(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 103678513 A
(43)申请公布日 2014.03.26

(21)申请号 201310611470.7

(22)申请日 2013.11.26

(71)申请人 安徽科大讯飞信息科技股份有限公司
地址 230088 安徽省合肥市高新开发区望江西路666号

(72)发明人 吴及 侯晋峰 吕萍 何婷婷 胡国平 胡郁

(74)专利代理机构 北京维澳专利代理有限公司 11252
代理人 王立民 吉海莲

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

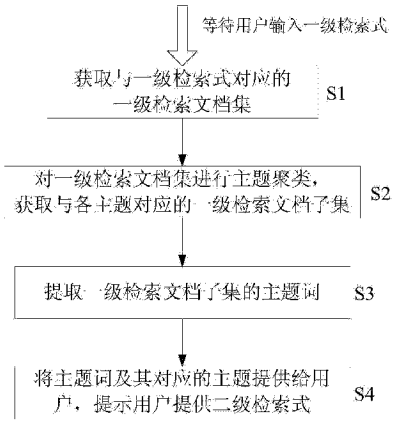
权利要求书4页 说明书10页 附图7页

(54)发明名称

一种交互式的检索式生成方法及系统

(57)摘要

本发明公开了一种交互式的检索式生成方法及系统，该方法包括：在接收到用户输入的一级检索式后获取与所述一级检索式相对应的一级检索文档集；对一级检索文档集进行主题聚类，获取与各主题一一对应的一级检索文档子集；提取所述一级检索文档子集中的主题词；将主题词及其对应的主题提供给用户，并提示用户利用所述主题词确定与其对应的主题相关的二级检索式。本发明的交互式的检索式生成方法及系统可以辅助用户生成复杂检索式，帮助专业检索领域的专业检索人员生成更为精确的检索式。



CN 103678513 A

1. 一种交互式的检索式生成方法，其特征在于，包括：

在接收到用户输入的一级检索式后获取与所述一级检索式相对应的一级检索文档集；

对所述一级检索文档集进行主题聚类，获取与各主题一一对应的一级检索文档子集；

提取所述一级检索文档子集中的主题词；

将所述主题词及其对应的主题提供给用户，并提示用户利用所述主题词确定与其对应的主题相关的二级检索式。

2. 根据权利要求 1 所述的方法，其特征在于，所述方法还包括：

在接收到用户输入的二级检索式后，获取与各二级检索式一一对应的二级检索文档集；

对二级检索式进行两两组合，使每组两个二级检索式成为两个待验证检索式；

对两个待验证检索式进行交叉验证，获取与两个待验证检索式一一对应的两个检索文档集的交叉文档集，若所述交叉文档集中文档的数目大于设定阈值，则对两个待验证检索式进行优化。

3. 根据权利要求 2 所述的方法，其特征在于，所述对两个待验证检索式进行优化包括：

对所述交叉文档集进行主题聚类，获取与所述交叉文档集的各主题一一对应的交叉文档子集；

提取交叉文档子集中的主题词，并将交叉文档子集的主题词及其对应的主题提供给用户，同时提示用户利用交叉文档子集的主题词优化两个待验证检索式，确定两个优化检索式。

4. 根据权利要求 3 所述的方法，其特征在于，所述方法还包括：

在接收到用户输入的两个优化检索式后，获取与两个优化检索式一一 对应的两个优化检索文档集；

使两个优化检索式作为两个待验证检索式进行所述交叉验证。

5. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，所述方法还包括：

为用户提供用于确定检索式的逻辑运算符，所述逻辑运算符包括“邻近”，所述“邻近”表示相“邻近”的两个关键词在文档中的距离在预设字数以内。

6. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，所述主题聚类的方法包括：

步骤 a1：设定待聚类文档集为被拆分类，提取被拆分类中文档的聚类特征，获取与文档一一对应的特征向量；

步骤 a2：在被拆分类中选择两个密度最大的文档作为种子文档，文档的密度为在被拆分类中与文档的余弦距离小于 0.5 的文档的个数，其中，文档间的余弦距离为文档的特征向量间的余弦距离；

步骤 a3：以两个所述种子文档作为种子，用 K 均值聚类算法将所述被拆分类拆分为两个主题；

步骤 a4：判断两个主题中是否有一个主题的文档数量小于被拆分类的预设百分比，如是则主题聚类结束，否则将两个主题中数量较多的一个主题作为被拆分类，继续执行步骤 a2。

7. 根据权利要求 6 所述的方法，其特征在于，在步骤 a1 中，对每篇文档计算特征词典

中每个词的 TF-IDF 值作为聚类特征，获取 k 维的特征向量，其中，k 等于特征词典中词的数量。

8. 根据权利要求 7 所述的方法，其特征在于，所述方法还包括：获取所述特征词典的方法为：

统计整个检索库中所有文档所包含的词及对应的词频，作为背景特征；

统计一级检索文档集中所有文档所包含的词及对应的词频，作为候选特征；

计算候选特征与背景特征之间的词的词频的差异度，选择差异度最大的预设数量的词构成所述特征词典。

9. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，提取经主题聚类得到的各文档子集的主题词包括：

提取各文档子集的候选主题词；

将同一候选主题词分配给词频最高的文档子集；

针对每个文档子集选择词频最高的 6 至 12 个候选主题词作为主题词。

10. 根据权利要求 9 所述的方法，其特征在于，所述提取各文档子集的候选主题词包括：

查找文档子集中距离在 m 个词以内的二元词组，在二元词组表中列出查找到的二元词组及对应的词频，其中 m 取 0 至 5 的整数；

查找二元词组表中的等同二元词组，所述等同二元词组由两个词相同、但语序不同的二元词组构成；在二元词组表中删除等同二元词组中词频较低的二元词组，并将词频较高的二元词组的词频更新为等同二元词组的词频；

在二元词组表中删除具有停词表中的停词的二元词组；

提取二元词组表中词频最高的 n 个二元词组作为文档子集的候选主题词，n 取 10 至 100 的整数。

11. 一种交互式的检索式生成系统，其特征在于，包括：

第一输入模块，用于接收用户输入的一级检索式；

第一检索模块，用于在接收到所述第一输入模块提供的一级检索式后获取与所述一级检索式相对应的一级检索文档集；

聚类模块，用于对所述一级检索文档集进行主题聚类，获取与各主题一一对应的一级检索文档子集；

主题词提取模块，用于提取所述一级检索文档子集中的主题词；

第一输出模块，用于将所述主题词及其对应的主题提供给用户，并提示用户利用所述主题词确定与其对应的主题相关的二级检索式。

12. 根据权利要求 11 所述的系统，其特征在于，所述系统还包括：

第二输入模块，用于接收用户输入的二级检索式；

第二检索模块，用于在接收到所述第二输入模块提供的二级检索式后， 获取与各二级检索式一一对应的二级检索文档集；

组合模块，用于对二级检索式进行两两组合，使每组两个二级检索式成为两个待验证检索式；以及，

交叉验证模块，用于对两个待验证检索式进行交叉验证，所述交叉验证模块包括：

统计单元，用于获取与两个待验证检索式一一对应的两个检索文档集的交叉文档集；

比较单元，用于将所述交叉文档集中文档的数目与设定阈值进行比较，如果交叉文档集中文档的数目大于设定阈值，则确定对两个待验证检索式进行优化。

13. 根据权利要求 12 所述的系统，其特征在于，

所述比较单元还用于在确定对两个待验证检索式进行优化后，将所述交叉文档集输入至所述聚类模块；

所述聚类模块还用于获取与所述交叉文档集的各主题一一对应的交叉文档子集；所述主题词提取模块还用于提取交叉文档子集的主题词；

所述系统还包括：

第二输出模块，用于将交叉文档子集的主题词及其对应的主题提供给用户，同时提示用户利用交叉文档子集的主题词优化两个待验证检索式，确定两个优化检索式。

14. 根据权利要求 13 所述的系统，其特征在于，所述系统还包括：

第三输入模块，用于接收用户输入的两个优化检索式，以及用于将两个优化检索式作为两个待验证检索式输入至所述交叉验证模块；

第三检索模块，用于在接收到所述第三输入模块提供的两个优化检索式后，获取与两个优化检索式一一对应的两个优化检索文档集。

15. 根据权利要求 11 至 14 中任一项所述的系统，其特征在于，所述聚类模块包括：

特征向量计算单元，用于设定待聚类文档集为被拆分类，提取被拆分类中各文档的聚类特征，获取与各文档一一对应的特征向量；

种子文档确定单元，用于在被拆分类中选择两个密度最大的文档作为种子文档，文档的密度为在被拆分类中与文档的余弦距离小于 0.5 的文档的个数，其中，文档间的余弦距离为文档的特征向量间的余弦距离；

