(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103136213 A
(43)申请公布日 2013.06.05

(21)申请号 201110376840.4

(22)申请日 2011.11.23

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼岛资本大厦一座四层847号邮箱

(72)发明人 钟灵 周祥军 申月 杨洁 蒋龙

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 郭润湘

(51)Int.Cl.
G06F 17/30(2006.01)

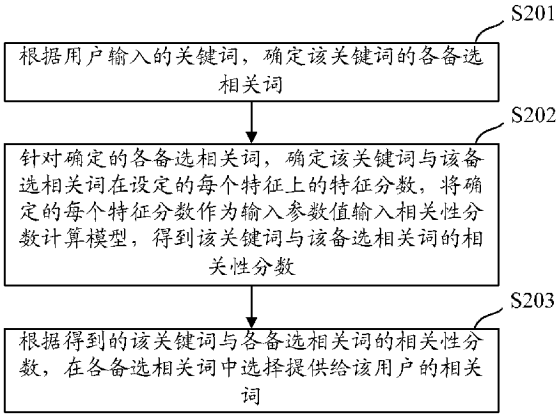
权利要求书4页 说明书9页 附图3页

(54)发明名称

一种提供相关词的方法及装置

(57)摘要

本申请公开了一种提供相关词的方法及装置，用以解决现有技术中提供的相关词不够准确的问题。该方法针对用户输入的关键词的各备选相关词，将该关键词与该备选相关词在设定的每个特征上的特征分数输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，并据此提供相关词，其中，该相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的。通过上述方法，即使用户输入的关键词未记录在搜索日志中，也可以通过将该关键词与各备选相关词的特征分数输入相关性分数计算模型，来获得该关键词与各备选相关词的相关性分数，从而为用户提供准确的相关词，使用户无需再次进行搜索，节省了服务器资源。



CN 103136213 A

1. 一种提供相关词的方法，其特征在于，包括：

根据用户输入的关键词，确定所述关键词的各备选相关词；

针对确定的各备选相关词，确定所述关键词与该备选相关词在设定的每个特征上的特征分数，将确定的每个特征分数作为输入参数值输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，其中，所述相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的；

根据得到的所述关键词与各备选相关词的相关性分数，在各备选相关词中选择提供给所述用户的相关词。

2. 如权利要求 1 所述的方法，其特征在于，根据设定数量的已计算出相关性分数的关键词与相关词，确定所述相关性分数计算模型，具体包括：

确定已计算出相关性分数的关键词与相关词作为训练样本，选择设定数量的训练样本；

针对选择的每个训练样本，根据所述设定的每个特征，确定该训练样本中的关键词和相关词在每个特征上的特征分数，将已计算出的该训练样本中的关键词和相关词的相关性分数确定为目标值，将确定的该训练样本中的关键词和相关词在每个特征上的特征分数确定为输入参数值；

根据针对每个训练样本确定的目标值和输入参数值，采用设定的算法进行回归运算，得到相关性分数计算模型。

3. 如权利要求 2 所述的方法，其特征在于，根据所述设定的每个特征，确定该训练样本中的关键词和相关词在每个特征上的特征分数，具体包括：

确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数；以及

确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数；以及

确定该训练样本中的关键词和相关词的编辑距离作为编辑距离分数；以及

确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

4. 如权利要求 3 所述的方法，其特征在于，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数，具体包括：

采用该训练样本中的关键词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索类目的搜索结果的个数与搜索结果总数的比值；

将采用该训练样本中的关键词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量确定为关键词类目向量；

采用该训练样本中的相关词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索结果类目的搜索结果的个数与搜索结果总数的比值；

将采用该训练样本中的相关词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量确定为相关词类目向量；

确定所述关键词类目向量与所述相关词类目向量的余弦值，将所述余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数。

5. 如权利要求 3 所述的方法，其特征在于，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数，具体包括：

根据该训练样本中的关键词，确定采用该训练样本中的关键词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第一集合；

根据该训练样本中的相关词，确定采用该训练样本中的相关词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第二集合；

确定所述第一集合与第二集合的交集以及并集，确定所述交集中包含的元素的个数与所述并集中包含的元素的个数的比值，将所述比值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数。

6. 如权利要求 3 所述的方法，其特征在于，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数，具体包括：

确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的每个相同的搜索结果；

针对每个相同的搜索结果，根据搜索日志中的记录，确定通过该训练样本中的关键词进行搜索时该搜索结果被点击的次数，确定通过该训练样本中的相关词进行搜索时该搜索结果被点击的次数；

将通过该训练样本中的关键词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量确定为关键词点击向量；

将通过该训练样本中的相关词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量确定为相关词点击向量；

确定所述关键词点击向量与所述相关词点击向量的余弦值，将所述余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

7. 如权利要求 2 ～ 6 任一所述的方法，其特征在于，采用设定的算法进行回归运算，得到相关性分数计算模型，具体包括：

采用支持向量机 SVM 算法进行回归运算，得到相关性分数计算模型；或者

采用评定模型 Logit 算法进行回归运算，得到相关性分数计算模型。

8. 一种提供相关词的装置，其特征在于，包括：

备选相关词确定模块，用于根据用户输入的关键词，确定所述关键词的各备选相关词；

相关性分数确定模块，用于针对确定的各备选相关词，确定所述关键词与该备选相关词在设定的每个特征上的特征分数，将确定的每个特征分数作为输入参数值输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，其中，所述相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的；

相关词提供模块，用于根据得到的所述关键词与各备选相关词的相关性分数，在各备选相关词中选择提供给所述用户的相关词。

9. 如权利要求 8 所述的装置，其特征在于，所述相关性分数确定模块包括 ：

确定选择子模块，用于确定已计算出相关性分数的关键词与相关词作为训练样本，选择设定数量的训练样本 ；

特征分数确定子模块，用于针对选择的每个训练样本，根据所述设定的每个特征，确定该训练样本中的关键词和相关词在每个特征上的特征分数，将已计算出的该训练样本中的关键词和相关词的相关性分数确定为目标值，将确定的该训练样本中的关键词和相关词在每个特征上的特征分数确定为输入参数值 ；

