(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 103853722 A
(43)申请公布日 2014.06.11

(21)申请号 201210497241.2

(22)申请日 2012.11.29

(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 王艳敏 王迪 赫南 张文斌 胡立新 刘小兵 胡景贺 朱建朋

(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 张驰 宋志强

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

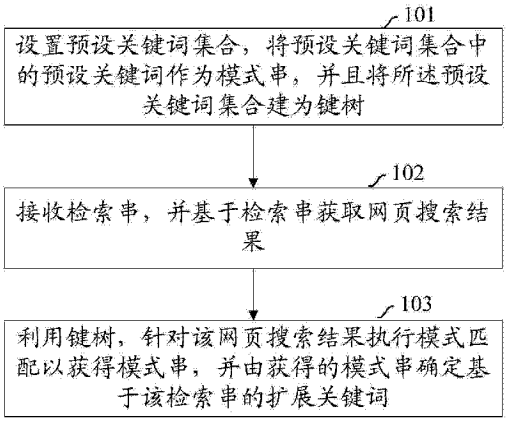
权利要求书3页 说明书10页 附图5页

(54)发明名称

一种基于检索串的关键词扩展方法、装置和系统

(57)摘要

本发明实施方式提出了一种基于检索串的关键词扩展方法、装置和系统。方法包括：设置预设关键词集合，将预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树；接收检索串，并基于检索串获取网页搜索结果；利用键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的模式串确定基于该检索串的扩展关键词。本发明实施方式丰富了关键词匹配结果，扩展了检索串的关键词，提高了展示内容的全面性。而且，量化了检索串与关键词之间的相似度，保证了展示内容的相关性。



CN 103853722 A

1. 一种基于检索串的关键词扩展方法，其特征在于，该方法包括：

设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树；

接收检索串，并基于所述检索串获取网页搜索结果；

利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

2. 根据权利要求 1 所述的关键词扩展方法，其特征在于，所述针对该网页搜索结果执行模式匹配以获得模式串包括：

针对该网页搜索结果执行多模式匹配以获得多个模式串，或者针对该网页搜索结果执行单模式匹配以获得单个模式串。

3. 根据权利要求 1 所述的关键词扩展方法，其特征在于，该方法进一步包括：

从所述检索串本身提取扩展关键词；

将所述由模式串确定的基于该检索串的扩展关键词以及从所述检索串本身提取的扩展关键词相聚合，以获得扩展关键词集合。

4. 根据权利要求 1 所述的关键词扩展方法，其特征在于，该方法进一步包括：利用检索串变换方式获取扩展关键词；

将所述由模式串确定的基于该检索串的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

5. 根据权利要求 1 所述的关键词扩展方法，其特征在于，该方法进一步包括：

从所述检索串本身提取扩展关键词，以及利用检索串变换方式获取扩展关键词；

将所述由模式串确定的基于该检索串的扩展关键词、从所述检索串本身提取的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

6. 根据权利要求 5 所述的关键词扩展方法，其特征在于，该方法进一步包括：

从所述扩展关键词集合中的扩展关键词和检索串，分别提取至少两个比较特征，所述比较特征包括文本特征、分类特征或语义特征；

基于所述扩展关键词集合中的扩展关键词和检索串的每个比较特征，计算所述扩展关键词集合中的扩展关键词和检索串之间的每个比较特征的相关性；

根据逻辑回归模型对各个比较特征的相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；

基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

7. 根据权利要求 6 所述的关键词扩展方法，其特征在于，

所述文本特征包括公共子串、共有语素或编辑距离；所述语义特征包括标题域的语义特征、摘要域的语义特征、标题域和摘要域整合后的语义特征、去掉检索串和关键词共有语素表示后标题域的语义特征、去掉检索串和关键词的共有语素后摘要域的语义特征及去掉检索串和关键词的共有语素后标题域和摘要域整合后的语义特征。

8. 一种基于检索串的关键词扩展装置，其特征在于，该装置包括键树建立单元、搜索结果获取单元和关键词扩展单元，其中：

键树建立单元，用于设置预设关键词集合，将所述预设关键词集合中的预设关键词作

为模式串，并且将所述预设关键词集合建为键树；

搜索结果获取单元，用于接收检索串，并基于所述检索串获取网页搜索结果；

关键词扩展单元，用于利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

9. 根据权利要求 8 所述的基于检索串的关键词扩展装置，其特征在于，

关键词扩展单元，用于针对该网页搜索结果执行多模式匹配以获得多个模式串，或者针对该网页搜索结果执行单模式匹配以获得单个模式。

10. 根据权利要求 8 所述的基于检索串的关键词扩展装置，其特征在于，进一步包括关键词提取单元和关键词聚合单元；

关键词提取单元，用于从所述检索串本身提取扩展关键词；

关键词聚合单元，用于将所述由模式串确定的基于该检索串的扩展关键词以及从所述检索串本身提取的扩展关键词相聚合，以获得扩展关键词集合。

11. 根据权利要求 8 所述的基于检索串的关键词扩展装置，其特征在于，进一步包括检索串变换单元和关键词聚合单元；

检索串变换单元，用于利用检索串变换方式获取扩展关键词；

关键词聚合单元，用于将所述由模式串确定的基于该检索串的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

12. 根据权利要求 8 所述的基于检索串的关键词扩展装置，其特征在于，进一步包括关键词提取单元、检索串变换单元和关键词聚合单元；其中：

关键词提取单元，用于从所述检索串本身提取扩展关键词；

检索串变换单元，用于利用检索串变换方式获取扩展关键词；

关键词聚合单元，用于将所述由模式串确定的基于该检索串的扩展关键词、从所述检索串本身提取的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

13. 根据权利要求 12 所述的基于检索串的关键词扩展装置，其特征在于，进一步包括相关性指标确定单元，其中：

相关性指标确定单元，用于从所述扩展关键词集合中的扩展关键词和检索串，分别提取至少两个比较特征，所述比较特征包括文本特征、分类特征或语义特征；基于所述扩展关键词集合中的扩展关键词和检索串的每个比较特征，计算所述扩展关键词集合中的扩展关键词和检索串之间的每个比较特征的相关性；根据逻辑回归模型对各个比较特征的相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

14. 一种基于检索串的关键词扩展系统，其特征在于，包括客户端、搜索引擎和关键词扩展装置，其中：

客户端，用于接收检索串，并基于所述检索串向搜索引擎查询网页搜索结果；

搜索引擎，用于向客户端提供对应于检索串的网页搜索结果；

关键词扩展装置，用于设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树，利用所述键树，针对该网页搜索结果执

行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

15. 根据权利要求 14 所述的基于检索串的关键词扩展系统，其特征在于，

关键词扩展装置，进一步用于从所述检索串本身提取扩展关键词，以及利用检索串变换方式获取扩展关键词；将所述由模式串确定的基于该检索串的扩展关键词、从所述检索串本身提取的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

