(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 103324644 A
(43)申请公布日 2013.09.25

(21)申请号 201210080590.4

(22)申请日 2012.03.23

(71)申请人 日电（中国）有限公司
地址 100191 北京市海淀区学院路 35 号世宁大厦 20 层

(72)发明人 李建强 刘春辰

(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 孔凡红

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

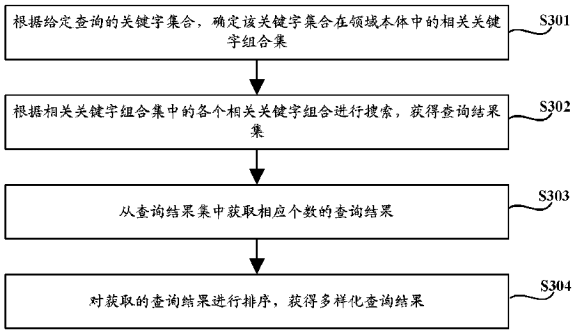
权利要求书4页 说明书8页 附图4页

(54)发明名称

一种查询结果多样化方法及装置

(57)摘要

本发明公开了一种查询结果多样化方法及装置，涉及信息检索技术，通过领域本体确定给定查询的关键字集合的相关关键字组合集，并使用这些相关关键字组合进行查询，避免使用不可靠的查询日志确定子查询关键字，从而使得多样化查询结果更加准确。



CN 103324644 A

1. 一种查询结果多样化方法，其特征在于，包括：

根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集；

根据所述相关关键字组合集中的各个相关关键字组合进行搜索，获得查询结果集；

从所述查询结果集中获取相应个数的查询结果；

对获取的查询结果进行排序，获得多样化查询结果。

2. 如权利要求 1 所述的方法，其特征在于，所述根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集，具体包括：

根据给定查询每个关键字，确定该关键字在所述领域本体中的相关关键字；

根据各个相关关键字，确定相关关键字组合集。

3. 如权利要求 2 所述的方法，其特征在于，根据各个相关关键字，确定相关关键字组合集，具体包括：

确定相关关键字组合集为：$S(Q) = \{(c_1, c_2, \ldots, c_m) | c_1 \in C_1 \&\& c_2 \in C_2 \&\& \ldots c_m \in C_m\}$，其中，$C_i$ 为给定查询中 m 个关键字的第 i 个关键字的相关关键字集合。

4. 如权利要求 1 所述的方法，其特征在于，在所述根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集后，还包括：

对于相关关键字组合集中的每个相关关键字组合，从领域本体中抽取连接各个关键字的最小子图，所述最小子图为实现连接各关键字的领域本体子图中，边数最少的子图；

所述根据相关关键字组合集中的各个相关关键字组合进行搜索，获得查询结果集，具体包括：

对于每个最小子图，确定由该最小子图中包括的关键字及其它节点构成的子查询；

根据每个子查询中包括的关键字及其它节点进行搜索，获得与最小子图数量相同的子查询结果集；

确定查询结果集为各个子查询结果集构成的集合。

5. 如权利要求 4 所述的方法，其特征在于，所述从所述查询结果集中获取相应个数的查询结果，具体包括：

根据每个子查询与给定查询的相关程度，从每个子查询结果集中获取相应个数的查询结果；

合并从各个子查询结果集中获取的查询结果。

6. 如权利要求 5 所述的方法，其特征在于，所述根据每个子查询与给定查询的相关程度，从每个子查询结果集中获取相应个数的查询结果，具体包括：

确定每个最小子图的子图权重为：$w = \sum_{i=1}^{m} r_i \Big/ m \times |E|$，其中 m 为查询关键字的数量，ri 为根据所述领域本体确定的相关关键字与相应的关键字的匹配值，E 为该子图包括的边的数量；

根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果。

7. 如权利要求 6 所述的方法，其特征在于，所述根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果，具体包括：

从该最小子图对应的子查询结果集中获取的查询结果为与该最小子图关联程度最大的前 a 个查询结果，a 为不大于当前最小子图的子图权重与所有最小子图的子图权重和的比值的最大整数。