模型确定子模块，用于根据针对每个训练样本确定的目标值和输入参数值，采用设定的算法进行回归运算，得到相关性分数计算模型。

10. 如权利要求 9 所述的装置，其特征在于，所述特征分数确定子模块具体用于，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数，以及，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数，以及，确定该训练样本中的关键词和相关词的编辑距离作为编辑距离分数，以及，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

11. 如权利要求 10 所述的装置，其特征在于，所述特征分数确定子模块具体用于，采用该训练样本中的关键词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索类目的搜索结果个数与搜索结果总数的比值，将采用该训练样本中的关键词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量确定为关键词类目向量 ；采用该训练样本中的相关词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索结果类目的搜索结果个数与搜索结果总数的比值，将采用该训练样本中的相关词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量确定为相关词类目向量 ；确定所述关键词类目向量与所述相关词类目向量的余弦值，将所述余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数。

12. 如权利要求 10 所述的装置，其特征在于，所述特征分数确定子模块具体用于，根据该训练样本中的关键词，确定采用该训练样本中的关键词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第一集合 ；根据该训练样本中的相关词，确定采用该训练样本中的相关词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第二集合 ；确定所述第一集合与第二集合的交集以及并集，确定所述交集中包含的元素的个数与所述并集中包含的元素的个数的比值，将所述比值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数。

13. 如权利要求 10 所述的装置，其特征在于，所述特征分数确定子模块具体用于，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的每个相同的搜索结果 ；针对每个相同的搜索结果，根据搜索日志中的记录，确定通过该训练样本中的关键词进行搜索时该搜索结果被点击的次数，确定通过该训练样本中的相关词进行搜索时该搜索结果被点击的次数 ；将通过该训练样本中的关键词进行搜索时，针对每个相同的搜索结果确定的每

个被点击的次数构成的向量确定为关键词点击向量；将通过该训练样本中的相关词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量确定为相关词点击向量；确定所述关键词点击向量与所述相关词点击向量的余弦值，将所述余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

14. 如权利要求 9 ～ 13 任一所述的装置，其特征在于，所述模型确定子模块具体用于，采用支持向量机 SVM 算法进行回归运算，得到相关性分数计算模型，或者，采用评定模型 Logit 算法进行回归运算，得到相关性分数计算模型。

一种提供相关词的方法及装置

技术领域

[0001] 本申请涉及通信技术领域，尤其涉及一种提供相关词的方法及装置。

背景技术

[0002] 目前，很多购物网站的服务器都提供了商品搜索的功能，用户输入想要搜索的商品的关键词，服务器则根据该关键词搜索相应的结果并返回给用户，也即向搜索到的商品信息返回给用户。

[0003] 由于用户输入的关键词往往不规范，只用该关键词搜索可能会搜索不到用户实际想要的搜索结果，因此，为了提供准确的搜索结果，服务器一般采用的搜索方法为，将用户输入的关键词进行归一化操作，使得归一化之后的关键词更加规范，使用归一化之后的关键词进行搜索，并提供搜索结果。

[0004] 其中，为了达到更好的搜索效果，服务器一般还会根据用户输入的关键词，查找该关键词对应的各个相关词，并把查找到的各个相关词提供给用户。采用提供相关词的方法，用户在未得到其满意的搜索结果时，就可以直接点击某个相关词，使用该相关词进行搜索，这样可以进一步提高搜索的准确性。

[0005] 在现有技术中，服务器提供相关词的方法如图 1 所示。图 1 为现有技术中提供相关词的过程，具体包括以下步骤：

[0006] S101：采用设定的分词算法，将用户输入的关键词拆分为若干个分词。

[0007] S102：根据每个分词的属性，在保存的各相关词中，确定该关键词的各备选相关词。

[0008] S103：查找搜索日志中记录的通过该关键词进行搜索得到的搜索结果中被点击的每个第一搜索结果，以及点击的次数。

[0009] S104：针对确定的每个备选相关词，查找搜索日志中记录的通过该备选相关词进行搜索得到的搜索结果中被点击的每个第二搜过结果，以及点击的次数。

[0010] S105：确定每个第一搜索结果和每个第二搜索结果中相同的各搜索结果，根据相同的各搜索结果对应关键词的被点击次数，以及对应该备选相关词的被点击次数，计算该关键词与该备选相关词的相关性分数。

[0011] S106：根据确定的该关键词与每个备选相关词的相关性分数，选择相关性分数较高的备选相关词，作为该关键词的相关词提供给用户。

[0012] 然而，采用现有技术中提供相关词的方法时，如果用户输入的关键词是新的关键词，其并未记录在搜索日志中，或者，如果在搜索日志中并未查找到该关键词对应的每个第一搜索结果与各备选相关词对应的每个第二搜索结果中有相同的搜索结果，则不能计算该关键词与各个备选相关词的相关性分数，因此提供的相关词不够准确，导致用户需要输入其他类似的相关词再次进行搜索，消耗了大量的服务器资源。

发明内容

[0013] 本申请实施例提供一种提供相关词的方法及装置，用以解决现有技术中提供的相关词不够准确，导致用户需要输入其他类似的相关词再次进行搜索，消耗了大量的服务器资源的问题。

[0014] 本申请实施例提供的一种提供相关词的方法，包括：

[0015] 根据用户输入的关键词，确定所述关键词的各备选相关词；

[0016] 针对确定的各备选相关词，确定所述关键词与该备选相关词在设定的每个特征上的特征分数，将确定的每个特征分数作为输入参数值输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，其中，所述相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的；