16. 根据权利要求 15 所述的基于检索串的关键词扩展系统，其特征在于，

关键词扩展装置，进一步用于从所述扩展关键词集合中的扩展关键词和检索串，分别提取至少两个比较特征，所述比较特征包括文本特征、分类特征或语义特征；基于所述扩展关键词集合中的扩展关键词和检索串的每个比较特征，计算所述扩展关键词集合中的扩展关键词和检索串之间的每个比较特征的相关性；根据逻辑回归模型对各个比较特征的相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

一种基于检索串的关键词扩展方法、装置和系统

技术领域

[0001] 本发明实施方式涉及信息处理技术领域，更具体地，涉及一种基于检索串的关键词扩展方法、装置和系统。

背景技术

[0002] 在当今的信息时代中，各种信息设备应运而生：有用于话音传输的固定电话、移动电话；有用于信息资源共享、处理的服务器和个人电脑；有用于视频数据显示的各种电视机等等。这些设备都是在特定领域内为解决实际的需求而产生的。随着电子消费、计算机、通信(3C)融合的到来，人们越来越多地将注意力放到了对各个不同领域的信息设备进行综合利用的研究上，以充分利用现有资源设备来为人们更好的服务。

[0003] 搜索引擎广告就是一种针对信息综合利用的示范性应用。搜索引擎广告也称为关键词广告，是指广告主根据自己的产品或服务的内容、特点等确定关键词，撰写广告内容并自主定价投放在搜索引擎上的广告。当用户搜索的内容与广告主的关键词匹配时，相应的广告就可能会被展示，并在用户点击后按照广告主对该关键词的出价收费。

[0004] 搜索引擎广告展示的过程概括如下：广告主向搜索引擎广告系统提交有效关键词，连同出价、广告物料(标题、描述)等信息以倒排索引的形式，加载到匹配系统中，匹配系统对用户提交的检索串(query)进行在线分析，找到对应各种匹配类型的关键词；再通过关键词的倒排信息，完成后续的广告拉取、精选、排序等竞价排名过程，最终展示给用户。

[0005] 然而，现有技术中仅从检索串字面抽取关键词，所获得的关键词数量有限，因此难以保证展示内容的全面性。

[0006] 而且，现有技术中所提取的关键词与检索串仅局限于文本上的关联，其他语义上相关的关键词很难被找到。从检索串本身抽取出来的关键词往往是检索串的一部分，两者字面上虽然有一定的相关性，然而很难保证意图上的一致。比如从检索串“北京鲜花快递哪里最便宜”里可以找到关键词“最便宜”，但是这两者的意图不完全匹配。如果直接用关键词“最便宜”去查询展示内容，容易展示出与用户本意相差较多的展示内容，从而导致展示内容并不相关。

发明内容

[0007] 本发明实施方式提出一种基于检索串的关键词扩展方法，从而扩展关键词，提高展示内容的全面性。

[0008] 本发明实施方式提出一种基于检索串的关键词扩展装置，从而扩展关键词，提高展示内容的全面性。

[0009] 本发明实施方式提出一种基于检索串的关键词扩展系统，从而扩展关键词，提高展示内容的全面性。

[0010] 本发明实施方式的具体方案如下：

[0011] 一种基于检索串的关键词扩展方法，该方法包括：

[0012] 设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树；

[0013] 接收检索串，并基于所述检索串获取网页搜索结果；

[0014] 利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

[0015] 一种基于检索串的关键词扩展装置，该装置包括键树建立单元、搜索结果获取单元和关键词扩展单元，其中：

[0016] 键树建立单元，用于设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树；

[0017] 搜索结果获取单元，用于接收检索串，并基于所述检索串获取网页搜索结果；

[0018] 关键词扩展单元，用于利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

[0019] 一种基于检索串的关键词扩展系统，其特征在于，包括客户端、搜索引擎、关键词扩展装置，其中：

[0020] 客户端，用于接收检索串，并基于所述检索串向搜索引擎查询网页搜索结果；

[0021] 搜索引擎，用于向客户端提供对应于检索串的网页搜索结果；

[0022] 关键词扩展装置，用于设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树，利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

[0023] 从上述技术方案可以看出，在本发明实施方式中，设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树；接收检索串，并基于所述检索串获取网页搜索结果；利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。由此可见，应用本发明实施方式以后，使用搜索结果对原始检索串进行扩充(比如文本扩充和语义扩充)，通过模式算法在搜索结果中查找关键词，从而极大地丰富了关键词匹配结果，扩展了检索串的相关关键词，提高了展示内容的全面性。

[0024] 而且，在本发明实施方式中，利用网页搜索结果表征检索串和关键词的语义特征，结合文本相关性和分类等特征，通过机器学习方法训练得到相似度计算模型，量化了检索串与关键词之间的相似度，从而保证了展示内容的相关性。

[0025] 另外，可以将本发明实施方式可以应用到各种终端中，可以跨平台跨终端使用本发明实施方式，适用范围非常广泛。

附图说明

[0026] 图 1 为根据本发明实施方式的基于检索串的关键词扩展方法流程图；

[0027] 图 2 为根据本发明实施方式的关键词扩展示意图；

[0028] 图 3 为根据本发明实施方式的关键词与检索串相关性度量示意图；

[0029] 图 4 为根据本发明实施方式的关键词扩展以及关键词与检索串相关性度量的示意图；

[0030] 图 5 为根据本发明实施方式的逻辑回归模型的训练流程图；

[0031] 图 6 为根据本发明实施方式的基于检索串的关键词扩展装置结构图；

[0032] 图 7 为根据本发明实施方式的基于检索串的关键词扩展系统结构图。

具体实施方式

[0033] 为使本发明的目的、技术方案和优点更加清楚，下面结合附图对本发明作进一步的详细描述。

[0034] 在现有技术中，经常涉及到由检索串到关键词的映射，也就是检索串的匹配。现有技术常见的检索串匹配方法主要包括：

[0035] (1)从检索串本身匹配关键词，如精确匹配、词组匹配和广泛匹配(这里特指有语素删除的广泛匹配，抽取出的关键词是检索串的一个或几个子串的组合)。例如有广告主提交了如下关键词：

[0036] <ABCDEF、ABC、CDE、ACD、CA>(其中 A、B、C、D、E、F 是单个语素)；

[0037] 当有用户输入检索串 ABCDEF，则搜索引擎的广告匹配系统通过精确匹配，可以找到关键词 ABCDEF；通过词组匹配，可以找到关键词 ABC 和 CDE；通过广泛匹配，可以找到关键词 ACD 和 CA。