8. 如权利要求 4 所述的方法，其特征在于，所述对获取的查询结果进行排序，获得多样化查询结果，具体包括：

对于每个查询结果，确定该查询结果与对应的最小子图的关联程度值；

对于每个查询结果，根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重；

根据所述查询结果的权重，对获取的查询结果进行排序，获得多样化查询结果。

9. 如权利要求 8 所述的方法，其特征在于，所述根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重，具体包括：

确定该查询结果的权重为该查询结果与对应的最小子图的关联程度值与该最小子图的子图权重的乘积。

10. 如权利要求 8 所述的方法，其特征在于，所述根据所述查询结果的权重，对获取的查询结果进行排序，具体包括：

直接按照所述查询结果的权重大小，对获取的查询结果进行排序；或者

确定权重最大的查询结果为排在第一位的查询结果，并确定每两个查询结果之间的相似程度值；对于其它查询结果，确定每个查询结果的相似权重为：$s\times\prod_{d'\in D}(1-similarity(d,d'))$，其中，s 为查询结果的权重，d 为当前查询结果，D 为已排序的查询结果构成的集合，similarity(d,d′) 为 d 和 d’的相似程度值；按照所述相似权重的大小，对除排在第一位的查询结果外的查询结果进行递归排序。

11. 一种查询结果多样化装置，其特征在于，包括：

关键字确定单元，用于根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集；

查询单元，用于根据所述相关关键字组合集中的各个相关关键字组合进行搜索，获得查询结果集；

查询结果获取单元，用于从所述查询结果集中获取相应个数的查询结果；

排序单元，用于对获取的查询结果进行排序，获得多样化查询结果。

12. 如权利要求 11 所述的装置，其特征在于，所述关键字确定单元具体用于：

根据给定查询每个关键字，确定该关键字在所述领域本体中的相关关键字；

根据各个相关关键字，确定相关关键字组合集。

13. 如权利要求 12 所述的装置，其特征在于，所述关键字确定单元根据各个相关关键字，确定相关关键字组合集，具体包括：

确定相关关键字组合集为：$S(Q)=\{(c_1,c_2,\ldots,c_m)\,|\,c_1\in C_1\&\&c_2\in C_2\&\&\ldots c_m\in C_m\}$，其中，$C_i$ 为给定查询中 m 个关键字的第 i 个关键字的相关关键字集合。

14. 如权利要求 11 所述的装置，其特征在于，所述关键字确定单元还用于：

在所述根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集后：

对于相关关键字组合集中的每个相关关键字组合，从领域本体抽取连接各个关键字的

最小子图，所述最小子图为实现连接各关键字的领域本体子图中，边数最少的子图；

所述查询单元具体用于：

对于每个最小子图，确定由该最小子图中包括的关键字及其它节点构成子查询；

根据每个子查询中包括的关键字及其它节点进行搜索，获得与最小子图数量相同的子查询结果集；

确定查询结果集为各个子查询结果集构成的集合。

15. 如权利要求 14 所述的装置，其特征在于，所述查询结果获取单元具体用于：

根据每个子查询给定查询的相关程度，从每个子查询结果集中获取相应个数的查询结果；

合并从各个子查询结果集中获取的查询结果。

16. 如权利要求 15 所述的装置，其特征在于，所述查询结果获取单元具体用于：

确定每个最小子图的子图权重为：$w = \sum_{i=1}^{m} r_i \Big/ m \times |E|$，其中 m 为查询关键字的数量，ri 为根据所述领域本体确定的相关关键字与相应的关键字的匹配值，E 为该子图包括的边的数量；

根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果；

合并从各个子查询结果集中获取的查询结果。

17. 如权利要求 16 所述的装置，其特征在于，所述查询结果获取单元根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果，具体包括：

从该最小子图对应的子查询结果集中获取的查询结果为与该最小子图关联程度最大的前 a 个查询结果，a 为不大于当前最小子图的子图权重与所有最小子图的子图权重和的比值的最大整数。

18. 如权利要求 14 所述的装置，其特征在于，所述排序单元具体用于：

对于每个查询结果，确定该查询结果与对应的最小子图的关联程度值；

对于每个查询结果，根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重；

根据所述查询结果的权重，对获取的查询结果进行排序，获得多样化查询结果。

19. 如权利要求 18 所述的装置，其特征在于，所述排序单元根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重，具体包括：