[0017] 根据得到的所述关键词与各备选相关词的相关性分数，在各备选相关词中选择提供给所述用户的相关词。

[0018] 本申请实施例提供的一种提供相关词的装置，包括：

[0019] 备选相关词确定模块，用于根据用户输入的关键词，确定所述关键词的各备选相关词；

[0020] 相关性分数确定模块，用于针对确定的各备选相关词，确定所述关键词与该备选相关词在设定的每个特征上的特征分数，将确定的每个特征分数作为输入参数值输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，其中，所述相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的；

[0021] 相关词提供模块，用于根据得到的所述关键词与各备选相关词的相关性分数，在各备选相关词中选择提供给所述用户的相关词。

[0022] 本申请实施例提供一种提供相关词的方法及装置，该方法针对用户输入的关键词的各备选相关词，将该关键词与该备选相关词在设定的每个特征上的特征分数输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，并据此提供相关词，其中，该相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的。通过上述方法，即使用户输入的关键词未记录在搜索日志中，也可以通过将该关键词与各备选相关词的特征分数输入相关性分数计算模型，来获得该关键词与各备选相关词的相关性分数，从而为用户提供准确的相关词，使用户无需再次进行搜索，节省了服务器资源。

附图说明

[0023] 图 1 为现有技术中提供相关词的过程；

[0024] 图 2 为本申请实施例提供的提供相关词的过程；

[0025] 图 3 为本申请实施例提供的确定相关性分数计算模型的过程；

[0026] 图 4 为本申请实施例提供的分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果示意图；

[0027] 图 5 为本申请实施例提供的提供相关词的装置结构示意图。

具体实施方式

[0028] 由于现有技术中服务器在为用户提供相关词时，主要是通过查找搜索日志中记录的分别通过该用户输入的关键词和备选相关词进行搜索时得到的相同的搜索结果，根据查

找到的各相同的搜索结果对应该关键词的被点击的次数和对应该备选相关词的被点击的次数，确定该关键词与该备选相关词的相关性分数，因此，如果搜索日志中未记录该用户输入的关键词，或者，搜索日志中记录的分别通过该关键词和备选相关词进行搜索时不存在相同的搜索结果，则现有技术无法计算该关键词与该备选相关词的相关性分数，导致提供的相关词不够准确。

[0029] 本申请实施例为了提高提供的相关词的准确性，服务器预先根据设定数量的已计算出相关性分数的关键词与相关词，确定相关性分数计算模型。在提供相关词时，确定用户输入的关键词与备选相关词在设定的每个特征上的特征分数，将该特征分数作为输入参数值输入到相关性分数计算模型中，得到的就是该关键词与该备选相关词的相关性分数，因此即使搜索日志中未记录该用户输入的关键词，或者，搜索日志中记录的分别通过该关键词和备选相关词进行搜索时不存在相同的搜索结果，服务器仍然可以通过确定用户输入的关键词与备选相关词的特征分数，确定二者的相关性分数，从而为用户提供准确的相关词。

[0030] 下面结合说明书附图，对本申请实施例进行详细描述。

[0031] 图 2 为本申请实施例提供的提供相关词的过程，具体包括以下步骤：

[0032] S201：根据用户输入的关键词，确定该关键词的各备选相关词。

[0033] 在本申请实施例中，服务器可以采用与现有技术类似的方法，确定用户输入的关键词的各备选相关词。

[0034] S202：针对确定的各备选相关词，确定该关键词与该备选相关词在设定的每个特征上的特征分数，将确定的每个特征分数作为输入参数值输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数。

[0035] 其中，该相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的。

[0036] 在本申请实施例中，服务器向用户提供相关词时，根据设定的每个特征，确定该用户输入的关键词与备选相关词在该每个特征上的特征分数，并基于确定的相关性分数计算模型，将确定的特征分数作为输入参数值输入到相关性分数计算模型中，得到的计算结果即为该关键词与该备选相关词的相关性分数。

[0037] S203：根据得到的该关键词与各备选相关词的相关性分数，在各备选相关词中选择提供给该用户的相关词。

[0038] 其中，可以选择相关性分数大于某个阈值的所有备选相关词作为该关键词的相关词提供给用户，也可以选择相关性分数较大的设定个数的备选相关词提供给用户。

[0039] 在本申请实施例中，服务器确定相关性分数计算模型的过程如图 3 所示。图 3 为本申请实施例提供的确定相关性分数计算模型的过程，具体包括以下步骤：

[0040] S301：确定已计算出相关性分数的关键词与相关词作为训练样本，选择设定数量的训练样本。

[0041] 在本申请实施例中，服务器查找已计算出相关性分数的关键词和相关词，也即查找已知相关性分数的关键词和相关词，其中，查找到的关键词和相关词的相关性分数是通过现有技术中计算相关性分数的算法可以计算得到的，例如通过现有技术中的 SimRank 算法或者 CosRank 算法可以计算得到的相关性分数。

[0042] 服务器将查找到的已知相关性分数的关键词和相关词作为训练样本，也即，在一

个训练样本中包括一个关键词和一个相关词，且二者的相关性分数是已知的。服务器选择设定数量的训练样本用于在后续步骤中得到相关性分数计算模型，其中，该设定数量可以根据需要进行设定，该设定数量越大，后续得到的相关性计算模型的准确性越高。

[0043] S302：针对选择的每个训练样本，根据设定的每个特征，确定该训练样本中的关键词和相关词在每个特征上的特征分数，将已计算出的该训练样本中的关键词和相关词的相关性分数确定为目标值，将确定的该训练样本中的关键词和相关词在每个特征上的特征分数确定为输入参数值。

[0044] 在本申请实施例中，服务器针对每个训练样本，将该训练样本中的关键词和相关词在设定的每个特征上的相似度量化，作为该训练样本中的关键词和相关词在每个特征上的特征分数，其中，设定的每个特征可以根据需要进行设定，特征越多，后续得到的相关性分数计算模型的准确性越高。本申请实施例中设定的每个特征包括：分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度，以及，分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度，以及，该训练样本中的关键词和相关词的编辑相似度，以及，分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度。当然，也可以是上述四种特征中的一种或几种特征的组合。