K 均值聚类单元，用于以两个所述种子文档作为种子，用 K 均值聚类算法将所述被拆分类拆分为两个主题；以及，

判断单元，用于判断两个主题中是否有一个主题的文档数量小于被拆分类的预设百分比，如是则主题聚类结束，否则将两个主题中数量较多的一个主题作为被拆分类输入至种子文档确定单元。

16. 根据权利要求 15 所述的系统，其特征在于，所述特征向量计算单元用于对每篇文档计算特征词典中每个词的 TF-IDF 值作为聚类特征，获取 k 维的特征向量，其中，k 等于特征词典中词的数量。

17. 根据权利要求 16 所述的系统，其特征在于，所述聚类模块还包括特征词典获取模块，所述特征词典生成模块包括：

背景特征统计单元，用于统计整个检索库中所有文档所包含的词及对应的词频，作为背景特征；

候选特征统计单元，用于统计一级检索文档集中所有文档所包含的词及对应的词频，作为候选特征；

差异度计算单元，用于计算候选特征与背景特征之间的词的词频的差异度，选择差异度最大的预设数量的词构成所述特征词典。

18. 根据权利要求 11 至 14 中任一项所述的系统，其特征在于，所述主题词提取模块包

括：

候选主题词提取单元，用于提取各文档子集的候选主题词；

分配单元，用于将同一候选主题词分配给词频最高的文档子集；

主题词选择单元，用于针对每个文档子集选择词频最高的6至12个候选主题词作为主题词。

19. 根据权利要求18所述的系统，其特征在于，所述候选主题词提取单元包括：

二元词组查找子单元，用于查找文档子集中距离在m个词以内的二元词组，在二元词组表中列出查找到的二元词组及对应的词频，其中m取0至5的整数；

合并子单元，用于查找二元词组表中的等同二元词组，所述等同二元词组由两个词相同、但语序不同的二元词组构成；在二元词组表中删除等同二元词组中词频较低的二元词组，并将词频较高的二元词组的词频更新为等同二元词组的词频；

删除子单元，用于在二元词组表中删除具有停词表中的停词的二元词组；

候选主题词选择子单元，用于提取二元词组表中词频最高的n个二元词组作为文档子集的候选主题词，n取10至100的整数。

一种交互式的检索式生成方法及系统

技术领域

[0001] 本发明涉及文本检索领域，尤其涉及一种交互式的检索式生成方法及系统。

背景技术

[0002] 随着现代社会各种信息量的高速增长以及存储技术的不断进步，从海量数据中快速有效的获取有用信息也越来越困难，大量的数据得不到有效的利用。检索是一种实现海量数据中有用信息快速获取的技术手段，其接收用户检索式输入，在数据库中搜索与所述检索式相关的内容。检索式即理解和运算的查词串，至少包括关键词，对于复杂的检索，通常还包括逻辑运算符、搜索指令（搜索语法）等，其中关键词是检索式的主体，逻辑运算符和搜索指令根据具体的查询要求从不同的角度对关键词进行搜索限定。

[0003] 显然构建更加高效的检索式可以提高检索的精确性，对一些专业检索领域尤其具有重要意义。如电话服务行业的录音数据，通过语音识别转化为文本以后，由语音识别带来的一些错误容易导致精确信息获取的困难，而通过构建更加专业鲁棒的检索式则可以帮助我们对数据进行更加精确的定位，获取更多的信息。

[0004] 用户在使用检索系统时通常需要人工生成检索式，然而即使是一些专业领域的专业检索人员，也只是靠自己多年的从业经验来生成比较好的检索式，且个体差异很大。对此，为了改善和提高信息检索的性能，目前在检索领域一般采用查询扩展的方法，以用户原查询为基础，把与原查询相关的词或者词组自动添加到原查询，得到比原查询更长的新查询，以便更完整地描述原查询所隐含的语义或者主题，帮助信息检索系统提供更多有利于判断文档相关性的信息。其具体流程如下所示：

[0005] 步骤 1：接收用户输入的检索式；

[0006] 步骤 2：根据所述检索式在数据库中搜索得到相关文档，作为初检结果；

[0007] 步骤 3：从所述初检结果中获取原检索式的扩展词，具体可以利用聚类技术、文本挖掘技术、关联规则等，从文本集或者用户查询日志中获取；

[0008] 步骤 4：根据所述扩展词以及原检索式，生成新的检索式；

[0009] 步骤 5：根据所述新的检索式重新检索。

[0010] 基于查询扩展的检索式生成方法，以全自动的方式获得扩展词，得到比原查询更完备的新查询，实现了对原查询所隐含的语义或者主题的更完整的描述，从而帮助信息检索系统提供更多有利于判断文档相关性的信息。然而该方法生成的检索式对用户完全不透明，因此，所产生的结果是无法预料的；其次，扩展词之间通常采用“或”的逻辑进行连接，对检索结果的性能提升有限，且对于检索结果没有有效的快速评估的方法，需要检索人员一条一条的浏览；再次，生成的检索式也无法重复利用，如果用户想在不同的数据集上检索得到该类数据，则需要重新进行构建检索式。

发明内容

[0011] 本发明的一个目的在于克服现有技术中的不足，提供了一种交互式的检索式生成

方法，以辅助用户生成复杂检索式，帮助专业检索领域的专业检索人员生成更为精确的检索式。

[0012] 为了实现上述目的，本发明采用的技术方案为：一种交互式的检索式生成方法，包括：

[0013] 在接收到用户输入的一级检索式后获取与所述一级检索式相对应的一级检索文档集；

[0014] 对所述一级检索文档集进行主题聚类，获取与各主题一一对应的一级检索文档子集；

[0015] 提取所述一级检索文档子集中的主题词；

[0016] 将所述主题词及其对应的主题提供给用户，并提示用户利用所述主题词确定与其对应的主题相关的二级检索式。

[0017] 优选的是，所述方法还包括：

[0018] 在接收到用户输入的二级检索式后，获取与各二级检索式一一对应的二级检索文档集；

[0019] 对二级检索式进行两两组合，使每组两个二级检索式成为两个待验证检索式；

[0020] 对两个待验证检索式进行交叉验证，获取与两个待验证检索式一一对应的两个检索文档集的交叉文档集，若所述交叉文档集中文档的数目大于设定阈值，则对两个待验证检索式进行优化。

[0021] 优选的是，所述对两个待验证检索式进行优化包括：

[0022] 对所述交叉文档集进行主题聚类，获取与所述交叉文档集的各主题一一对应的交叉文档子集；

[0023] 提取交叉文档子集中的主题词，并将交叉文档子集的主题词及其对应的主题提供给用户，同时提示用户利用交叉文档子集的主题词优化两个待验证检索式，确定两个优化检索式。

[0024] 优选的是，所述方法还包括：

[0025] 在接收到用户输入的两个优化检索式后，获取与两个优化检索式一一对应的两个检索文档集；

[0026] 使两个优化检索式作为两个待验证检索式进行所述交叉验证。

[0027] 优选的是，所述方法还包括：

[0028] 为用户提供用于确定检索式的逻辑运算符，所述逻辑运算符包括“邻近”，所述“邻近”表示相“邻近”的两个关键词在文档中的距离在预设字数以内。

[0029] 优选的是，所述主题聚类的方法包括：

[0030] 步骤 a1：设定待聚类文档集为被拆分类，提取被拆分类中文档的聚类特征，获取与文档一一对应的特征向量；

[0031] 步骤 a2：在被拆分类中选择两个密度最大的文档作为种子文档，文档的密度为在被拆分类中与文档的余弦距离小于 0.5 的文档的个数，其中，文档间的余弦距离为文档的特征向量间的余弦距离；

[0032] 步骤 a3：以两个所述种子文档作为种子，用 K 均值聚类算法将所述被拆分类拆分为两个主题；

[0033] 步骤 a4 :判断两个主题中是否有一个主题的文档数量小于被拆分类的预设百分比,如是则主题聚类结束,否则将两个主题中数量较多的一个主题作为被拆分类,继续执行步骤 a2。

[0034] 优选的是,在步骤 a1 中,对每篇文档计算特征词典中每个词的 TF-IDF 值作为聚类特征,获取 k 维的特征向量,其中,k 等于特征词典中词的数量。

[0035] 优选的是,所述方法还包括 :获取所述特征词典的方法为 :

