。在本实施例中，多个分词组合合并后输出权重值的格式举例：

[0099] *B*E*：x|xx|xxx|xxxx，x，|x|xx|xxx|xxx|xxxx，x，x|xx|xxx|xxxx

[0100] 上面这个例子中每一个x表示一个实际的数，可能是0或者1，表示当前<查询词，标题>对中某个分词是否出现在标题中的统计。

[0101] 用“|”分隔的表示某个区间里1个、2个、3个、4个分词在标题出现的情况，例如一开

始的3个“|”分别记录B前面只有一个分词时这个分词是否在标题中出现、有2个分词时这2个分词的出现情况等等，用逗号隔开了表示在模式B、E之间可替换的分词在标题里的出现情况，以及B和E在标题中的出现情况；在本实施例中，综合了多个分词组合的分词权重得到模式中分词的权重，数据量减少更加适于存储和使用。

[0102] 本发明的一个实施例中提供一种计算查询词模式中分词权重的装置，相比于前述的实施例，本实施例的计算查询词模式中分词权重的装置，

[0103] 权重计算模块240获取多个模式中查找相同的模式，对所述相同模式的权重进行合并。

[0104] 在本实施例中，在每个<查询词，标题>中，能够得到模式的一个值；最后把相同模式的不同值进行合并，主要是处理不同分词的情况，例如：

[0105] *B*E*:1,0,11,0

[0106] *B*E*:11,1,1,0,1

[0107] 合并后为

[0108] *B*E:1|11,0.5,1|11,0,1

[0109] 第一个1|11，表示B前面存在一个分词和2个分词这两种情况，且他们都在标题里出现；

[0110] 第二个0.5，表示B在标题中出现的概率是0.5；

[0111] 第三个1|11表示B和E之间存在一个分词和2个分词这两种情况，且他们都在标题出现；

[0112] 第四个0表示E没有在标题出现；

[0113] 第五个1表示E后面有一个分词，并且在标题出现。

[0114] 在本实施例中，用户可能多次输入同一个查询词而点击了不同的搜索结果，则根据查询词和单次点击的搜索结果的网址标题计算模式的分词权重可能存在不准确的情况；而本实施例中对相同模式的分词权重组合，相当于综合了用户点击同一查询词以及用多次点击的搜索结果的网址标题来计算查询词模式的分词权重，所以计算结果更加准确。

[0115] 如图3所示，本发明的一个实施例中提供一种计算查询词模式中分词权重的装置，相比于前述的实施例，本实施例的计算查询词模式中分词权重的装置，还包括：

[0116] 过滤模块310，检测所述模式在已知多个查询词中是否出现，根据检测结果判断是否保留所述模式。

[0117] 在本实施例中，通过模式在所有<查询词，标题>的出现次数进行过滤，最后得到大概1亿个模式，清除了重复的数据。综合以上实施例，可以大规模地挖掘查询词的模式，并且同时包含模式的分词在网址标题里的出现概率，这个概率可以作为分词权重的重要特征，例如：

[0118] 查询词:但怎么造句，可以匹配如下模式：

[0119] *怎么*造句*:0.79|0.72 0.73|0.64 0.65 0.65|0.67 0.61 0.62 0.63,0.29…

[0120] 通过这个模式，我们能够发现“但”这个单字，并且是停用词的单字，在这个查询词里有重要的作用，因为当“怎么”前面只有一个分词时，这个分词在标题中的出现概率是0.79；利用这个信息来改进分词的权重值，有利于节省对查询词的分析，搜索结果的质量能够取得明显改进。

[0121] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述，构造这类系统所要求的结构是显而易见的。此外，本发明也不针对任何特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本发明的内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0122] 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

[0123] 类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

[0124] 本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0125] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0126] 本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明实施例的计算查询词模式中分词权重的装置中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如，计算机程序和计算机程序产品)。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

[0127] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的

元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中，这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