[0045] 下面以搜索商品信息为例进行说明。

[0046] 当搜索商品信息时，各训练样本中的关键词和相关词均是为了搜索某个商品的关键词和相关词，则上述特征具体如下：

[0047] 分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度具体为：分别采用关键词和相关词进行搜索，得到的商品信息在商品类目上的相似度。例如，该训练样本中的关键词为：A 品牌 a 型号手机，相关词为：A 品牌 b 型号手机，采用关键词（A 品牌 a 型号手机）进行搜索得到的搜索结果为：手机 1、手机 2、手机 3，采用相关词（A 品牌 b 型号手机）进行搜索得到的搜索结果为：手机 4、手机 5、手机 1，则该特征即为这两个搜索结果在商品类目上的相似度。

[0048] 分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度具体为：分别采用关键词和相关词进行搜索，得到的商品信息在商品属性上的相似度。继续沿用上例，该特征即为这两个搜索结果在商品属性上的相似度。

[0049] 该训练样本中的关键词和相关词的编辑相似度具体为：关键词与相关词的编辑距离。继续沿用上例，该特征即为关键词（A 品牌 a 型号手机）与相关词（A 品牌 b 型号手机）的编辑距离。

[0050] 分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度具体为：分别采用关键词和相关词进行搜索，得到的相同的商品信息分别对应关键词和相关词被点击的次数的相似度。继续沿用上例，这两个搜索结果中相同的搜索结果为：手机 1，该特征即为通过关键词（A 品牌 a 型号手机）搜索时，手机 1 被点击的次数，与通过相关词（A 品牌 b 型号手机）搜索时，手机 1 被点击的次数的相似度。

[0051] 量化上述相似度得到该训练样本中的关键词和相关词在每个特征上的特征分数，将该特征分数确定为输入参数值，将已知的二者的相关性分数作为目标值。

[0052] S303：根据针对每个训练样本确定的目标值和输入参数值，采用设定的算法进行

回归运算，得到相关性分数计算模型。

[0053] 在本申请实施例中，回归运算后得到的相关性分数计算模型满足：对于任意训练样本，将确定的该训练样本中的关键词和相关词在每个特征上的特征分数作为输入参数值输入到该模型中后，得到的计算结果与已知的该训练样本中的关键词和相关词的相关性分数相同。也即，根据针对每个训练样本确定的目标值和输入参数值，拟合出将输入参数值输入后，能够得到相应目标值的模型，作为相关性分数计算模型。并且，可以采用支持向量机(SVM)算法进行回归运算，也可以采用评定模型(Logit)算法进行回归运算，当然也可以采用其他回归算法进行回归运算。

[0054] 在上述过程中，服务器预先选择设定数量的已知相关性分数的关键词和相关词作为训练样本，针对每个训练样本，量化该训练样本中的关键词和相关词在设定的每个特征上的相似度，也即确定该训练样本中的关键词和相关词在每个特征上的特征分数，以已知的该训练样本的相关性分数作为目标值，以确定的每个特征分数作为输入参数值，根据针对每个训练样本确定的目标值和输入参数值，采用设定的算法进行回归运算，得到相关性分数计算模型，使得到的相关性分数计算模型满足，将针对任意训练样本确定的输入参数值输入该模型后，得到的结果与已知的该训练样本的目标值相同。在提供相关词时，则可以将用户输入的关键词与备选相关词在每个特征上的特征分数输入到相关性分数计算模型中，得到该用户输入的关键词与备选相关词的相关性分数，并据此提供相关词。因此即使用户输入的关键词未记录在搜索日志中，或者，搜索日志中记录的分别通过用户输入的关键词和备选相关词进行搜索时不存在相同的搜索结果，本申请实施例提供的方法也可以准确的确定出用户输入的关键词与备选相关词的相关性分数，从而可以据此为用户提供准确的相关词，使用户无需再次进行搜索，节省了服务器资源。

[0055] 在上述图3所示的步骤S302中，服务器需要量化该训练样本中的关键词和相关词在每个特征上的相似度，得到相应的特征分数。也即，确定下述各特征分数：

[0056] 确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数；以及

[0057] 确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数；以及

[0058] 确定该训练样本中的关键词和相关词的编辑距离作为编辑距离分数；以及

[0059] 确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

[0060] 其中，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数的方法具体为：采用该训练样本中的关键词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索类目的搜索结果的个数与搜索结果总数的比值；将采用该训练样本中的关键词进行搜索，针对每个搜索结果确定的每个比值构成的向量确定为关键词类目向量；采用该训练样本中的相关词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索结果类目的搜索结果的个数与搜索结果总数的比值；将采用该训练样本中的相关词进行搜索，针对每个搜索结果确定的每个比值构成的向量确定为相关词类目向量；确定关键词类

目向量与相关词类目向量的余弦值，将该余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数。

[0061] 例如，搜索结果类目共有四个，分别为：类目 1、类目 2、类目 3、类目 4，采用该训练样本中的关键词进行搜索，得到的搜索结果总数为 N，这 N 个搜索结果中，属于类目 1 的搜索结果的个数为 n1，则针对类目 1 确定的比值为 n1/N，相应的，属于类目 2 的搜索结果的个数为 n2，则针对类目 2 确定的比值为 n2/N，属于类目 3 的搜索结果的个数为 n3，则针对类目 3 确定的比值为 n3/N，属于类目 4 的搜索结果的个数为 n4，则针对类目 4 确定的比值为 n4/N，其中，n1、n2、n3、n4 的和值为 N，n1、n2、n3、n4 均为不小于 0 且不大于 N 的正整数。则采用该训练样本中的关键词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量即为 (n1/N,n2/N,n3/N,n4/N)，该向量即为关键词类目向量。相应的，采用该训练样本中的相关词进行搜索，得到的搜索结果总数为 M，其中属于类目 1、类目 2、类目 3、类目 4 的搜索结果的个数分别为 m1、m2、m3、m4，则针对每个搜索结果类目确定的比值分别为 m1/N、m2/N、m3/N、m4/N，采用该训练样本中的相关词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量即为 (m1/N,m2/N,m3/N,m4/N)，该向量即为相关词类目向量。因此，确定向量 (n1/N,n2/N,n3/N,n4/N) 与向量 (m1/N,m2/N,m3/N,m4/N) 的余弦值，将该余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数。