[0036] 统计整个检索库中所有文档所包含的词及对应的词频,作为背景特征 ;

[0037] 统计一级检索文档集中所有文档所包含的词及对应的词频,作为候选特征 ;

[0038] 计算候选特征与背景特征之间的词的词频的差异度,选择差异度最大的预设数量的词构成所述特征词典。

[0039] 优选的是,提取经主题聚类得到的各文档子集的主题词包括 :

[0040] 提取各文档子集的候选主题词 ;

[0041] 将同一候选主题词分配给词频最高的文档子集 ;

[0042] 针对每个文档子集选择词频最高的 6 至 12 个候选主题词作为主题词 ;

[0043] 优选的是,所述提取各文档子集的候选主题词包括 :

[0044] 查找文档子集中距离在 m 个词以内的二元词组,在二元词组表中列出查找到的二元词组及对应的词频,其中 m 取 0 至 5 的整数 ;

[0045] 查找二元词组表中的等同二元词组,所述等同二元词组由两个词相同、但语序不同的二元词组构成 ;在二元词组表中删除等同二元词组中词频较低的二元词组,并将词频较高的二元词组的词频更新为等同二元词组的词频 ;

[0046] 在二元词组表中删除具有停词表中的停词的二元词组 ;

[0047] 提取二元词组表中词频最高的 n 个二元词组作为文档子集的候选主题词,n 取 10 至 100 的整数。

[0048] 本发明的另一个目的在于克服现有技术中的不足,提供了一种交互式的检索式生成系统,以辅助用户生成复杂检索式,帮助专业检索领域的专业检索人员生成更为精确的检索式。

[0049] 为实现上述目的,本发明采用的技术方案为 :一种交互式的检索式生成系统,包括 :

[0050] 第一输入模块,用于接收用户输入的一级检索式 ;

[0051] 第一检索模块,用于在接收到所述第一输入模块提供的一级检索式后获取与所述一级检索式相对应的一级检索文档集 ;

[0052] 聚类模块,用于对所述一级检索文档集进行主题聚类,获取与各主题一一对应的一级检索文档子集 ;

[0053] 主题词提取模块,用于提取所述一级检索文档子集中的主题词 ;

[0054] 第一输出模块,用于将所述主题词及其对应的主题提供给用户,并提示用户利用所述主题词确定与其对应的主题相关的二级检索式。

[0055] 优选的是,所述系统还包括 :

[0056] 第二输入模块,用于接收用户输入的二级检索式 ;

[0057] 第二检索模块,用于在接收到所述第二输入模块提供的二级检索式后,获取与各

二级检索式一一对应的二级检索文档集；

[0058] 组合模块，用于对二级检索式进行两两组合，使每组两个二级检索式成为两个待验证检索式；以及，

[0059] 交叉验证模块，用于对两个待验证检索式进行交叉验证，所述交叉验证模块包括：

[0060] 统计单元，用于获取与两个待验证检索式一一对应的两个检索文档集的交叉文档集；

[0061] 比较单元，用于将所述交叉文档集中文档的数目与设定阈值进行比较，如果交叉文档集中文档的数目大于设定阈值，则确定对两个待验证检索式进行优化。

[0062] 优选的是，所述比较单元还用于在确定对两个待验证检索式进行优化后，将所述交叉文档集输入至所述聚类模块；所述聚类模块还用于获取与所述交叉文档集的各主题一一对应的交叉文档子集；所述主题词提取模块还用于提取交叉文档子集的主题词；所述系统还包括：

[0063] 第二输出模块，用于将交叉文档子集的主题词及其对应的主题提供给用户，同时提示用户利用交叉文档子集的主题词优化两个待验证检索式，确定两个优化检索式。

[0064] 优选的是，所述系统还包括：

[0065] 第三输入模块，用于接收用户输入的两个优化检索式，以及用于将两个优化检索式作为两个待验证检索式输入至所述交叉验证模块；

[0066] 第三检索模块，用于在接收到所述第三输入模块提供的两个优化检索式后，获取与两个优化检索式一一对应的两个优化检索文档集。

[0067] 优选的是，所述聚类模块包括：

[0068] 特征向量计算单元，用于设定待聚类文档集为被拆分类，提取被拆分类中各文档的聚类特征，获取与各文档一一对应的特征向量；

[0069] 种子文档确定单元，用于在被拆分类中选择两个密度最大的文档作为种子文档，文档的密度为在被拆分类中与文档的余弦距离小于 0.5 的文档的个数，其中，文档间的余弦距离为文档的特征向量间的余弦距离；

[0070] K 均值聚类单元，用于以两个所述种子文档作为种子，用 K 均值聚类算法将所述被拆分类拆分为两个主题；以及，

[0071] 判断单元，用于判断两个主题中是否有一个主题的文档数量小于被拆分类的预设百分比，如是则主题聚类结束，否则将两个主题中数量较多的一个主题作为被拆分类输入至种子文档确定单元。

[0072] 优选的是，所述特征向量计算单元用于对每篇文档计算特征词典中每个词的 TF-IDF 值作为聚类特征，获取 k 维的特征向量，其中，k 等于特征词典中词的数量。

[0073] 优选的是，所述聚类模块还包括特征词典获取模块，所述特征词典生成模块包括：

[0074] 背景特征统计单元，用于统计整个检索库中所有文档所包含的词及对应的词频，作为背景特征；

[0075] 候选特征统计单元，用于统计一级检索文档集中所有文档所包含的词及对应的词频，作为候选特征；

[0076] 差异度计算单元，用于计算候选特征与背景特征之间的词的词频的差异度，选择差异度最大的预设数量的词构成所述特征词典。

[0077] 优选的是，所述主题词提取模块包括：

[0078] 候选主题词提取单元，用于提取各文档子集的候选主题词；

[0079] 分配单元，用于将同一候选主题词分配给词频最高的文档子集；

[0080] 主题词选择单元，用于针对每个文档子集选择词频最高的6至12个候选主题词作为主题词。

[0081] 优选的是，所述候选主题词提取单元包括：

[0082] 二元词组查找子单元，用于查找文档子集中距离在m个词以内的二元词组，在二元词组表中列出查找到的二元词组及对应的词频，其中m取0至5的整数；

[0083] 合并子单元，用于查找二元词组表中的等同二元词组，所述等同二元词组由两个词相同、但语序不同的二元词组构成；在二元词组表中删除等同二元词组中词频较低的二元词组，并将词频较高的二元词组的词频更新为等同二元词组的词频；

[0084] 删除子单元，用于在二元词组表中删除具有停词表中的停词的二元词组；

[0085] 候选主题词选择子单元，用于提取二元词组表中词频最高的n个二元词组作为文档子集的候选主题词，n取10至100的整数。

[0086] 本发明的有益效果在于，本发明提出的交互式的检索式生成方法及系统，可以辅助用户生成复杂检索式，帮助专业检索领域的专业检索人员生成更为精确的检索式；可让计算机等参与到检索式的生成过程中，通过文本挖掘的技术为检索人员提供作为候选检索词的主题词，辅助检索人员生成更加复杂及精确的检索式；还可以辅助检索人员对检索结果进行验证，快速对检索性能进行有效评估，获得更加精确的检索结果；进一步地，通过本发明的方法生成的检索式在同一类数据上可以重复利用，大大减轻了检索人员的负担，提高了检索的准确率。

附图说明

[0087] 图1示出了根据本发明所述交互式的检索式生成方法的一种实施方式的流程图；

[0088] 图2示出了根据本发明所述交互式的检索式生成方法的另一种实施方式的流程图；

[0089] 图3示出了进行图2中所示交叉验证的方法；

[0090] 图4示出了根据本发明所述交互式的检索式生成方法的第三种实施方式的流程图；

[0091] 图5示出了实现图4所示第三种实施方式的一个具体实施步骤；

[0092] 图6示出了根据本发明所述交互式的检索式生成系统的一种实施结构；

[0093] 图7示出了根据本发明所述交互式的检索式生成系统的另一种实施结构；

[0094] 图8示出了根据本发明所述交互式的检索式生成系统的第三种实施结构；

[0095] 图9示出了根据本发明所述交互式的检索式生成系统的第四种实施结构。

具体实施方式

[0096] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终

相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。

[0097] 如图 1 所示，本发明的交互式的检索式生成方法包括：

[0098] 步骤 S1：在接收到用户输入的一级检索式后获取与所述一级检索式相对应的一级检索文档集。

[0099] 步骤 S2：对所述一级检索文档集进行主题聚类，获取与各主题一一对应的一级检索文档子集，即按照确定的各主题将一级检索文档集拆分成各一级检索文档子集；该主题聚类可采用现有的主题聚类方法。