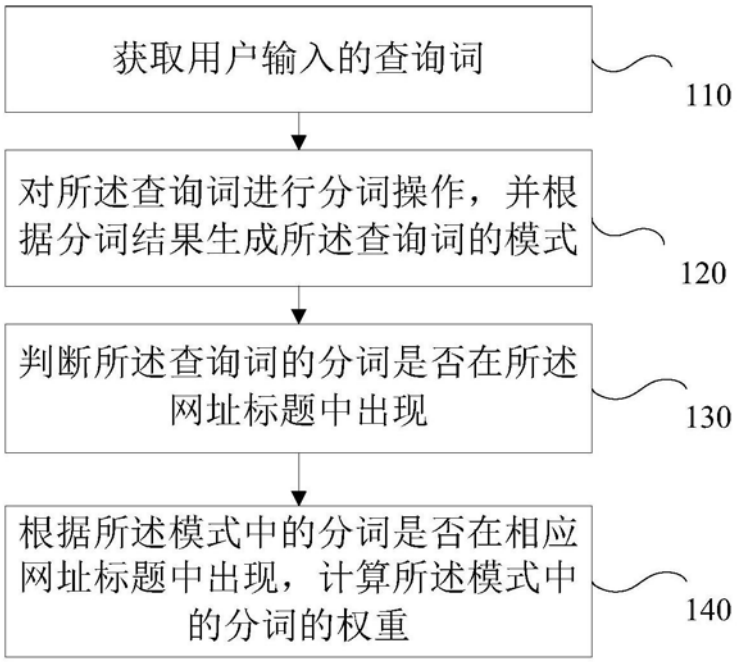

获取用户输入的查询词
110
对所述查询词进行分词操作，并根据分词结果生成所述查询词的模式
120
判断所述查询词的分词是否在所述网址标题中出现
130
根据所述模式中的分词是否在相应网址标题中出现，计算所述模式中的分词的权重
140

图1

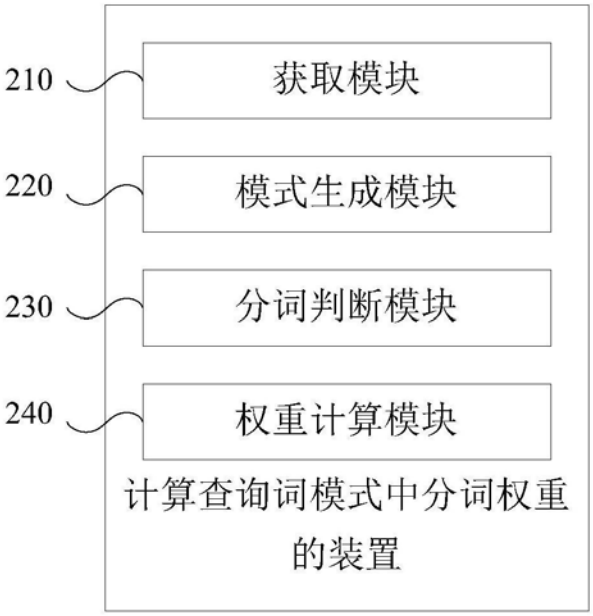

210
获取模块
220
模式生成模块
230
分词判断模块
240
权重计算模块
计算查询词模式中分词权重的装置

图2

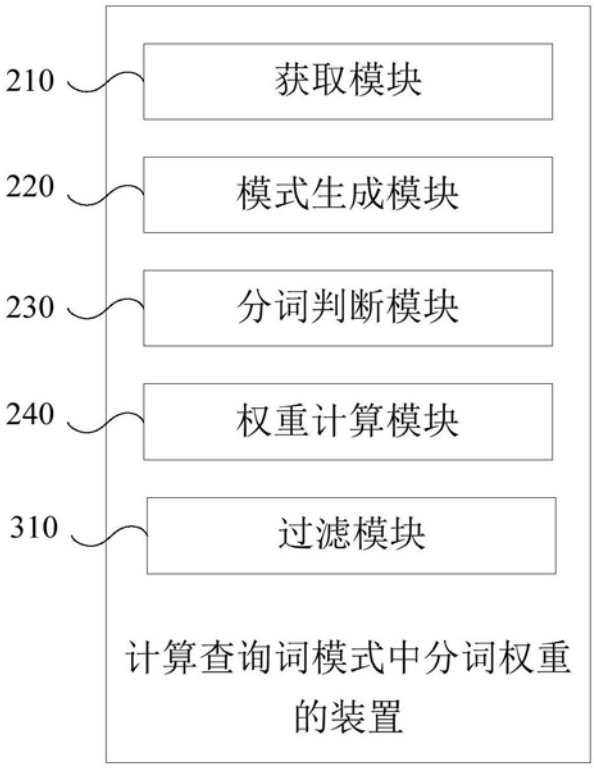

210
获取模块
220
模式生成模块
230
分词判断模块
240
权重计算模块
310
过滤模块
计算查询词模式中分词权重的装置

图3