[0062] 确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数的方法具体为：根据该训练样本中的关键词，确定采用该训练样本中的关键词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第一集合；根据该训练样本中的相关词，确定采用该训练样本中的相关词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第二集合；确定第一集合与第二集合的交集以及并集，确定该交集中包含的元素的个数与该并集中包含的元素的个数的比值，将该比值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数。

[0063] 例如，采用该训练样本中的关键词进行搜索所得到的搜索结果对应的搜索结果属性包括两种，分别为：属性 1、属性 2，则以这两种搜索结果属性为元素构成的第一集合为 { 属性 1，属性 2}。相应的，采用该训练样本中的相关词进行搜索所得到的搜索结果对应的搜索结果属性也包括两种，分别为属性 2、属性 3，则以这两种搜索结果属性为元素构成的第二集合为 { 属性 2，属性 3}。第一集合与第二集合的交集为 { 属性 2}，并集为 { 属性 1，属性 2，属性 3}，可见交集中包含的元素的个数为 1，并集中包含的元素的个数为 3，二者的比值为 1/3，则将该比值 1/3 确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数。

[0064] 确定该训练样本中的关键词和相关词的编辑距离作为编辑距离分数的方法具体为：确定将该关键词变更为该相关词所需要的操作次数，作为该关键词与相关词的编辑距离，其中，将该关键词中的一个字符删除的操作，以及，向该关键词中添加一个字符的操作为一次操作。

[0065] 例如，该训练样本中的关键词为：A 品牌 a 型号手机，相关词为：A 品牌 b 型号手机，则要将该关键词变更为该相关词所要做的操作为：将关键词中的字符“a”删除，在该关键

词中添加字符“b”，因此操作次数为 2 次，也即该训练样本中的关键词和相关词的编辑距离为 2，编辑距离分数为 2。

[0066] 确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数的方法具体为：确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的每个相同的搜索结果；针对每个相同的搜索结果，根据搜索日志中的记录，确定通过该训练样本中的关键词进行搜索时该搜索结果被点击的次数，确定通过该训练样本中的相关词进行搜索时该搜索结果被点击的次数；将通过该训练样本中的关键词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量确定为关键词点击向量；将通过该训练样本中的相关词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量确定为相关词点击向量；确定关键词点击向量与相关词点击向量的余弦值，将该余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

[0067] 例如，如图 4 所示，图 4 为本申请实施例提供的分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果示意图。在图 4 中，采用该训练样本中的关键词进行搜索，得到的搜索结果为：结果 1、结果 2、结果 3、结果 4，采用该训练样本中的相关词进行搜索，得到的搜索结果为：结果 2、结果 3、结果 5、结果 6。则相同的搜索结果为：结果 2、结果 3。针对结果 2，根据搜索日志中的记录，确定通过该关键词搜索到结果 2 时，该结果 2 被点击的次数 i2，确定通过该相关词搜索到结果 2 时，该结果 2 被点击的次数 j2。针对结果 3，根据搜索日志中的记录，确定通过该关键词搜索到结果 3 时，该结果 3 被点击的次数 i3，确定通过该相关词搜索到结果 3 时，该结果 3 被点击的次数 j3。则通过该训练样本中的关键词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量为 (i2，i3)，该向量即为关键词点击向量。通过该训练样本中的相关词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量为 (j2，j3)，该向量即为相关词点击向量。确定关键词点击向量 (i2，i3) 与关键词点击向量 (j2，j3) 的余弦值，将该余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

[0068] 上述已经确定了训练样本中的关键词和相关词的每个特征分数，在后续的步骤中，则可以针对每个训练样本确定的特征分数，以及每个训练样本已知的相关性分数，采用设定的算法进行回归运算得到相关性分数计算模型，并基于得到的相关性分数计算模型，确定用户输入的关键词与各备选相关词的相关性分数，并据此向用户提供相关词。

[0069] 在本申请实施例中，还可以在上述四个设定的特征的基础上，根据需要增加更细粒度的其他特征，例如当搜索商品信息时，还可以在上述四个特征的基础上，增加分别采用该训练样本中的关键词和相关词进行搜索，所得到的商品信息在品牌上的相似度分数、在型号上的相似度分数、在商品颜色上的相似度分数、在商品质地上的相似度分数等等，这里就不再一一赘述。

[0070] 图 5 为本申请实施例提供的提供相关词的装置结构示意图，具体包括：

[0071] 备选相关词确定模块 501，用于根据用户输入的关键词，确定所述关键词的各备选相关词；

[0072] 相关性分数确定模块 502，用于针对确定的各备选相关词，确定所述关键词与该备

选相关词在设定的每个特征上的特征分数，将确定的每个特征分数作为输入参数值输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，其中，所述相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的；

[0073] 相关词提供模块 503，用于根据得到的所述关键词与各备选相关词的相关性分数，在各备选相关词中选择提供给所述用户的相关词。

[0074] 所述相关性分数确定模块 502 包括：

[0075] 确定选择子模块 5021，用于确定已计算出相关性分数的关键词与相关词作为训练样本，选择设定数量的训练样本；

[0076] 特征分数确定子模块 5022，用于针对选择的每个训练样本，根据所述设定的每个特征，确定该训练样本中的关键词和相关词在每个特征上的特征分数，将已计算出的该训练样本中的关键词和相关词的相关性分数确定为目标值，将确定的该训练样本中的关键词和相关词在每个特征上的特征分数确定为输入参数值；