[0100] 步骤 S3：提取所述一级检索文档子集中的主题词。

[0101] 步骤 S4：将所述主题词及其对应的主题提供给用户，并提示用户利用所述主题词确定与其对应的主题相关的二级检索式。在此，用户可对每一个其认为有意义的主题提供一个二级检索式，用户可以选取该主题下的与自身的检索目的相关的主题词，以“与”、“或”、“非”、“near”（即“邻近”）等逻辑运算符进行组合，获取该主题的二级检索式；以上的逻辑运算符“near”表示相“near”的两个关键词在文档中的距离在预设字数以内，该预定字数通常选择为 0 至 5 间的整数，最常用的选择是 3。

[0102] 在此，主题词的提取不仅可以告诉用户各主题（或者称为子类）中的文档内容，而且可以帮助用户生成与各主题相关的检索式。在人类语言中，二元词组比单个词表达意思更加明确，比如“开通 - 流量”比单独的“流量”更加清晰，但是如果用“开通 & 流量”这样的检索式在检索库中进行检索，将会产生很多的虚警，比如一个文档中出现了“开通来电显示”，同时“取消流量”，就会被误检到，如果限定“开通”和“流量”两个词之间的距离，则可以大大的提高准确率，因此，本发明为用户提供了“near”这个逻辑运算符，用以限定两个词之间的距离。

[0103] 本发明的方法还可在以上提供的一次交互的基础上进行更深层次的交互，为此，如图 2 所示，该方法还包括：

[0104] 步骤 S5：在接收到用户输入的二级检索式后，获取与各二级检索式一一对应的二级检索文档集。

[0105] 步骤 S6：对二级检索式进行两两组合，使每组两个二级检索式成为两个待验证检索式，以依次对各组二级检索式进行交叉验证；举例说明该处所指的两两组合的含义，例如用户输入三个二级检索式，分别为二级检索式 a、b、c，则组合形式为：第一组：二级检索式 a、b；第二组：二级检索式 a、c；第三组：二级检索式 b、c。

[0106] 步骤 S7：对两个待验证检索式进行交叉验证，其中，如图 3 所示，对两个待验证检索式进行交叉验证的方法包括：

[0107] 步骤 S71：获取与两个待验证检索式一一对应的两个检索文档集的交叉文档集，其中，如果待验证检索式为二级检索式，则与其对应的检索文档集则为二级检索文档集，如果待验证检索式为在二级检索式基础上优化得到的优化检索式，则与其对应的检索文档集则为优化检索文档集。

[0108] 步骤 S72：判断所述交叉文档集中文档的数目是否大于设定阈值，如是则执行步骤 S73，如否则执行步骤 S74，在此，该设定阈值通常为与两个待验证检索式一一对应的两个检索文档集的总文档数的百分比，例如总文档数的 10％至 50％，设定阈值的比例越低，

检索结果越准确，最终提供的检索结果中的文档数量也会越少，但相应地检索优化速度也会降低，本实施例选择总文档数的 30％。

[0109] 步骤 S73 ：对两个待验证检索式进行优化。

[0110] 步骤 S74 ：告知用户无需对两个待验证检索式作进一步优化。

[0111] 如图 4 所示，步骤 S73 中对两个待验证检索式进行优化的方法可包括 ：

[0112] 步骤 S731 ：对交叉文档集进行主题聚类，获取与所述交叉文档集的各主题一一对应的交叉文档子集。

[0113] 步骤 S732 ：提取交叉文档子集的主题词。

[0114] 步骤 S733 ：将交叉文档子集的主题词及其对应的主题提供给用户，同时提示用户利用交叉文档子集的主题词优化两个待验证检索式，确定两个优化检索式，用户可以根据自身的检索需求将所提供的主题词通过适当的逻辑运算符加入两个待验证检索式中，以尽量降低两个优化检索式的交叉文档集中的文档的数目。例如，用户可根据交叉文档子集的各主题的主题词判断交叉文档子集的内容，如果两个待验证检索式所代表的主题分别出现在为交叉文档子集确定的两个主题中，则用户可将交叉文档子集的主题词通过各种逻辑运算符加入到两个待验证检索式，形成两个优化检索式。

[0115] 步骤 S734 ：判断用户是否输入两个优化检索式，如是则执行步骤 S745 ；如否则结束对两个待验证检索式作进一步优化 ；

[0116] 步骤 S735 ：获取与两个优化检索式一一对应的两个优化检索文档集，两个优化检索式作为两个待验证检索式执行步骤 S71。

[0117] 以下给出一种对步骤 S6 确定的一组两个二级检索式进行交叉验证的实施方法，以便于更好地理解以上的交叉验证步骤，如图 5 所示，在步骤 S6 后，赋值 i=0，之后执行以下各步骤 ：

[0118] 步骤 S7a ：获取两个二级检索式的交叉文档集。

[0119] 步骤 S7b ：判断所述交叉文档集中文档的数目是否大于设定阈值，如是则执行步骤 S7c，如否则告知用户无需对两个二级检索式作进一步优化。步骤 S7c ：对交叉文档集进行主题聚类，获取与所述交叉文档集的各主题一一对应的交叉文档子集。

[0120] 步骤 S7d ：提取交叉文档子集的主题词。

[0121] 步骤 S7e ：赋值 i=i+1。

[0122] 步骤 S7f ：将交叉文档子集的主题词及其对应的主题提供给用户，同时提示用户利用交叉文档子集的主题词优化两个二级检索式，确定两个 i 级优化检索式，在此，由于各级优化检索式均是在二级检索式的基础上进行优化得到的，因此，对各级检索式的优化均被认为是对两个二级检索式的进一步优化。

[0123] 步骤 S7g ：判断用户是否输入两个 i 级优化检索式，如是则执行步骤 S7h ；如否则结束对两个二级检索式作进一步优化 ；

[0124] 步骤 S7h ：获取与两个 i 级优化检索式一一对应的两个 i 级优化检索文档集。

[0125] 步骤 S7i ：获取两个 i 级优化检索式的交叉文档集，之后继续执行步骤 S7b。

[0126] 以下提供一种进行上述主题聚类的方法，其可包括 ：

[0127] 步骤 a1 ：设定待聚类文档集为被拆分类，提取被拆分类中文档的聚类特征，获取与文档一一对应的特征向量 ；

[0128] 步骤 a2 :在被拆分类中选择两个密度最大的文档作为种子文档,文档的密度为在被拆分类中与文档的余弦距离小于 0.5 的文档的个数,其中,文档间的余弦距离为文档的特征向量间的余弦距离 ;

[0129] 步骤 a3 :以两个所述种子文档作为种子,用 K 均值聚类算法将所述被拆分类分成两个主题 ;

[0130] 步骤 a4 :判断两个主题中是否有一个主题的文档数量小于被拆分类的预设百分比,如是则主题聚类结束,否则将两个主题中数量较多的一个主题作为被拆分类,继续执行步骤 a2。该预设百分比可根据聚类要求进行选择,本实施例选择为 10%。

[0131] 在上述步骤 a1 中,对每篇文档计算特征词典中每个词的 TF-IDF(term frequency - inverse document frequency,词频 - 逆向文档频率)值作为聚类特征,获取 k 维的特征向量,其中,k 等于特征词典中词的数量。

[0132] 本发明还提供了一种获取上述特征词典的方法,具体包括 :

[0133] 步骤 b1 :统计整个检索库中所有文档所包含的词及对应的词频,作为背景特征。

[0134] 步骤 b2 :统计一级检索文档集中所有文档所包含的词及对应的词频,作为候选特征。

[0135] 步骤 b3 :计算候选特征与背景特征之间的词的词频的差异度,选择差异度最大的预设数量的词构成所述特征词典,该预设数量通常为 300 至 500 间的整数。

[0136] 本发明还提供了一种提取经主题聚类得到的各文档子集的主题词的方法,具体包括 :

[0137] 步骤 c1 :提取各文档子集的候选主题词。

[0138] 步骤 c2 :将同一候选主题词分配给词频最高的文档子集,即在步骤 c1 中可能存在不同的文档子集具有相同候选主题词的情况,步骤 c2 即是对该种情况的处理。

[0139] 步骤 c3 :针对每个文档子集选择词频最高的 6 至 12 个候选主题词作为主题词。

[0140] 以上的提取各文档子集的候选主题词可包括 :