[0038] (2)利用特定的分析技术，如 word 删除、检索串替换(querysubstitution)、纠错、词干抽取、共同点击等方法，对变换后的检索串(query)匹配关键词。例如，检索串替换可以利用会话日志(session log)数据，通过统计方法得到相似检索串或片段(phrase)，对检索串进行改写替换，并给出改写后的串和原串之间的相似度衡量。某些情况下，用户在使用搜索引擎的过程中出现拼写错误或不规范的输入，导致没有合适的关键词触发出来，纠错就是针对性地对输入检索串进行修正、改写，正确表达检索意图。

[0039] 然而，上述方式中都是仅从检索串字面抽出关键词，因此关键词数量有限，从而难于保证展示内容的全面性。而且，通过上述方式所找到的关键词仅仅局限于文本上的关联，很难找到其他语义上相关的关键词，因此容易展示出与用户本意相差较多的展示内容，导致展示内容并不相关。

[0040] 在本发明实施方式中，可以根据检索串的自然搜索结果，将搜索结果全文，或者预定数目的标题和 / 或摘要汇集在一起，以作为检索串扩展内容，并从中找出扩展关键词。

[0041] 图 1 为根据本发明实施方式的基于检索串的关键词扩展方法流程图。

[0042] 如图 1 所示，该方法包括：

[0043] 步骤 101：设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树。

[0044] 在这里，可以预先设置预设关键词集合，在该预设关键词集合中包含有预设的关键词。比如，在广告搜索引擎应用中，该预设关键词可以具体为广告商预先提供的关键词。可以将预设关键词集合中的预设关键词作为模式串，并且将预设关键词集合建为键树(trie)。

[0045] 步骤 102：接收检索串，并基于所述检索串获取网页搜索结果。

[0046] 在这里，可以从用户接收检索串，并基于所述检索串获取网页搜索结果。比如，可以由搜索引擎基于该检索串从互联网上爬取对应于检索串的网页搜索结果。

[0047] 步骤 103：利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由

获得的所述模式串确定基于该检索串的扩展关键词。

[0048] 在这里，可以根据检索串的自然搜索结果，将网页搜索结果的全文作为检索串扩展内容，从中找出扩展关键词；或者将网页搜索结果中预定数目的标题和 / 或摘要汇集在一起以作为检索串扩展内容，从中找出扩展关键词。

[0049] 具体地，网页搜索结果的全文或者网页搜索结果中的预定数目的标题和摘要汇集通常为较长的文本，而从较长的文本中找出关键词，可以将整块内容切词，判断每个语素是否为包含于预设关键词集合中的关键词，如果是作为扩展关键词，不是，则丢弃该语素并继续判断下一个语素。不过，这种方式只能找出单语素(关键词切词后是其自身)的关键词，而预设关键词集合中的关键词(比如：广告系统中的关键词)大部分由多语素组成。

[0050] 在本发明实施方式中，还可以将单语素组合在一起，如 A、B、C 组合成 ABC，并判断 ABC 是否为关键词，如果是，则认定找到了多语素的关键词 ABC。不过，由于长文本中可能有几百个单语素，通过排列组合的方式进行验证，可能复杂度会较高。

[0051] 在本发明实施方式中，优选根据多模式匹配算法(比如 AC 算法)从检索串的扩展知识中抽取出扩展关键词。

[0052] 下面以 AC 算法为例对多模式匹配算法从检索串的扩展知识中抽取出扩展关键词进行示范性详细说明。

[0053] AC 算法即 Aho-Corasick 算法，是一个经典的多模式匹配算法。对于给定的长度为 n 的文本，和模式集合 P{p1,p2,...pm}，可以在 O(n) 时间复杂度内，找到文本中的所有目标模式，而与模式集合的规模 m 无关。AC 算法的原理是用多模式串建立一个确定性的树形有限状态机，以被抽取串作为该有限状态机的输入，使状态机进行状态转换。当到达某些特定的状态时，说明发生模式匹配，获得的模式串即可确定为基于该检索串的扩展关键词，即找到了扩展关键词。

[0054] 本发明实施方式中，可以将预设关键词集合中的每个预设关键词作为一个模式串，所有的预设关键词建成一棵键树，检索串及其网页扩展内容作为有限状态机的输入，在状态转换过程中，当匹配到某一个模式时，该模式对应的关键词就是检索串扩展的扩展关键词。

[0055] 以上虽然以 AC 算法为实例对本发明实施方式进行了详细描述，本领域技术人员可以意识到，这种描述仅是示范性的，并不用于对本发明实施方式进行限定。

[0056] 而且，在本发明实施方式中，既可以针对该网页搜索结果执行多模式匹配以获得多个模式串，也可以每次针对该网页搜索结果执行单模式匹配以获得单个模式串。

[0057] 在一个实施方式中，还可以进一步从所述检索串本身提取扩展关键词，再将所述由模式串确定的基于该检索串的扩展关键词以及从所述检索串本身提取的扩展关键词相聚合，以获得扩展关键词集合。

[0058] 在一个实施方式中，还可以进一步利用检索串变换(rewrite)方式获取扩展关键词；再将所述由模式串确定的基于该检索串的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

[0059] 在一个实施方式中，还可以进一步从所述检索串本身提取扩展关键词，以及利用检索串变换方式获取扩展关键词，再将所述由模式串确定的基于该检索串的扩展关键词、从所述检索串本身提取的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，

以获得扩展关键词集合。

[0060] 图 2 为根据本发明实施方式的关键词扩展示意图。

[0061] 比如，可以利用 AC 算法从检索串的网页搜索结果中扩展出关键词，这些关键词形成集合 E，<e1...en>；再从检索串本身扩展出关键词 Z，<z1..zm>；还可以利用检索串变换技术扩展的关键词 S，<s1...s>。然后将这三种方式扩展出来的关键词聚合在一起，经过去重，就形成了关键词的最终候选集 K，<k1....kt>。

[0062] 即 K ＝ E ∪ Z ∪ S，其中，n、m、j 为相应方法扩展出的关键词数，t 为聚合后关键词总数。

[0063] 通过本发明实施方式可以找到较多的扩展关键词，从而可以提高展示内容的全面性。

[0064] 进一步地，为了保证扩展关键词的质量，即保证扩展关键词与搜索串之间的相关性，本发明还可以利用网页搜索结果对检索串和关键词进行语义表示，结合文本相似度和分类相似度等特征，通过机器学习的方法训练出相关性模型，从而实现对关键词匹配质量的量化度量。

[0065] 图 3 为根据本发明实施方式的关键词与检索串相关性度量示意图。

[0066] 由图 3 可见，在本发明实施方式中，进一步包括：从扩展关键词集合中的扩展关键词和检索串，分别提取至少两个比较特征，所述比较特征包括文本特征、分类特征或语义特征。

[0067] 然后，基于所述扩展关键词集合中的扩展关键词和检索串的每个比较特征，计算所述扩展关键词集合中的扩展关键词和检索串之间的每个比较特征的相关性；再根据逻辑回归模型对各个比较特征的相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；然后，基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