[0077] 模型确定子模块 5023，用于根据针对每个训练样本确定的目标值和输入参数值，采用设定的算法进行回归运算，得到相关性分数计算模型。

[0078] 所述特征分数确定子模块 5022 具体用于，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数，以及，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数，以及，确定该训练样本中的关键词和相关词的编辑距离作为编辑距离分数，以及，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

[0079] 所述特征分数确定子模块 5022 具体用于，采用该训练样本中的关键词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索类目的搜索结果个数与搜索结果总数的比值，将采用该训练样本中的关键词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量确定为关键词类目向量；采用该训练样本中的相关词进行搜索，针对每个搜索结果类目，确定所得到的属于该搜索结果类目的搜索结果的个数，以及得到的搜索结果总数，确定属于该搜索结果类目的搜索结果个数与搜索结果总数的比值，将采用该训练样本中的相关词进行搜索，针对每个搜索结果类目确定的每个比值构成的向量确定为相关词类目向量；确定所述关键词类目向量与所述相关词类目向量的余弦值，将所述余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果类目上的相似度分数。

[0080] 所述特征分数确定子模块 5022 具体用于，根据该训练样本中的关键词，确定采用该训练样本中的关键词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第一集合；根据该训练样本中的相关词，确定采用该训练样本中的相关词进行搜索所得到的搜索结果对应的每个属性，以确定的每个属性为元素构成第二集合；确定所述第一集合与第二集合的交集以及并集，确定所述交集中包含的元素的个数与所述并集中包含的元素的个数的比值，将所述比值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果属性上的相似度分数。

[0081] 所述特征分数确定子模块 5022 具体用于，确定分别采用该训练样本中的关键词和相关词进行搜索，所得到的每个相同的搜索结果；针对每个相同的搜索结果，根据搜索日

志中的记录，确定通过该训练样本中的关键词进行搜索时该搜索结果被点击的次数，确定通过该训练样本中的相关词进行搜索时该搜索结果被点击的次数；将通过该训练样本中的关键词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量确定为关键词点击向量；将通过该训练样本中的相关词进行搜索时，针对每个相同的搜索结果确定的每个被点击的次数构成的向量确定为相关词点击向量；确定所述关键词点击向量与所述相关词点击向量的余弦值，将所述余弦值确定为分别采用该训练样本中的关键词和相关词进行搜索，所得到的搜索结果在搜索结果点击上的相似度分数。

[0082] 所述模型确定子模块 5022 具体用于，采用支持向量机 SVM 算法进行回归运算，得到相关性分数计算模型，或者，采用评定模型 Logit 算法进行回归运算，得到相关性分数计算模型。

[0083] 具体的上述提供相关词的装置可以位于服务器中。

[0084] 本申请实施例提供一种提供相关词的方法及装置，该方法针对用户输入的关键词的各备选相关词，将该关键词与该备选相关词在设定的每个特征上的特征分数输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数，并据此提供相关词，其中，该相关性分数计算模型为根据设定数量的已计算出相关性分数的关键词与相关词确定的。通过上述方法，即使用户输入的关键词未记录在搜索日志中，也可以通过将该关键词与各备选相关词的特征分数输入相关性分数计算模型，来获得该关键词与各备选相关词的相关性分数，从而为用户提供准确的相关词，使用户无需再次进行搜索，节省了服务器资源。