[0141] 步骤 c11 :查找文档子集中距离在 m 个词以内的二元词组,在二元词组表中列出查找到的二元词组及对应的词频,其中 m 取 0 至 5 的整数,本实施例选为 3。

[0142] 步骤 c12 :查找二元词组表中的等同二元词组,所述等同二元词组由两个词相同、但语序不同的二元词组构成,例如“开通—流量”与“流量—开通”即为等同二元词组 ;在二元词组表中删除等同二元词组中词频较低的二元词组,并将等同二元词组中词频较高的二元词组的词频更新为等同二元词组的词频。

[0143] 步骤 c13 :在二元词组表中删除具有停词表中的停词的二元词组,该停词表可以是人工获得的词典,词典中通常包含了一些无意义的词,例如“嗯”,“啊”等。

[0144] 步骤 c14 :提取二元词组表中词频最高的 n 个二元词组作为文档子集的候选主题词,n 取 10 至 100 的整数,本实施例中 n 取 50 个。

[0145] 本发明还提供了一种可以实现上述方法的一种交互式的检索式生成系统,如图 6 所示,该系统包括第一输入模块 1、第一检索模块 2、聚类模块 3、主题词提取模块 4 和第一输出模块 5,其中,第一输入模块 1 用于接收用户输入的一级检索式 ;第一检索模块 2 用于在接收到第一输入模块 1 提供的一级检索式后获取与一级检索式相对应的一级检索文档集 ;聚类模块 3 用于对一级检索文档集进行主题聚类获取与各主题一一对应的一级检索文档

子集；主题词提取模块 4 用于提取一级检索文档子集中的主题词；第一输出模块 5 用于将主题词及其对应的主题提供给用户，并提示用户利用所述主题词确定与其对应的主题相关的二级检索式。

[0146] 如图 7 所示，本发明的系统还可以包括第二输入模块 6、第二检索模块 12、组合模块 8 和交叉验证模块 7，其中，第二输入模块 6 用于接收用户输入的二级检索式；第二检索模块 12 用于在接收到第二输入模块 6 提供的二级检索式后，获取与各二级检索式一一对应的二级检索文档集；组合模块 8 用于对二级检索式进行两两组合，使每组两个二级检索式成为两个待验证检索式；交叉验证模块 7 用于对两个待验证检索式进行交叉验证，该交叉验证模块 7 包括统计单元 71 和比较单元 72，其中，统计单元 71 用于获取与两个待验证检索式一一对应的两个检索文档集的交叉文档集；比较单元 72 用于将交叉文档集中文档的数目与设定阈值进行比较，如果交叉文档集中文档的数目大于设定阈值，则确定对两个待验证检索式进行优化。

[0147] 图 8 所示的系统提供了一种对两个待验证检索式进行优化的具体结构，在该结构下，以上比较单元 72 还用于在确定对两个待验证检索式进行优化后，将交叉文档集输入至聚类模块 3；聚类模块 3 还用于获取与交叉文档集的各主题一一对应的交叉文档子集；主题词提取模块 4 还用于提取交叉文档子集的主题词；对于如图 8 所示的实施方式，本发明的系统还包括第二输出模块 9，第二输出模块 9 用于将交叉文档子集的主题词及其对应的主题提供给用户，同时提示用户利用交叉文档子集的主题词优化两个待验证检索式，确定两个优化检索式。

[0148] 如图 9 所示的可与用户进行进一步交互的实施方式，本发明的系统还包括：

[0149] 第三输入模块 10，用于接收用户输入的两个优化检索式，以及用于将两个优化检索式作为两个待验证检索式输入至所述交叉验证模块；

[0150] 第三检索模块 11，用于在接收到第三输入模块 10 提供的两个优化检索式后，获取与两个优化检索式一一对应的两个优化检索文档集。

[0151] 以上聚类模块可包括特征向量计算单元、种子文档确定单元、K 均值聚类单元和判断单元，其中，特征向量计算单元用于设定待聚类文档集为被拆分类，提取被拆分类中各文档的聚类特征，获取与各文档一一对应的特征向量；种子文档确定单元用于在被拆分类中选择两个密度最大的文档作为种子文档，文档的密度为在被拆分类中与文档的余弦距离小于 0.5 的文档的个数，其中，文档间的余弦距离为文档的特征向量间的余弦距离；K 均值聚类单元用于根据 K 均值聚类算法将所述被拆分类分成由种子文档决定的两个主题；判断单元用于判断两个主题中是否有一个主题的文档数量小于被拆分类的预设百分比，如是则主题聚类结束，如否，则将两个主题中数量较多的一个主题作为被拆分类输入至种子文档确定单元。

[0152] 以上特征向量计算单元具体用于对每篇文档计算特征词典中每个词的 TF-IDF 值作为聚类特征，获取 k 维的特征向量，其中，k 等于特征词典中词的数量。

[0153] 以上聚类模块还可包括特征词典获取模块，以便于获取满足用户使用要求的特征词典，该特征词典生成模块包括背景特征统计单元、候选特征统计单元和差异度计算单元，其中，背景特征统计单元用于统计整个检索库中所有文档所包含的词及对应的词频，作为背景特征；候选特征统计单元用于统计一级检索文档集中所有文档所包含的词及对应的词

频，作为候选特征；差异度计算单元用于计算候选特征与背景特征之间的词的词频的差异度，选择差异度最大的预设数量的词构成所述特征词典。

[0154] 以上主题词提取模块可包括候选主题词提取单元、分配单元和主题词选择单元，其中，候选主题词提取单元用于提取各文档子集的候选主题词；分配单元用于将同一候选主题词分配给词频最高的文档子集；主题词选择单元用于针对每个文档子集选择词频最高的 6 至 12 个候选主题词作为主题词。

[0155] 以上候选主题词提取单元可包括二元词组查找子单元、合并子单元、删除子单元和候选主题词选择子单元，其中，二元词组查找子单元用于查找文档子集中距离在 m 个词以内的二元词组，在二元词组表中列出查找到的二元词组及对应的词频，其中 m 取 0 至 5 的整数；合并子单元用于查找二元词组表中的等同二元词组（定义请参见上述说明），在二元词组表中删除等同二元词组中词频较低的二元词组，并将等同二元词组中词频较高的二元词组的词频更新为等同二元词组的词频；删除子单元用于在二元词组表中删除具有停词表中的停词的二元词组；候选主题词选择子单元用于提取二元词组表中词频最高的 n 个二元词组作为文档子集的候选主题词，n 取 10 至 100 的整数。

[0156] 以上第一输入模块 1、第二输入模块 6 和第三输入模块 10 可为同一输入模块，也可为单独设置的不同模块，这些输入模块为用户提供用于确定检索式的逻辑运算符，该逻辑运算符包括“邻近”（其定义请参见上述说明）。

[0157] 同理，第一输出模块 5 和第二输出模块 9 可为同一输出模块；第一检索模块 2、第二检索模块 12 和第三检索模块 11 也可为同一检索模块。

[0158] 以上依据图式所示的实施例详细说明了本发明的构造、特征及作用效果，以上所述仅为本发明的较佳实施例，但本发明不以图面所示限定实施范围，凡是依照本发明的构想所作的改变，或修改为等同变化的等效实施例，仍未超出说明书与图示所涵盖的精神时，均应在本发明的保护范围内。

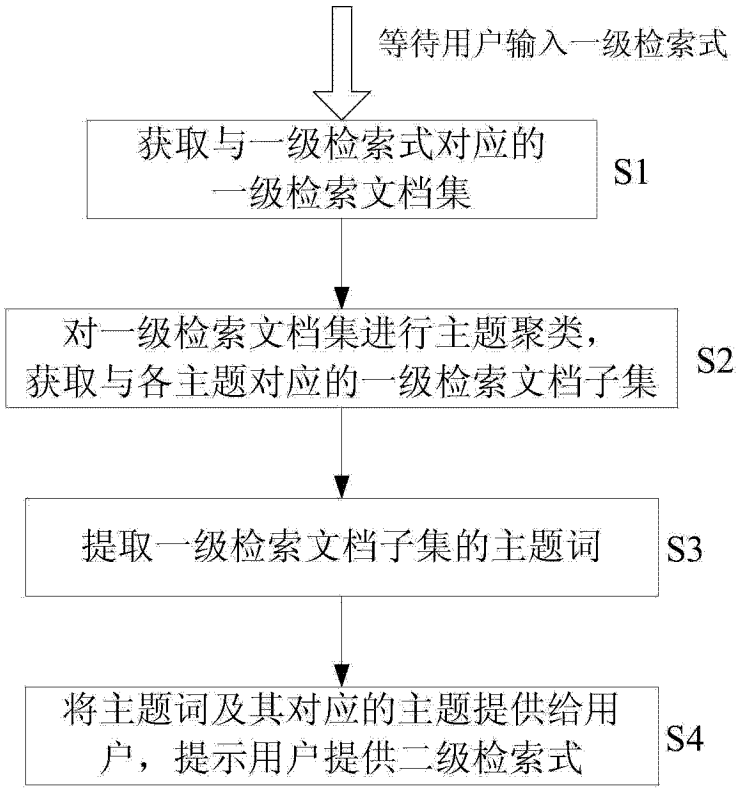

等待用户输入一级检索式
获取与一级检索式对应的一级检索文档集
S1
对一级检索文档集进行主题聚类，获取与各主题对应的一级检索文档子集
S2
提取一级检索文档子集的主题词
S3
将主题词及其对应的主题提供给用户，提示用户提供二级检索式
S4