确定该查询结果的权重为该查询结果与对应的最小子图的关联程度值与该最小子图的子图权重的乘积。

20. 如权利要求 18 所述的装置，其特征在于，所述排序单元根据所述查询结果的权重，对获取的查询结果进行排序，具体包括：

直接按照所述查询结果的权重大小，对获取的查询结果进行排序；或者

确定权重最大的查询结果为排在第一位的查询结果，并确定每两个查询结果之间的相似程度值；对于其它查询结果，确定每个查询结果的相似权重为：$s \times \prod_{d' \in D} (1 - similarity(d, d'))$，

其中，s 为查询结果的权重，d 为当前查询结果，D 为已排序的查询结果构成的集合，similarity(d,d′) 为 d 和 d’的相似程度值；按照所述相似权重的大小，对除排在第一位的查询结果外的查询结果进行递归排序。

一种查询结果多样化方法及装置

技术领域

[0001] 本发明涉及信息检索技术，尤其涉及一种查询结果多样化方法及装置。

背景技术

[0002] 传统的信息检索技术主要是通过对文献检索进行后处理或重新排序的步骤实现多样化，如搜索结果的聚类或分类，根据均值－方差分析进行重新排序的结果等。

[0003] 而随着信息检索技术的发展，用户对信息检索的搜索结果多样化和查询消歧的要求也越来越高。其中，搜索结果多样化是指：用户输入的查询关键字可能有多个解释，在获得查询结果时，应该产生包括这些不同解释的结果，搜索结果多样化的目的是通过平衡搜索结果的相关性和新颖性，最大限度地减少用户不满的风险。查询消歧是指：根据用户的输入的关键字确定所有可能的查询意图，并通过更准确的方式表示这些意图。

[0004] 查询消歧作为一种新的方式支持搜索多样化，有效地节省了计算成本并使结果更容易理解，尤其是当结果规模较大的时候。现有技术中，主要采用了对查询日志的统计分析（或机器学习等）实现多元化搜索。

[0005] 具体的，目前进行查询结果多样化的方法使用查询－查询的转化形式，如图 1 所示，包括：

[0006] 步骤 S101、对于给定的查询 Q，根据查询日志的分析大样本生成 k 个相关查询 R(Q)；

[0007] 步骤 S102、通过从每个查询结果集提取 n/(k+1) 个结果获得初始 DOC 列表（文档用户的数量可以视为 n）；

[0008] 步骤 S103、通过相关反馈方法重排序初始 DOC 列表。

[0009] 相应的搜索结果多样化装置如图 2 所示，包括：

[0010] 查询单元 201，用于存储用户的查询关键字；

[0011] 查询日志存储单元 202，用于存储用户的查询日志；

[0012] 查询消歧单元 203，用于根据用户的查询关键字和查询日志确定与目标查询相关的查询关键字；

[0013] 子查询存储单元 204，用于存储和目标查询相关的查询关键字；

[0014] 文档存储单元 205，用于存储所搜索的文档；

[0015] 关键字搜索单元 206，用于使用子查询的关键字搜索文档存储单元 205 中的文档；

[0016] 子查询结果存储单元 207，用于存储对每个子查询进行搜索的查询结果；

[0017] 查询结果合并单元 208，用于对各查询结果进行合并；

[0018] 查询结果存储单元 209，用于存储合并后的查询结果；

[0019] 查询结果排队单元 210，用于对合并后的查询结果进行排队处理；

[0020] 多样化排名列表存储单元 211，用于存储对目标查询的最终多样化查询结果。

[0021] 具体的，例如，用于给出查询关键字“window”，目标查询为 q ＝ (window)，则根据该查询关键字和查询日志获得子查询的关键字“window XP”“house window”.....，则 q

的子查询集合为 R(q) ＝{(q_1, q, window XP), (q_2, q, house window).....},根据对目标查询 q 进行搜索以及对子查询集合为 R(q) 中的各个子查询进行搜索,分别获得文档列表,形成文档列表集合 S(q) ＝{(q, document list1), (q_1, document list2), (q_2, document list3).....},从每个文档列表中选取 n/(k+1) 个数的文档,形成对于 q 的新的查询结果集合 RF(q),其中,n 表示结果规模,为预先设定的值,k 表示子查询的数量,根据文档和用户兴趣的匹配程度,对 RF(q) 中的文档进行排序,获得用户查询的多样化查询结果。