[0068] 其中，比较特征具体可以为文本特征、分类特征或语义特征，等等。

[0069] 在本发明实施方式中，可以由至少两个比较特征来计算扩展关键词和检索串的相关性指标。比如：可以根据文本特征和分类特征来计算扩展关键词和检索串的相关性指标；也可以根据分类特征和语义特征来计算扩展关键词和检索串的相关性指标；还可以根据文本特征和语义特征来计算扩展关键词和检索串的相关性指标，还可以根据文本特征、分类特征和语义特征来计算扩展关键词和检索串的相关性指标。

[0070] 比如，可以从扩展关键词集合中的扩展关键词和检索串分别提取文本特征、分类特征和语义特征；再基于所述扩展关键词集合中的扩展关键词和检索串的文本特征，计算所述扩展关键词集合中的扩展关键词和检索串的文本相关性；基于所述扩展关键词集合中的扩展关键词和检索串的分类特征，计算所述扩展关键词集合中的扩展关键词和检索串的分类相关性；基于所述扩展关键词集合中的扩展关键词和检索串的语义特征，计算所述扩展关键词集合中的扩展关键词和检索串的语义相关性。

[0071] 然后，根据逻辑回归模型对所述文本相关性、分类相关性和语义相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标。最后，基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

[0072] 具体地，本发明实施方式可以使用有监督的机器学习方法，通过训练、评测和对比优化，得到衡量检索串和扩展关键词相关性的逻辑回归模型，并应用到各种工业系统中。其中有两处关键技术：

[0073] (1)特征选取：

[0074] 本发明实施方式可以选用三类特征，包括文本特征、分类特征和语义特征。

[0075] 文本特征是从检索串和关键词本身的字面相似度进行衡量，主要包括公共子串、共有语素、编辑距离等。

[0076] 分类特征是检索串和关键词在类别上的重合程度。

[0077] 语义特征，顾名思义，是跟检索串和关键词的语义相关的。检索串和关键词本身都是短文本，蕴含的语义信息有限，因此需要外部知识的补充。和检索串一样，关键词的语义特征也是从其对应的网页搜索结果中提取出来的。

[0078] 具体地，获取检索串 / 关键词的网页搜索结果的预设数目标题和摘要，将这些内容中重要的语素提取出来，形成一个向量，以表示其语义信息，其中重要语素的选取是根据语素本身的重要性和语素在搜索结果中出现的频率来进行的，语素本身越重要、出现的频率越高，越能代表检索串的语义。这样每一对检索串和关键词，都有了代表各自语义的向量，通过向量的余弦相似度，就可以得到检索串和关键词间的语义相似度。

[0079] 本发明实施方式中，优选的语义特征可以有六个，分别是标题域的语义特征、摘要域的语义特征、标题域和摘要域整合后的语义特征、去掉检索串和关键词共有语素后标题域的语义特征、去掉检索串和关键词的共有语素后摘要域的语义特征及去掉检索串和关键词的共有语素后标题域和摘要域整合后的语义特征。

[0080] (2)逻辑回归模型的训练：

[0081] 基于上述技术(1)确定特征后，就可以进行逻辑回归模型的训练和测试。在训练逻辑回归模型之前，选取出训练数据和测试数据由编辑人员进行标注。标注完毕后，获取标注数据的各个子特征，再利用逻辑回归算法对训练数据进行训练，得到模型各个特征的权重，然后再利用测试数据进行评测，评测效果符合预期则可应用，不符合预期则对特征进行调整后再次训练。

[0082] 图 4 为根据本发明实施方式的关键词扩展以及关键词与检索串相关性度量的示意图。图 5 为根据本发明实施方式的逻辑回归模型的训练流程图。

[0083] 如图 5 所示，该方法包括：

[0084] 步骤 501：确定所使用的逻辑回归模型及其特征。

[0085] 步骤 502：选取针对该逻辑回归模型的训练数据以及测试数据。

[0086] 步骤 503：制定评测标准，并由用户进行评测。

[0087] 步骤 504：获取训练数据和测试数据的子特征。

[0088] 步骤 505：训练该逻辑回归模型，以得到逻辑回归模型各个参数值。

[0089] 步骤 506：使用评测数据对该逻辑回归模型进行评测。

[0090] 步骤 507：判断评测结果是否已经达到预期，如果是则执行步骤 508 并结束本流程，如果不是则执行步骤 509 并结束本流程。

[0091] 步骤 508：输出该逻辑回归模型。

[0092] 步骤 509：增加、删除或优化该逻辑回归模型算法，或者补充评测数据。

[0093] 在本发明实施方式中，利用逻辑回归模型，计算出来的是检索串和关键词的综合相似度，这个相似度可以作为衡量扩展关键词跟检索串是否相关的标准。针对扩展关键词，可以按综合相似度排序，其中得分较高的，作为最终的匹配结果。特别的，对于相似度得分较低的关键词，如果是原始检索串的子串，还可以用来过滤坏词，不让其作为传统的匹配结果触发其他的匹配流程。

[0094] 可以将本发明实施方式的相似度计算方法应用到各种应用情形中，比如网络搜索广告系统。而且，可以利用本发明实施方式的相似度计算方法实现任意两个短串之间的相似度衡量。

[0095] 另外，本发明实施方式可以应用于各种终端实体。比如，终端可以包括但是不局限于：功能手机、智能手机、掌上电脑、个人电脑(PC)、平板电脑或个人数字助理(PDA)，等等。

[0096] 基于上述详细分析，本发明实施方式还提出了一种基于检索串的关键词扩展装置。

[0097] 图 6 为根据本发明实施方式的基于检索串的关键词扩展装置结构图。

[0098] 如图 6 所示，该装置包括键树建立单元 601、搜索结果获取单元 602 和关键词扩展单元 603，其中：

[0099] 键树建立单元 601，用于设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树；

[0100] 搜索结果获取单元 602，用于接收检索串，并基于所述检索串获取网页搜索结果；

[0101] 关键词扩展单元 603，用于利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

[0102] 在一个实施方式中，关键词扩展单元 603，用于针对该网页搜索结果执行多模式匹配以获得多个模式串，或者针对该网页搜索结果执行单模式匹配以获得单个模式。

[0103] 优选地，进一步包括关键词提取单元 604 和关键词聚合单元 605；

[0104] 关键词提取单元 604，用于从所述检索串本身提取扩展关键词；

[0105] 关键词聚合单元 605，用于将所述由模式串确定的基于该检索串的扩展关键词以及从所述检索串本身提取的扩展关键词相聚合，以获得扩展关键词集合。

[0106] 在一个实施方式中，进一步包括检索串变换单元 606 和关键词聚合单元 605；

[0107] 检索串变换单元 606，用于利用检索串变换(rewrite)方式获取扩展关键词；

[0108] 关键词聚合单元 605，用于将所述由模式串确定的基于该检索串的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