[0085] 显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

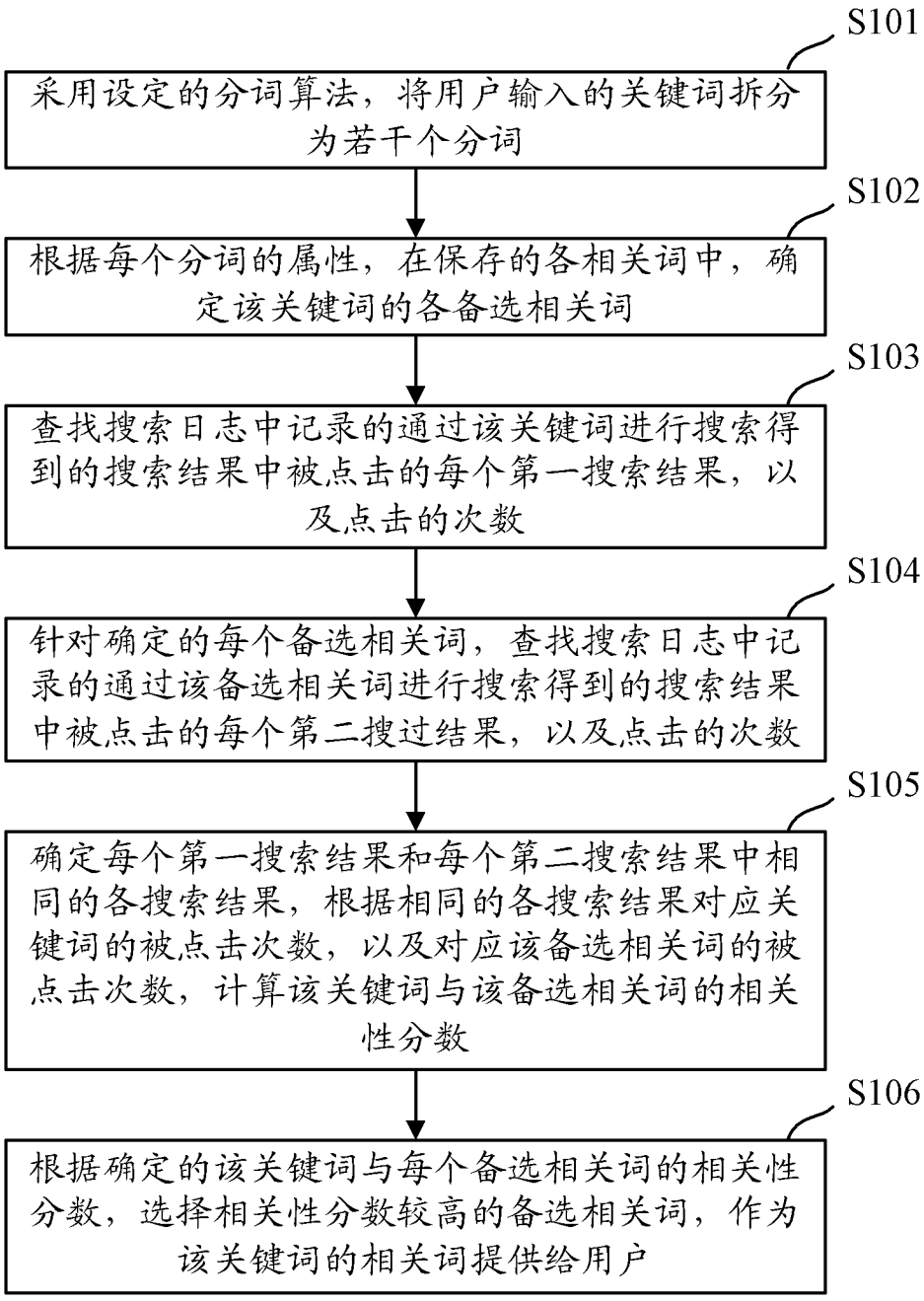

S101
采用设定的分词算法，将用户输入的关键词拆分为若干个分词
S102
根据每个分词的属性，在保存的各相关词中，确定该关键词的各备选相关词
S103
查找搜索日志中记录的通过该关键词进行搜索得到的搜索结果中被点击的每个第一搜索结果，以及点击的次数
S104
针对确定的每个备选相关词，查找搜索日志中记录的通过该备选相关词进行搜索得到的搜索结果中被点击的每个第二搜过结果，以及点击的次数
S105
确定每个第一搜索结果和每个第二搜索结果中相同的各搜索结果，根据相同的各搜索结果对应关键词的被点击次数，以及对应该备选相关词的被点击次数，计算该关键词与该备选相关词的相关性分数
S106
根据确定的该关键词与每个备选相关词的相关性分数，选择相关性分数较高的备选相关词，作为该关键词的相关词提供给用户

图 1

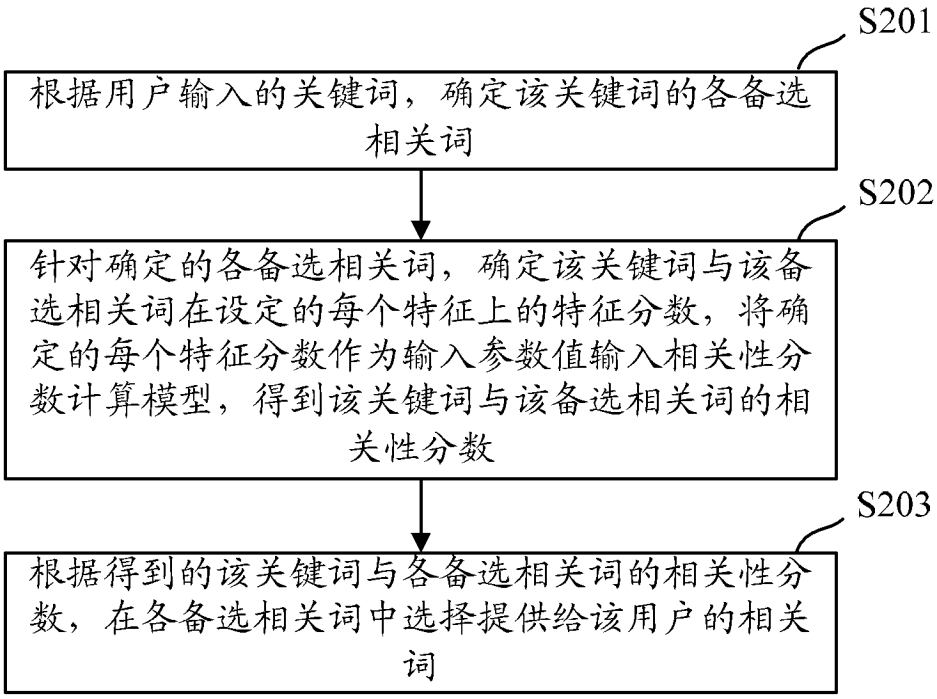

S201
根据用户输入的关键词，确定该关键词的各备选相关词
S202
针对确定的各备选相关词，确定该关键词与该备选相关词在设定的每个特征上的特征分数，将确定的每个特征分数作为输入参数值输入相关性分数计算模型，得到该关键词与该备选相关词的相关性分数
S203
根据得到的该关键词与各备选相关词的相关性分数，在各备选相关词中选择提供给该用户的相关词