图 1

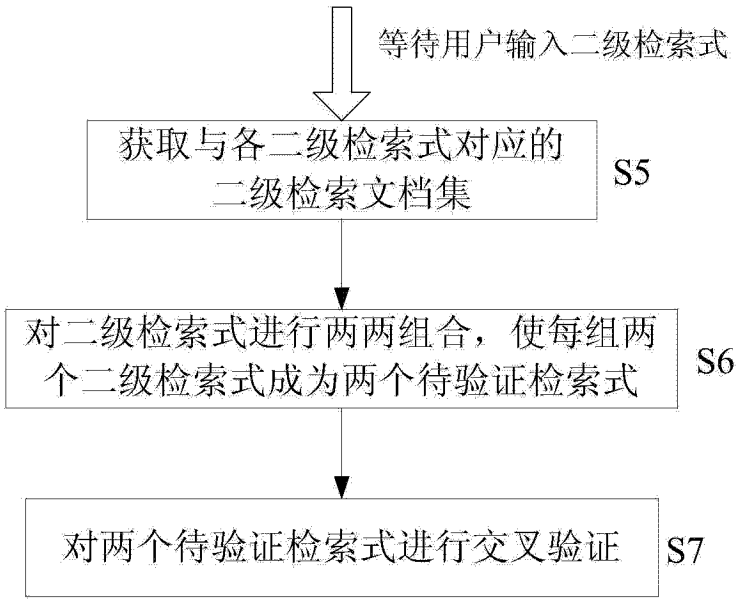

等待用户输入二级检索式
获取与各二级检索式对应的二级检索文档集
S5
对二级检索式进行两两组合，使每组两个二级检索式成为两个待验证检索式
S6
对两个待验证检索式进行交叉验证
S7

图 2

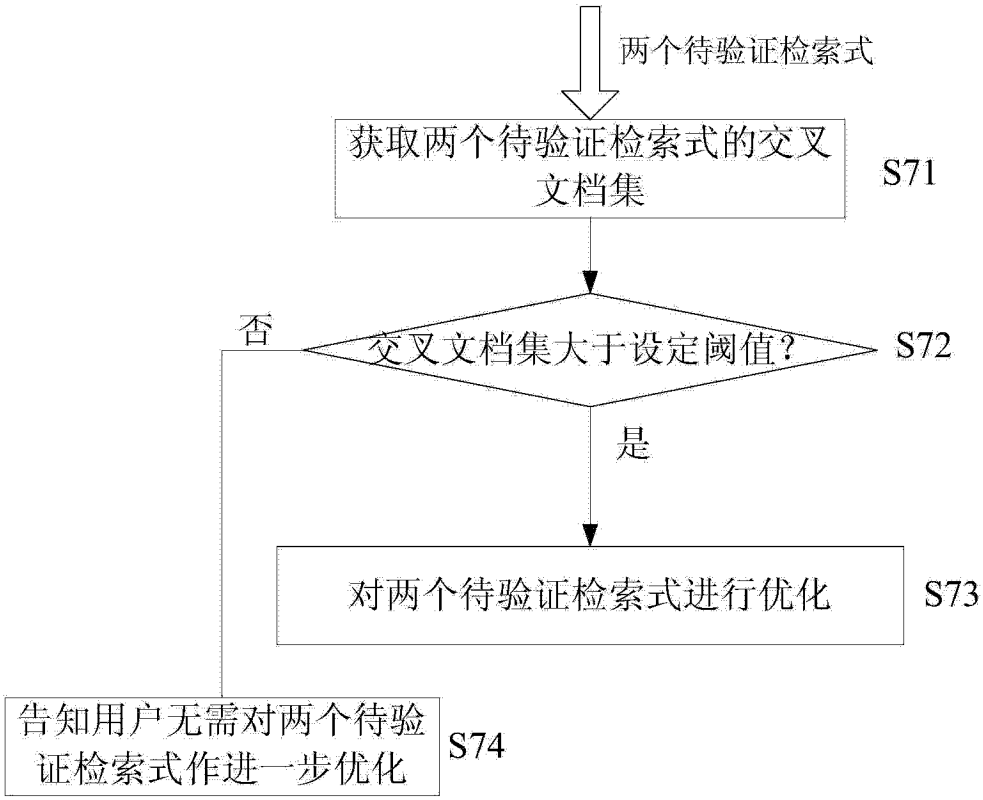

两个待验证检索式
获取两个待验证检索式的交叉文档集
S71
否
交叉文档集大于设定阈值?
S72
是
对两个待验证检索式进行优化
S73
告知用户无需对两个待验证检索式作进一步优化
S74

图 3

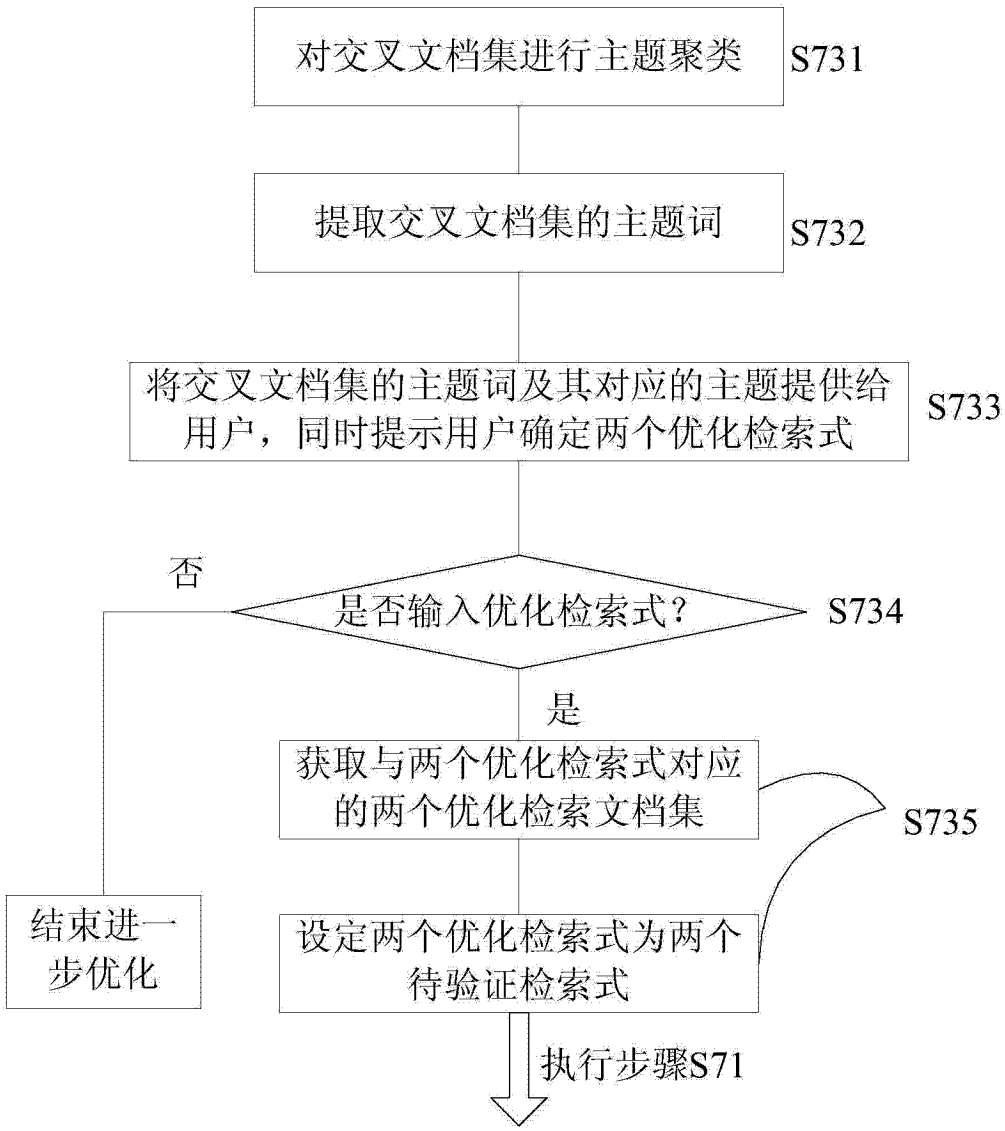

对交叉文档集进行主题聚类
S731
提取交叉文档集的主题词
S732
将交叉文档集的主题词及其对应的主题提供给用户，同时提示用户确定两个优化检索式
S733
否
是否输入优化检索式？
S734
是
获取与两个优化检索式对应的两个优化检索文档集
S735
结束进一步优化
设定两个优化检索式为两个待验证检索式
执行步骤S71