[0109] 在一个实施方式中，进一步包括关键词提取单元 604、检索串变换单元 606 和关键词聚合单元 605；其中：

[0110] 关键词提取单元 604，用于从所述检索串本身提取扩展关键词；

[0111] 检索串变换单元 606，用于利用检索串变换(rewrite)方式获取扩展关键词；

[0112] 关键词聚合单元 605，用于将所述由模式串确定的基于该检索串的扩展关键词、从所述检索串本身提取的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

[0113] 进一步地，该装置还可以包括相关性指标确定单元 607，其中：

[0114] 相关性指标确定单元 607，用于从所述扩展关键词集合中的扩展关键词和检索串，

分别提取至少两个比较特征，所述比较特征包括文本特征、分类特征或语义特征；基于所述扩展关键词集合中的扩展关键词和检索串的每个比较特征，计算所述扩展关键词集合中的扩展关键词和检索串之间的每个比较特征的相关性；根据逻辑回归模型对各个比较特征的相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

[0115] 优选地，相关性指标确定单元 607，用于从所述扩展关键词集合中的扩展关键词和检索串分别提取文本特征、分类特征和语义特征；基于所述扩展关键词集合中的扩展关键词和检索串的文本特征，计算所述扩展关键词集合中的扩展关键词和检索串的文本相关性；基于所述扩展关键词集合中的扩展关键词和检索串的分类特征，计算所述扩展关键词集合中的扩展关键词和检索串的分类相关性；基于所述扩展关键词集合中的扩展关键词和检索串的语义特征，计算所述扩展关键词集合中的扩展关键词和检索串的语义相关性；根据逻辑回归模型对所述文本相关性、分类相关性和语义相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

[0116] 基于上述详细分析，本发明实施方式还提出了一种基于检索串的关键词扩展系统。

[0117] 图 7 为根据本发明实施方式的基于检索串的关键词扩展系统结构图。

[0118] 如图 7 所示，包括客户端 701、搜索引擎 702 和关键词扩展装置 703，其中：

[0119] 客户端 701，用于接收检索串，并基于所述检索串向搜索引擎查询网页搜索结果；

[0120] 搜索引擎 702，用于向客户端提供对应于检索串的网页搜索结果；

[0121] 关键词扩展装置 703，用于设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树，利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。

[0122] 在一个实施方式中，关键词扩展装置 703，进一步用于从所述检索串本身提取扩展关键词，以及利用检索串变换方式获取扩展关键词；将所述由模式串确定的基于该检索串的扩展关键词、从所述检索串本身提取的扩展关键词以及利用检索串变换方式获取的扩展关键词相聚合，以获得扩展关键词集合。

[0123] 优选地，关键词扩展装置 703，进一步用于从所述扩展关键词集合中的扩展关键词和检索串，分别提取至少两个比较特征，所述比较特征包括文本特征、分类特征或语义特征；基于所述扩展关键词集合中的扩展关键词和检索串的每个比较特征，计算所述扩展关键词集合中的扩展关键词和检索串之间的每个比较特征的相关性；根据逻辑回归模型对各个比较特征的相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

[0124] 更优选地，关键词扩展装置 703，进一步用于从所述扩展关键词集合中的扩展关键词和检索串分别提取文本特征、分类特征和语义特征；基于所述扩展关键词集合中的扩展

关键词和检索串的文本特征，计算所述扩展关键词集合中的扩展关键词和检索串的文本相关性；基于所述扩展关键词集合中的扩展关键词和检索串的分类特征，计算所述扩展关键词集合中的扩展关键词和检索串的分类相关性；基于所述扩展关键词集合中的扩展关键词和检索串的语义特征，计算所述扩展关键词集合中的扩展关键词和检索串的语义相关性；根据逻辑回归模型对所述文本相关性、分类相关性和语义相关性进行特征拟合，以得到扩展关键词集合中的扩展关键词和检索串的相关性指标；基于所述扩展关键词集合中的扩展关键词和检索串的相关性指标，从所述扩展关键词集合中确定符合预定相关性指标门限值的扩展关键词。

[0125] 可以将图 7 所示系统应用于各种应用场景中，比如应用于搜索引擎广告系统中。

[0126] 而且，客户端 701、搜索引擎 702 和关键词扩展装置 703 相互之间可以采用的通信协议包括但是不局限于：传输控制协议 / 网际协议（TCP/IP）、超文本传输协议（HTTP）、简单邮件传输协议 (SMTP)、邮局协议的第 3 个版本（POP3），等等。

[0127] 还可以将图 6 所示装置集成到各种通信网络的硬件实体当中。比如，可以将事务提醒装置集成到：功能手机、智能手机、掌上电脑、个人电脑（PC）、平板电脑或个人数字助理（PDA）、网络服务器、广告服务器、搜索引擎等等设备之中。

[0128] 实际上，可以通过多种形式来具体实施本发明实施方式所提出的基于检索串的关键词扩展装置。比如，可以遵循一定规范的应用程序接口，将基于检索串的关键词扩展装置编写为安装到移动终端、智能手机、掌上电脑、个人电脑（PC）、平板电脑或个人数字助理（PDA）、网络服务器、广告服务器、搜索引擎中的插件程序，也可以将其封装为应用程序以供用户自行下载使用。当编写为插件程序时，可以将其实施为 ocx、dll、cab 等多种插件形式。也可以通过 Flash 插件、RealPlayer 插件、MMS 插件、MIDI 五线谱插件、ActiveX 插件等具体技术来实施本发明实施方式所提出的基于检索串的关键词扩展装置。

[0129] 可以通过指令或指令集存储的储存方式将本发明实施方式所提出的基于检索串的关键词扩展方法存储在各种存储介质上。这些存储介质包括但是不局限于：软盘、光盘、DVD、硬盘、闪存、U 盘、CF 卡、SD 卡、MMC 卡、SM 卡、记忆棒（Memory Stick）、xD 卡等。

[0130] 另外，还可以将本发明实施方式所提出的基于检索串的关键词扩展方法应用到基于闪存（Nand fash）的存储介质中，比如 U 盘、CF 卡、SD 卡、SDHC 卡、MMC 卡、SM 卡、记忆棒、xD 卡等。

[0131] 综上所述，在本发明实施方式中，设置预设关键词集合，将所述预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树；接收检索串，并基于所述检索串获取网页搜索结果；利用所述键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的所述模式串确定基于该检索串的扩展关键词。由此可见，应用本发明实施方式以后，使用搜索结果对原始检索串进行文本和语义扩充，通过模式算法在搜索结果中查找关键词，从而极大地丰富了关键词匹配结果，扩展了检索串的关键词，提高了展示内容的全面性。

[0132] 而且，在本发明实施方式中，利用网页搜索结果表征检索串和关键词的语义特征，结合文本相关性和分类等特征，通过机器学习方法训练得到相似度计算模型，量化了检索串与关键词之间的相似度，从而保证了展示内容的相关性。