图 2

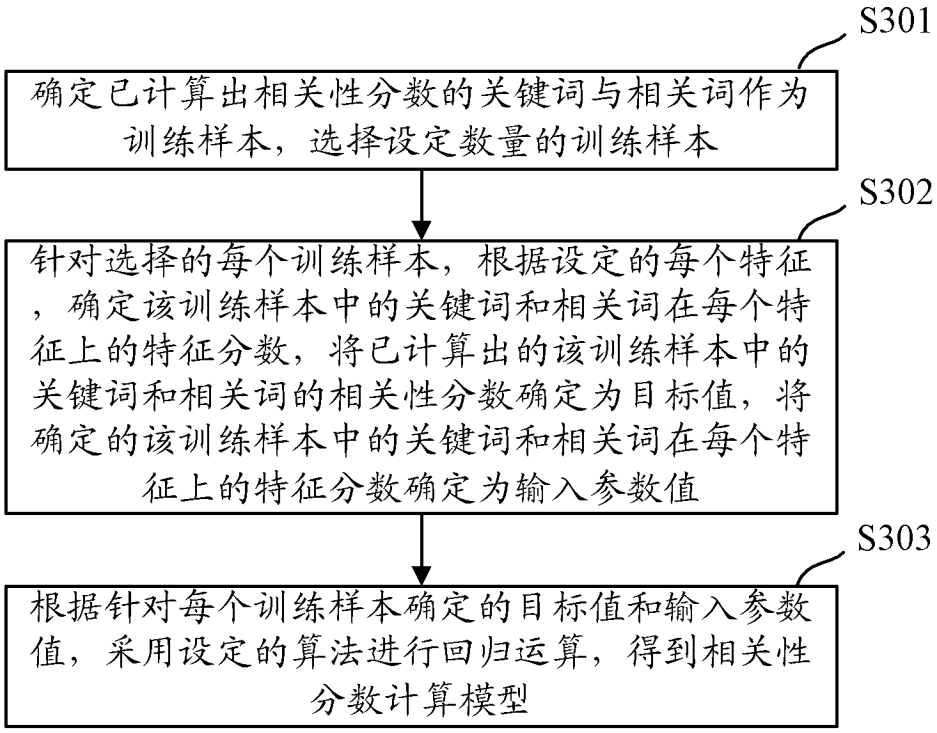

S301
确定已计算出相关性分数的关键词与相关词作为训练样本，选择设定数量的训练样本
S302
针对选择的每个训练样本，根据设定的每个特征，确定该训练样本中的关键词和相关词在每个特征上的特征分数，将已计算出的该训练样本中的关键词和相关词的相关性分数确定为目标值，将确定的该训练样本中的关键词和相关词在每个特征上的特征分数确定为输入参数值
S303
根据针对每个训练样本确定的目标值和输入参数值，采用设定的算法进行回归运算，得到相关性分数计算模型

图 3

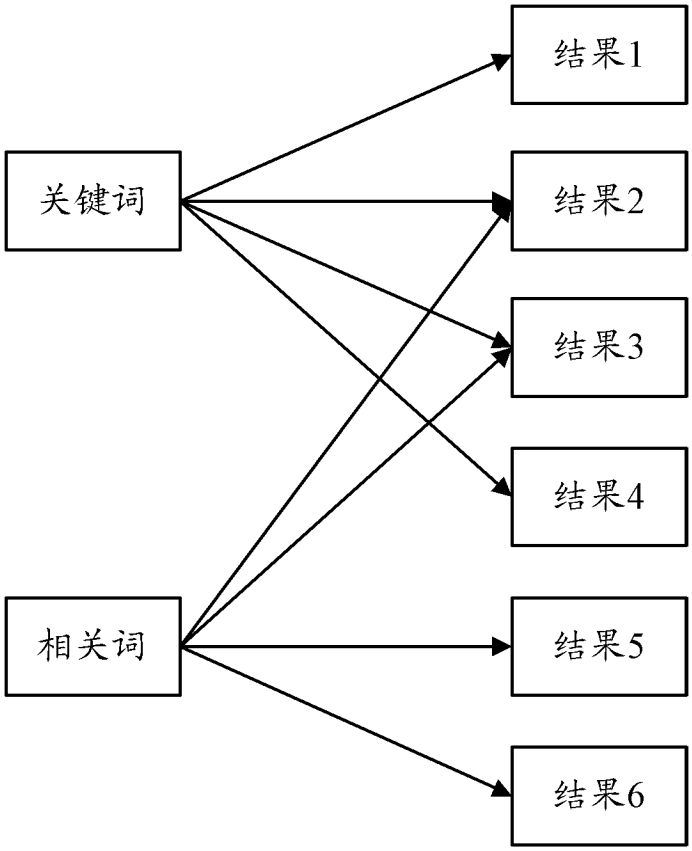

关键词
相关词
结果1
结果2
结果3
结果4
结果5
结果6

图 4

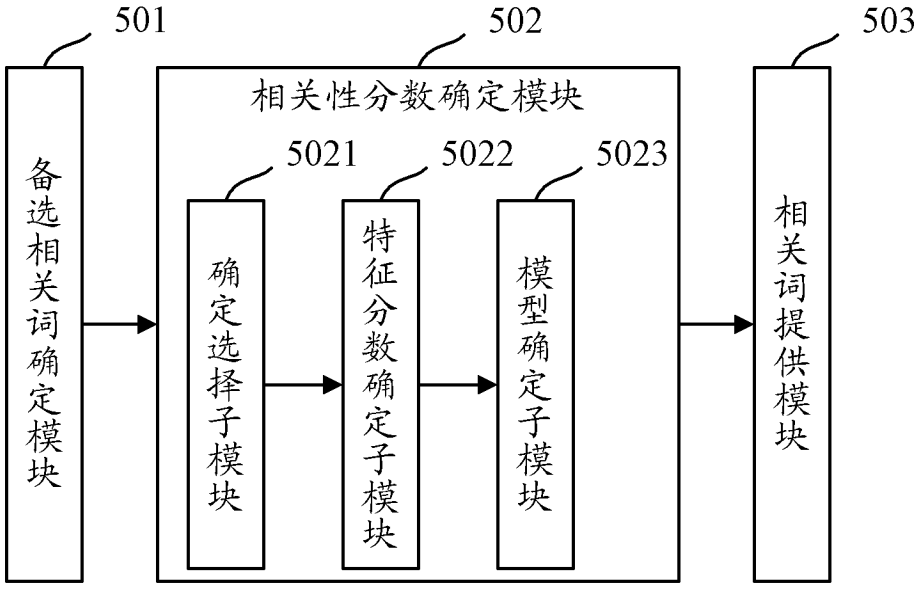

501
502
503
备选相关词确定模块
相关性分数确定模块
5021
5022
5023
确定选择子模块
特征分数确定子模块
模型确定子模块
相关词提供模块

图 5