[0022] 根据上述查询结果多样化的方法可知,现有技术中是基于查询日志来确定子查询集合的,但是,本发明的发明人发现,由于查询日志是基于用户输入查询关键字生成的,而查询关键字并不能准确代表当时用户实际的查询意图,同时,对于企业搜索等某些搜索环境,查询日志不可用或查询日志的规模不足以支持查询消歧,所以,查询日志是不可靠的数据来源,导致查询结果多样化后产生的查询结果并不准确。

发明内容

[0023] 本发明实施例提供一种查询结果多样化方法及装置,以获得较准确的多样化查询结果。

[0024] 一种查询结果多样化方法,包括 :

[0025] 根据给定查询的关键字集合,确定该关键字集合在领域本体中的相关关键字组合集 ;

[0026] 根据所述相关关键字组合集中的各个相关关键字组合进行搜索,获得查询结果集 ;

[0027] 从所述查询结果集中获取相应个数的查询结果 ;

[0028] 对获取的查询结果进行排序,获得多样化查询结果。

[0029] 一种查询结果多样化装置,包括 :

[0030] 关键字确定单元,用于根据给定查询的关键字集合,确定该关键字集合在领域本体中的相关关键字组合集 ;

[0031] 查询单元,用于根据所述相关关键字组合集中的各个相关关键字组合进行搜索,获得查询结果集 ;

[0032] 查询结果获取单元,用于从所述查询结果集中获取相应个数的查询结果 ;

[0033] 排序单元,用于对获取的查询结果进行排序,获得多样化查询结果。

[0034] 本发明实施例提供一种查询结果多样化方法及装置,通过领域本体确定给定查询的关键字集合的相关关键字组合集,并使用这些相关关键字组合进行查询,避免使用不可靠的查询日志确定子查询关键字,从而使得多样化查询结果更加准确。

附图说明

[0035] 图 1 为现有技术中查询结果多样化方法流程图 ;

[0036] 图 2 为现有技术中查询多样化装置结构示意图 ;

[0037] 图 3 为本发明实施例提供的查询结果多样化方法流程图 ;

[0038] 图 4 为本发明实施例提供的最小子图获取方法流程图 ;

[0039] 图 5 为本发明实施例提供的查询结果集确定方法流程图 ;

[0040] 图 6 为本发明实施例提供的查询结果获取方法流程图；

[0041] 图 7 为本发明实施例提供的排序方法流程图；

[0042] 图 8 为本发明实施例提供的根据相似程度进行排序的方法流程图；

[0043] 图 9 为本发明实施例提供的查询结果多样化装置结构示意图。

具体实施方式

[0044] 本发明实施例提供一种查询结果多样化方法及装置，通过领域本体确定给定查询的关键字集合的相关关键字组合集，并使用这些相关关键字组合进行查询，避免使用不可靠的查询日志确定子查询关键字，从而使得多样化查询结果更加准确。

[0045] 如图 3 所示，本发明实施例提供的查询结果多样化方法包括：

[0046] 步骤 S301、根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集；

[0047] 步骤 S302、根据相关关键字组合集中的各个相关关键字组合进行搜索，获得查询结果集；

[0048] 步骤 S303、从查询结果集中获取相应个数的查询结果；

[0049] 步骤 S304、对获取的查询结果进行排序，获得多样化查询结果。

[0050] 由于通过领域本体来进行各个相关关键字的确定，所以使得相关关键字的选取更加准确，更接近用户的意图，进而使得多样化查询结果更加准确，其中，领域本体为专业性的本体，描述的是特定领域中的概念和概念之间的关系，提供了某个专业学科领域中概念的词表以及概念间的关系，或在该领域里占主导地位的理论。

[0051] 具体的，步骤 S301 中，可以先根据给定查询每个关键字，确定该关键字在所述领域本体中的相关关键字；再根据各个相关关键字，确定相关关键字组合集。所确定的相关关键字组合集为：$S(Q) = \{(c_1, c_2, \ldots, cm) | c_1 \in C_1 \&\& c_2 \in C_2 \&\& \ldots c_m \in C_m\}$，其中，$C_i$ 为给定查询中 m 个关键字的第 i 个关键字的相关关键字集合。