[0133] 另外，可以将本发明实施方式可以应用到各种终端中，可以跨平台跨终端使用本

发明实施方式，适用范围非常广泛。

[0134] 以上所述，仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

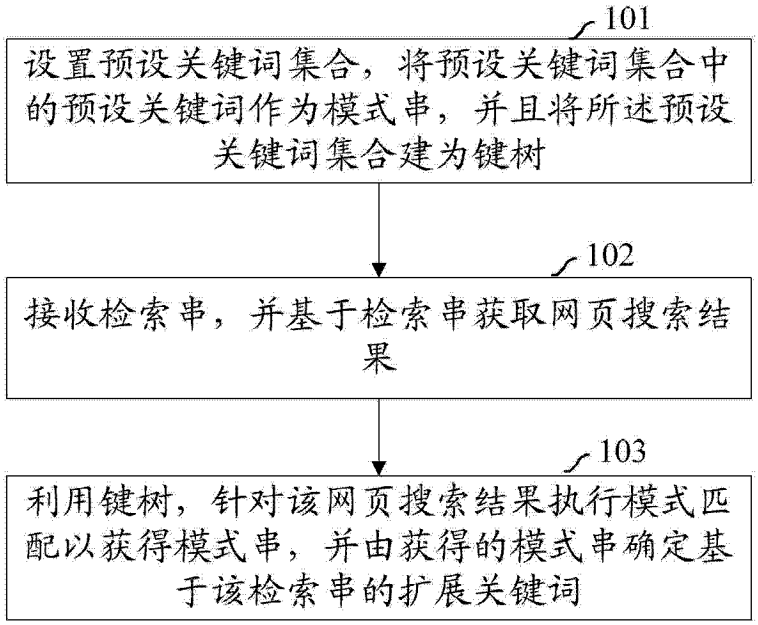

101
设置预设关键词集合，将预设关键词集合中的预设关键词作为模式串，并且将所述预设关键词集合建为键树
102
接收检索串，并基于检索串获取网页搜索结果
103
利用键树，针对该网页搜索结果执行模式匹配以获得模式串，并由获得的模式串确定基于该检索串的扩展关键词

图 1

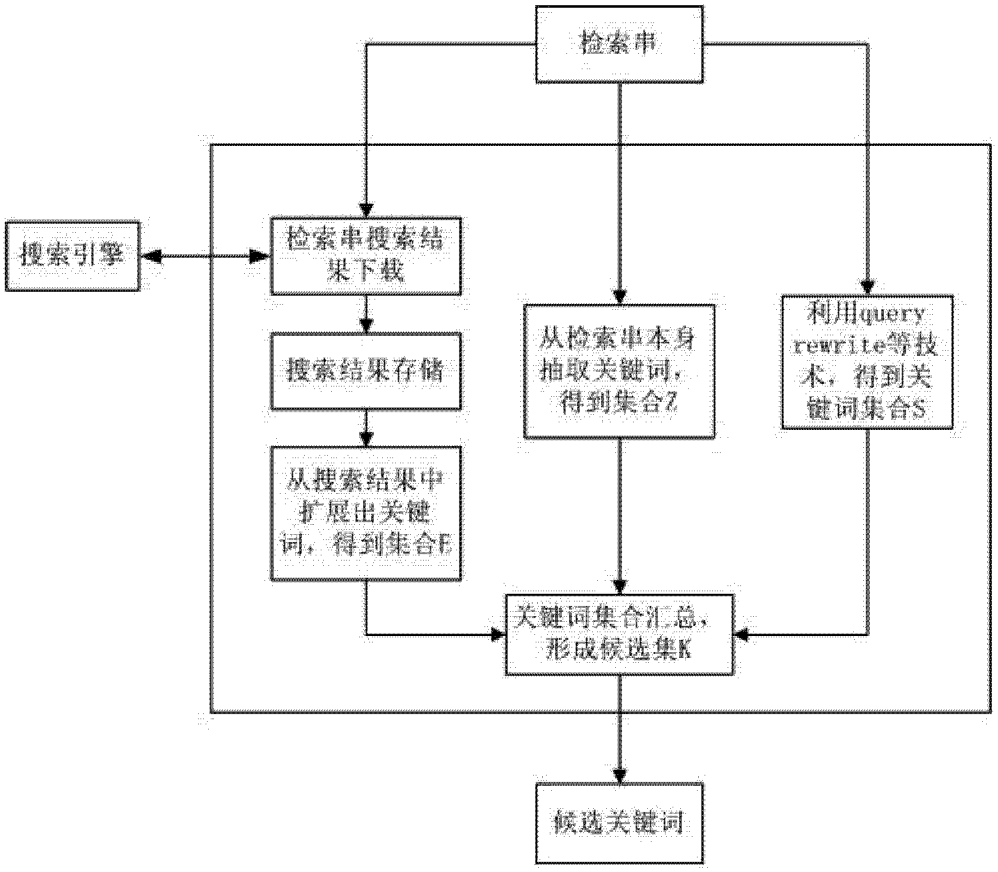

检索串
搜索引擎
检索串搜索结果下载
搜索结果存储
从搜索结果中扩展出关键词，得到集合E
从检索串本身抽取关键词，得到集合Z
利用query rewrite等技术，得到关键词集合S
关键词集合汇总，形成候选集K
候选关键词