图 4

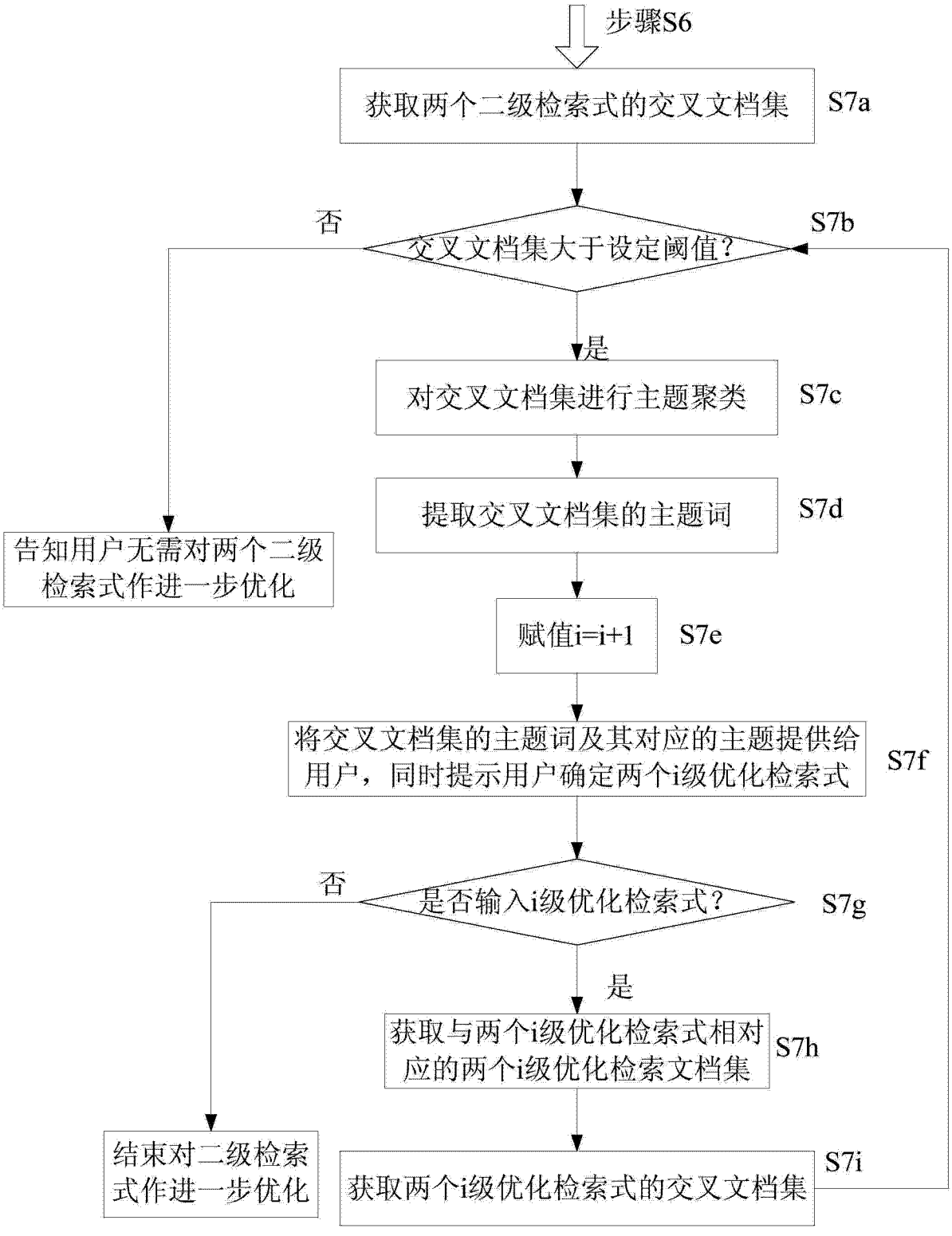

步骤S6
获取两个二级检索式的交叉文档集
S7a
否
交叉文档集大于设定阈值?
S7b
是
对交叉文档集进行主题聚类
S7c
提取交叉文档集的主题词
S7d
告知用户无需对两个二级检索式作进一步优化
赋值i=i+1
S7e
将交叉文档集的主题词及其对应的主题提供给用户，同时提示用户确定两个i级优化检索式
S7f
否
是否输入i级优化检索式?
S7g
是
获取与两个i级优化检索式相对应的两个i级优化检索文档集
S7h
结束对二级检索式作进一步优化
获取两个i级优化检索式的交叉文档集
S7i