[0052] 在确定关键字在领域本体中的相关关键字时，可以确定领域本体中包括该关键字的概念为相关关键字，也可以确定领域本体中与该关键字相关的相关节点作为相关关键字，当然，本领域技术人员也可以根据其它方式从领域本体中确定相关关键字。

[0053] 为了能够使得查询结果更加准确，可以进一步对相关关键字以及给定查询中的关键字的组合进行筛选，从而获得更加符合用户意图的关键字组合。

[0054] 具体的，在步骤 S301 根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集后，还包括：

[0055] 对于相关关键字组合集中的每个相关关键字组合，从领域本体中抽取连接各个关键字的最小子图，其中，最小子图为实现连接各关键字的领域本体子图中，边数最少的子图。

[0056] 如图 4 所示，假设相关关键字组合中包括 5 个关键字，所抽取的子图中，连接了全部 5 个关键字，且边数最少。

[0057] 此时，如图 5 所示，在步骤 S302 中，根据相关关键字组合集中的各个相关关键字组合进行搜索，获得查询结果集，具体包括：

[0058] 步骤 S501、对于每个最小子图，确定由该最小子图中包括的关键字及其它节点构

成子查询；

[0059] 步骤 S502、根据每个子查询中包括的关键字及其它节点进行搜索，获得与最小子图数量相同的子查询结果集；

[0060] 步骤 S503、确定查询结果集为各个子查询结果集构成的集合。

[0061] 例如，用户输入查询关键字，其中包括 m 个关键字，为 $Q = \{k_1, \ldots\ldots, k_m\}$，对于任一个关键字 k_i 都能在领域本体中确定一组相关的关键字 $C_i = \{c_{i1}, c_{i2}, \ldots\ldots, c_{ini}\}$，该组关键字包括 ni 个关键字，根据领域本体还可以得到每个相关关键字与 k_i 的相关程度值 $R_i = \{r_{i1}, r_{i2}, \ldots\ldots, r_{ini}\}$，此时，对于用户输入的查询关键字可以确定出 $\prod_{i=1}^{m} n_i$ 个查询组合，$S(Q) = \{(c1, c2, \ldots, cm) | c1 \in C1 \&\& c2 \in C2 \&\& \ldots cm \in Cm\}$。

[0062] 对于每个子查询，可以根据领域本体确定查询语义图，该查询语义图中包括该子查询中的各个关键字，每个关键字都作为查询语义图的节点，为使得各关键字能够连接起来，该查询语义图中也包括其它节点。对于每个查询语义图，获取连接各个关键字的最小子图，其中，最小子图为实现连接各关键字的子图中，边的条数最少的子图。

[0063] 在获取最小子图时，可以在查询语义图中随机选取一个关键字，遍历该关键字连接其它节点的每条路径，选择与目标节点之间最短的路径作为最小子图中的路径，直至确定出连接各个关键字的最小子图，若两个节点之间具有两条边数相同的路径，则可以随机选择一条。

[0064] 在步骤 S303 中，从查询结果集中获取相应个数的查询结果，可以从每个子查询的子查询结果集中获取设定个数的查询结果，也可以进一步根据子查询关键字与查询关键字的相关程度，从查询结果集中获取相应个数的查询结果，从而使得相关程度高的查询结果数量较多，更容易与用户的查询意图匹配。

[0065] 具体的，如图 6 所示，根据每个子查询与给定查询的相关程度，从每个子查询结果集中获取相应个数的查询结果，具体包括：

[0066] 步骤 S601、确定每个最小子图的子图权重，该子图权重为：$w = \sum_{i=1}^{m} r_i \Big/ m \times |E|$，其中 m 为查询关键字的数量，ri 为根据领域本体确定的相关关键字与相应的关键字的匹配值，E 为该子图包括的边的数量；

[0067] 步骤 S602、根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果。

[0068] 在步骤 S602 中，根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果，可以具体为：

[0069] 从该最小子图对应的子查询结果集中获取的查询结果为与该最小子图关联程度最大的前 a 个查询结果，a 为当前最小子图的子图权重与所有最小子图的子图权重和的比值。