图 2

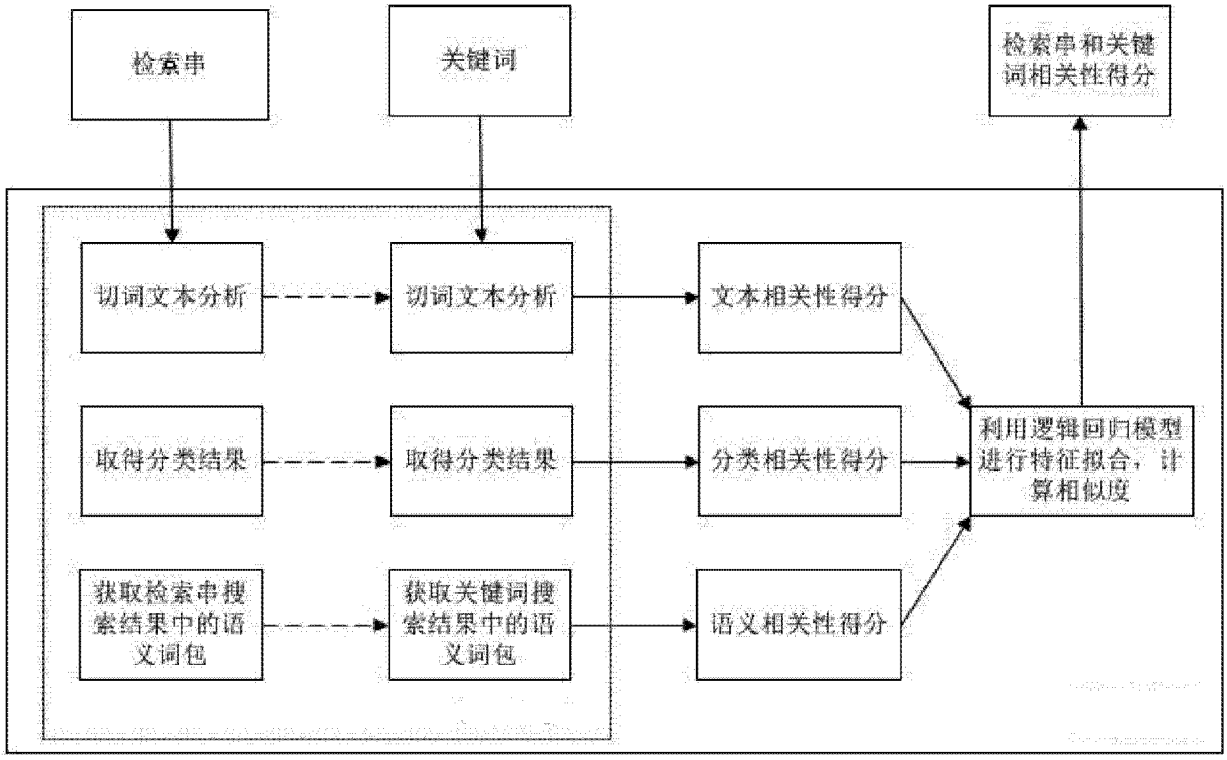

检索串
关键词
检索串和关键词相关性得分
切词文本分析
切词文本分析
文本相关性得分
取得分类结果
取得分类结果
分类相关性得分
利用逻辑回归模型进行特征拟合，计算相似度
获取检索串搜索结果中的语义词包
获取关键词搜索结果中的语义词包
语义相关性得分

图 3

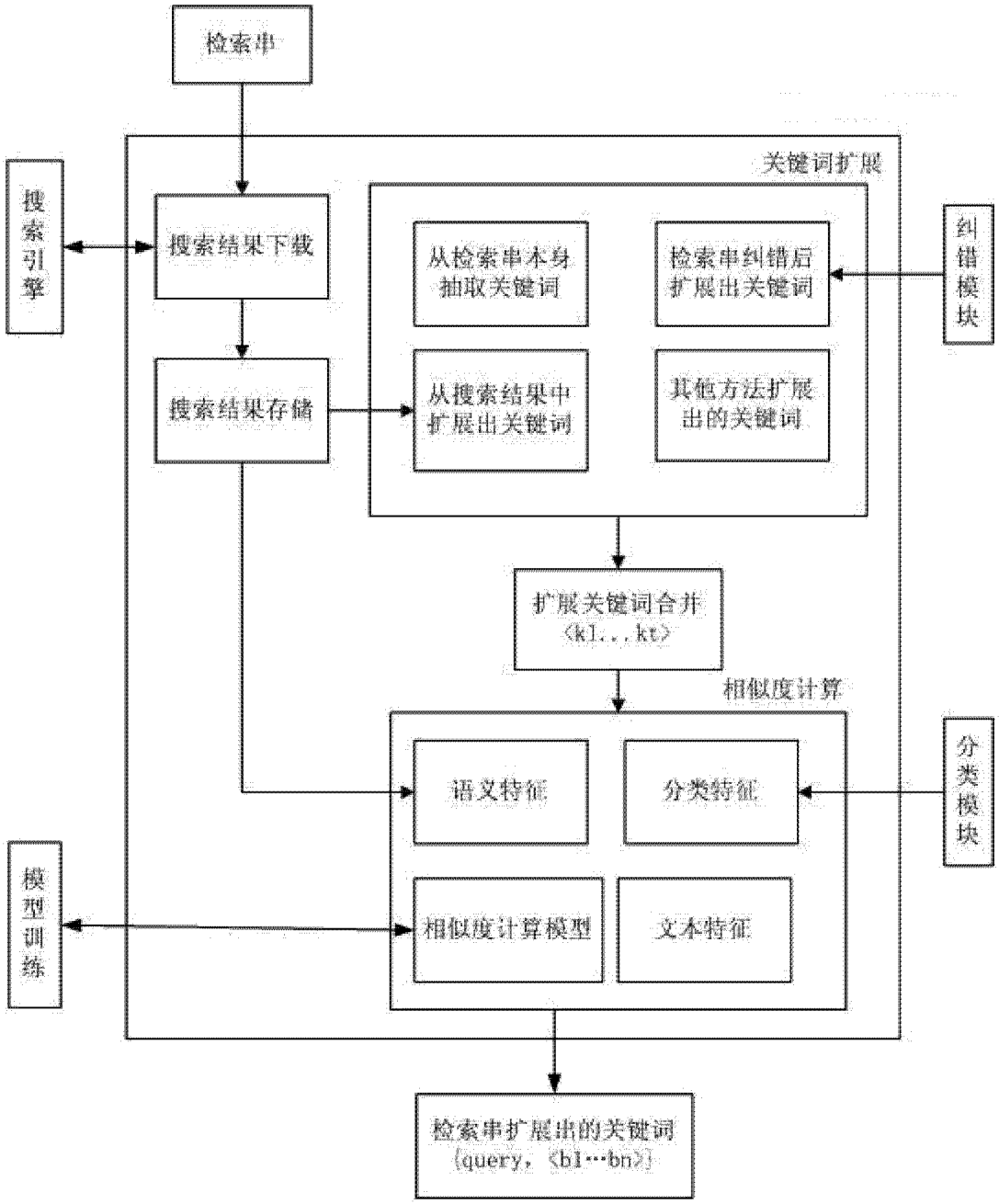

检索串
关键词扩展
搜索引擎
搜索结果下载
从检索串本身抽取关键词
检索串纠错后扩展出关键词
纠错模块
搜索结果存储
从搜索结果中扩展出关键词
其他方法扩展出的关键词
扩展关键词合并
〈k1...kt〉
相似度计算
语义特征
分类特征
分类模块
模型训练
相似度计算模型
文本特征
检索串扩展出的关键词
{query，〈b1…bn〉}