图 5

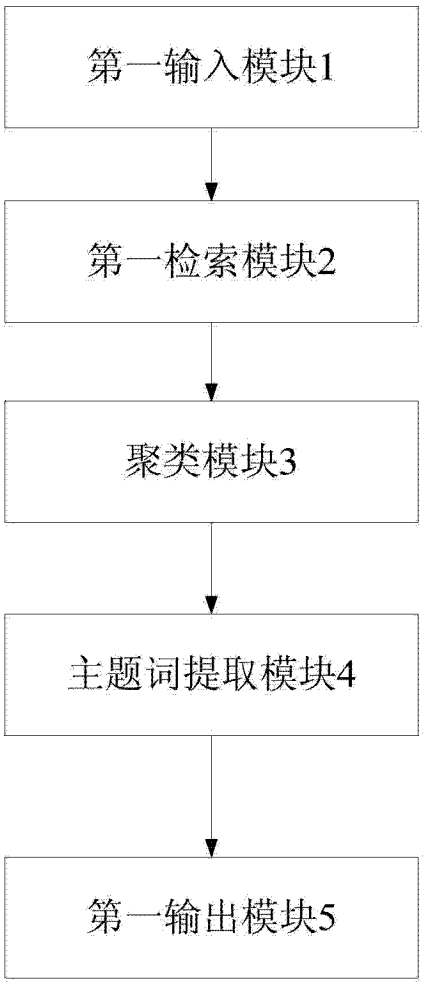

第一输入模块1
第一检索模块2
聚类模块3
主题词提取模块4
第一输出模块5

图 6

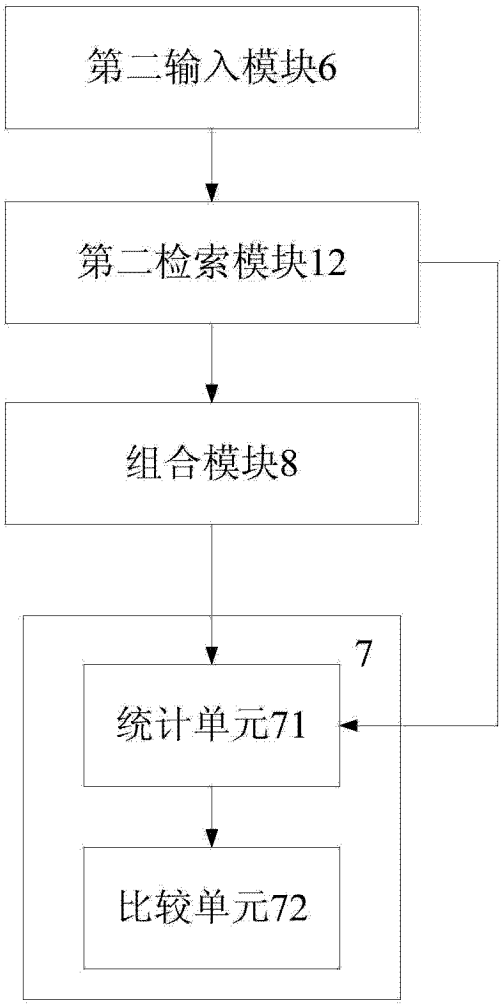

第二输入模块6
第二检索模块12
组合模块8
7
统计单元71
比较单元72

图 7

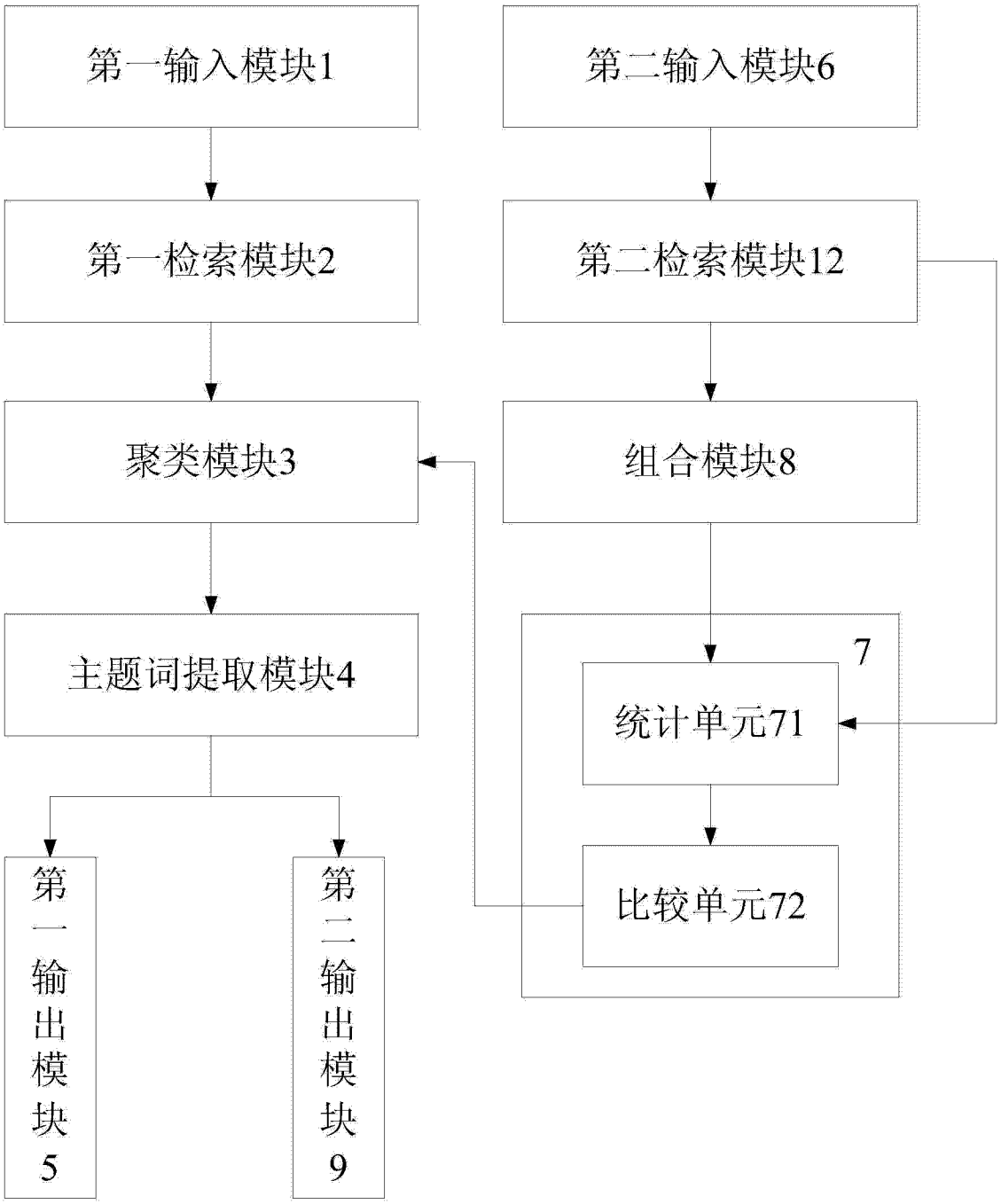

第一输入模块1
第二输入模块6
第一检索模块2
第二检索模块12
聚类模块3
组合模块8
主题词提取模块4
7
统计单元71
比较单元72
第一输出模块5
第二输出模块9

图 8

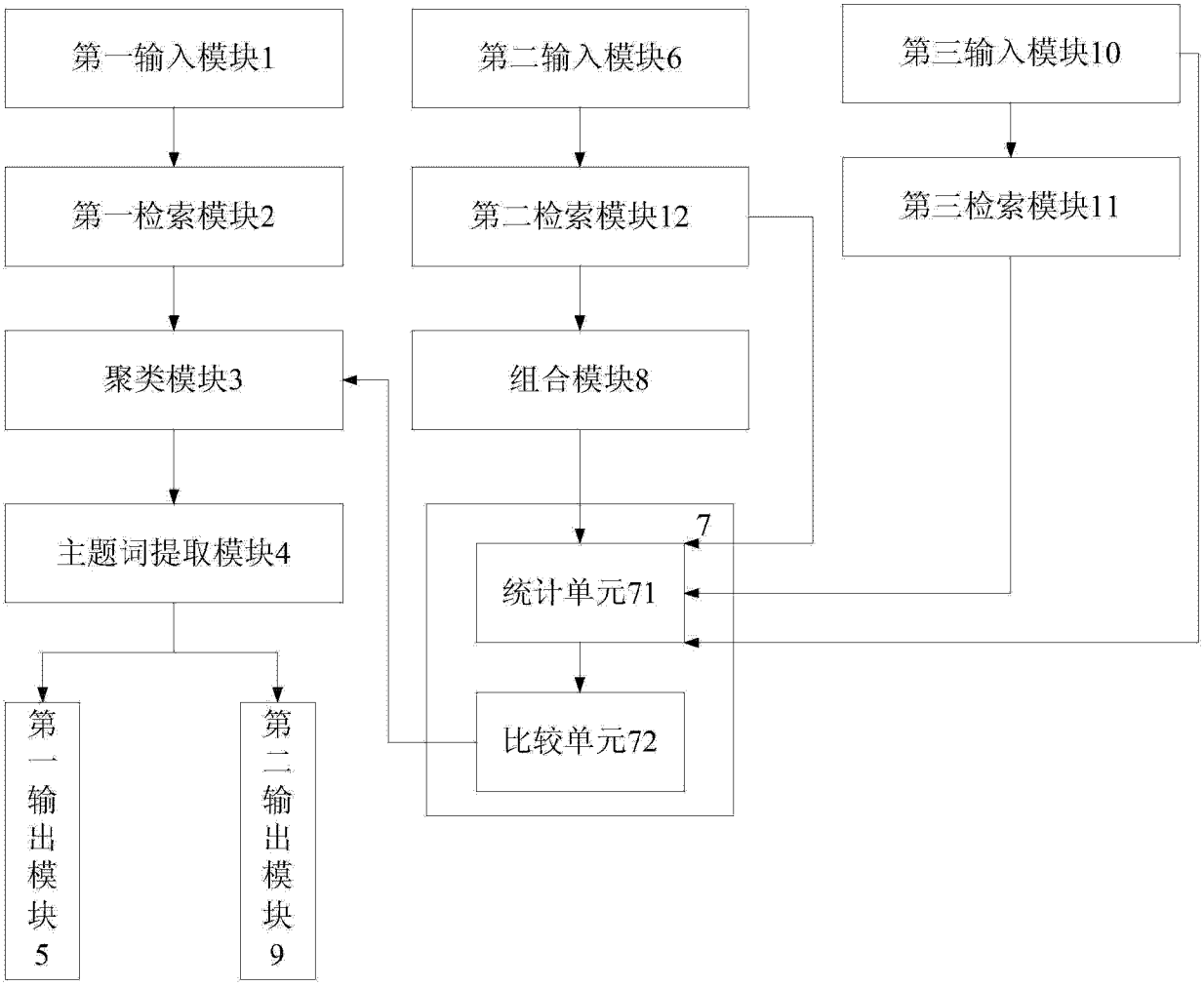

第一输入模块1
第二输入模块6
第三输入模块10
第一检索模块2
第二检索模块12
第三检索模块11
聚类模块3
组合模块8
主题词提取模块4
7
统计单元71
比较单元72
第一输出模块5
第二输出模块9

图 9