图 4

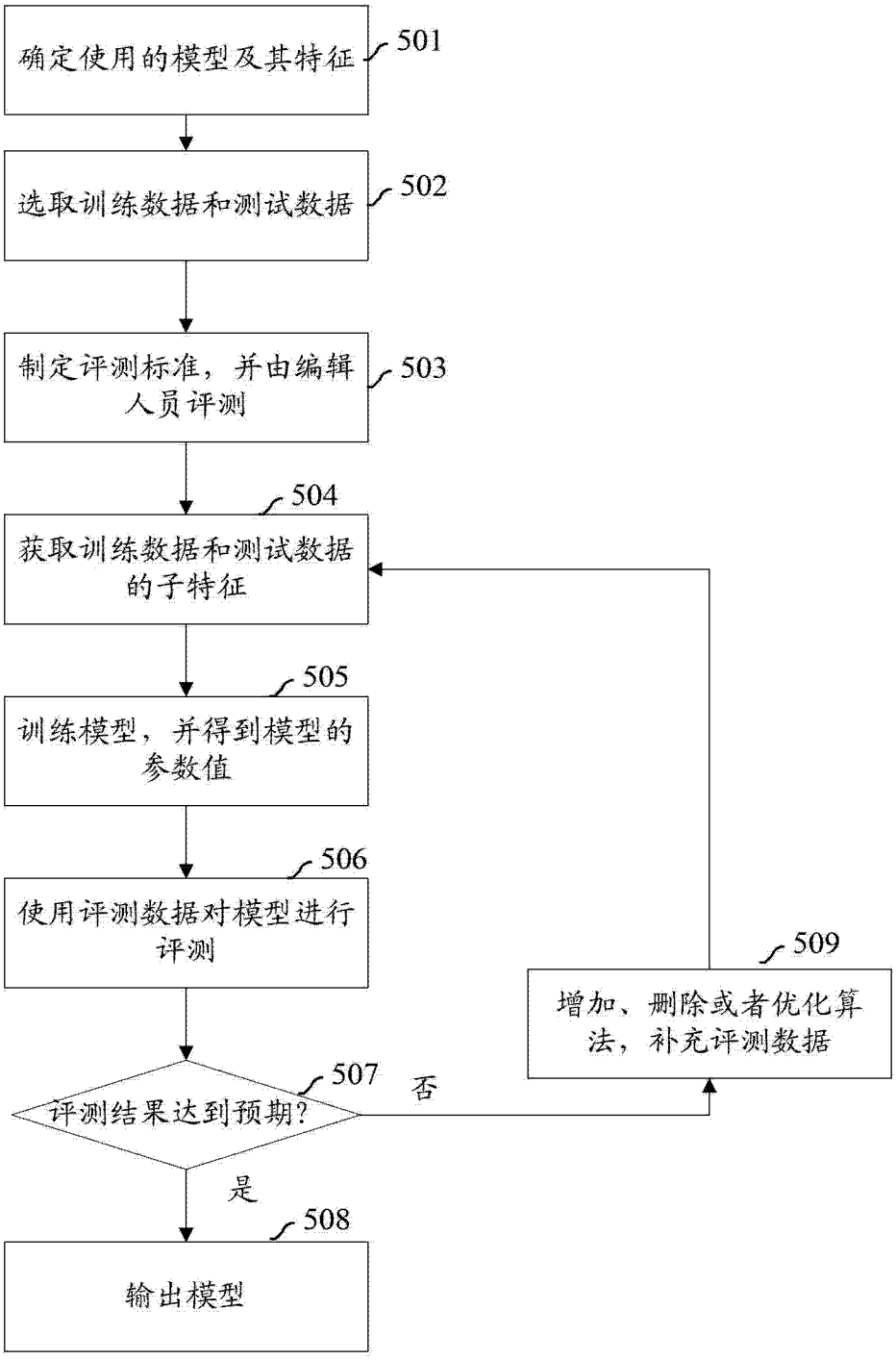

确定使用的模型及其特征
501
选取训练数据和测试数据
502
制定评测标准，并由编辑人员评测
503
504
获取训练数据和测试数据的子特征
505
训练模型，并得到模型的参数值
506
使用评测数据对模型进行评测
509
增加、删除或者优化算法，补充评测数据
评测结果达到预期?
507
否
是
508
输出模型

图 5

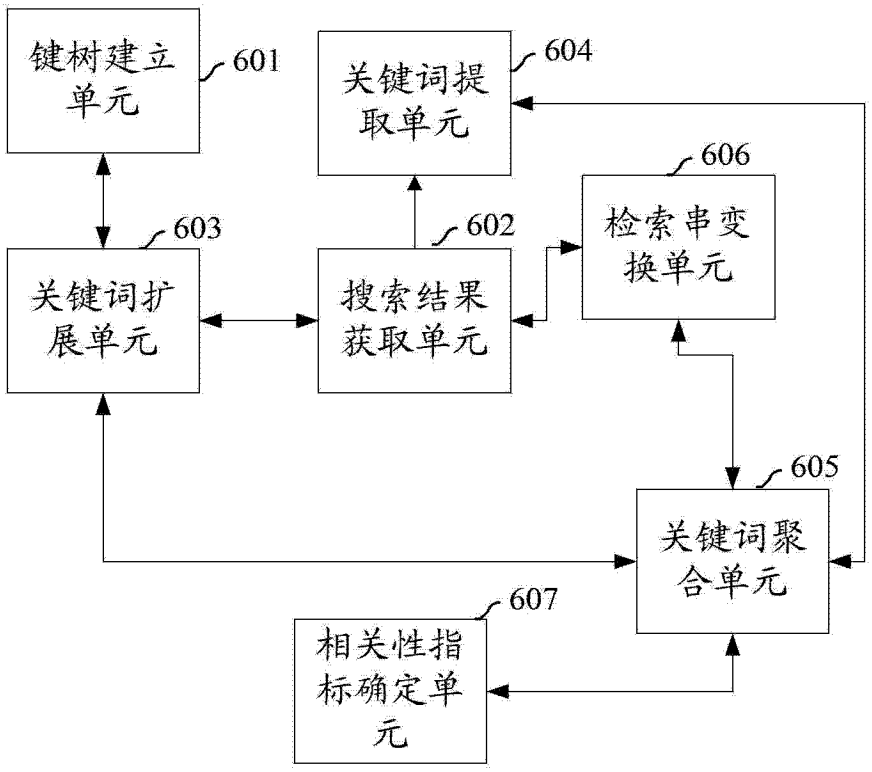

键树建立单元
601
关键词提取单元
604
606
检索串变换单元
603
602
关键词扩展单元
搜索结果获取单元
605
关键词聚合单元
607
相关性指标确定单元

图 6

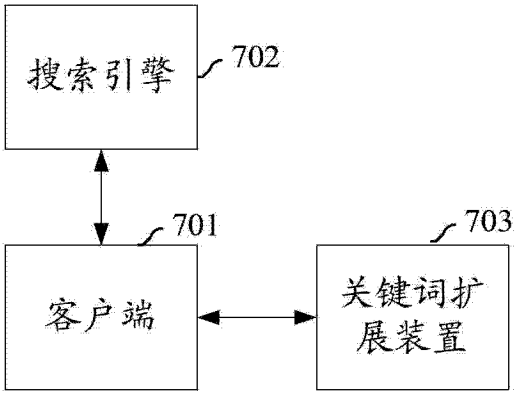

搜索引擎
702
701
703
客户端
关键词扩展装置

图 7