[0070] 进一步，为使得用户能够更方便的看到较符合查询意图的查询结果，本发明实施例提供相应的对查询结果排序的方法，此时，如图 7 所示，步骤 S304 对获取的查询结果进行排序，获得多样化查询结果，具体包括：

[0071] 步骤 S701、对于每个查询结果，确定该查询结果与对应的最小子图的关联程度

值；

[0072] 步骤 S702、对于每个查询结果，根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重；

[0073] 步骤 S703、根据查询结果的权重，对获取的查询结果进行排序，获得多样化查询结果。

[0074] 其中，步骤 S702 中，根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重，具体包括：

[0075] 确定该查询结果的权重为该查询结果与对应的最小子图的关联程度值与该最小子图的子图权重的乘积。

[0076] 进一步，在步骤 S703 中，根据查询结果的权重，对获取的查询结果进行排序，可以直接按照查询结果的权重大小，对获取的查询结果进行排序；也可以进一步考虑查询结果之间的相似性，使得用户能够较方便的获取多样化的查询结果，此时，如图 8 所示，步骤 S703 具体包括：

[0077] 步骤 S801、确定权重最大的查询结果为排在第一位的查询结果，并确定每两个查询结果之间的相似程度值；

[0078] 步骤 S802、对于其它查询结果，确定每个查询结果的相似权重为：$s\times\prod_{d'\in D}(1-similarity(d,d'))$，其中，s 为查询结果的权重，d 为当前查询结果，D 为已排序的查询结果构成的集合，similarity(d，d′) 为 d 和 d’ 的相似程度值；

[0079] 步骤 S803、按照相似权重的大小，对除排在第一位的查询结果外的查询结果进行递归排序。

[0080] 下面通过一个具体实例对本发明实施例提供的查询结果多样化方法进行说明：

[0081] 若用户给定查询的关键字为“牡丹”、“北京”时，可以通过领域本体确定 C(“牡丹”) = {(“牡丹花”,0.5)，(“牡丹电视”,0.2)，(“牡丹江”,0.2)，...}，C(“北京”) = {(“北京市”,0.8)，(“北京牌手表”,0.07)，(“北京故事”,0.05)...}，其中 (“牡丹花”,0.5) 表示“牡丹”的相关关键字“牡丹花”与“牡丹”的匹配值。

[0082] 确定各个相关关键字组合后，获取连接各个关键字的最小子图，例如最小子图集合为：S(graph) = {(g1，牡丹花、北京市，0.65)，(g2，牡丹电视、北京市，0.5)，(g3，牡丹花、李勤勤、北京故事，0.138)...}，容易推算，最小子图 g1 的子图权重为 0.65，g2 的子图权重为 0.5，g3 的子图权重为 0.138。

[0083] 根据每个子图中的关键字及其它节点进行搜索，获得各个子查询结果集，例如，result(g1) = {(doc1，ωg = 0.65，ωr = 0.9)，(doc2，ωg = 0.65，ωr = 0.7)，...}，result(g2) = {(doc3，ωg = 0.5，ωr = 0.8)，(doc4，ωg = 0.5，ωr = 0.6)...}.....，对于查询结果集中的每个文档，wg 表示其对应的最小子图的子图权重，wr 表示该文档与该最小子图的关联程度值，每个子查询结果集中的文档按 wr 排序。

[0084] 从该最小子图对应的子查询结果集中获取的查询结果为与该最小子图关联程度最大的前 a 个查询结果，例如，从 result(g1) 中选择排名为前$\left\lfloor {}^{0.65}\!/_{0.65+0.5+0.135+...}\right\rfloor$的文档加入查询结果集合 RF(q) 中，从 result(g2) 中选择排名为前$\left\lfloor {}^{0.5}\!/_{0.65+0.5+0.135+...}\right\rfloor$的文档加入查询结果集合 RF(q) 中。

[0085] 假设 RF(q) 为 RF(q) = {(doc1,0.65,0.9),(doc2,0.65,0.7),(doc3,0.5,0.8)}，则：

[0086] 可以直接根据查询结果的权重大小，对获取的查询结果进行排序，由于三个文档的权重分别为：s1 = 0.65×0.9，s2 = 0.65×0.7，s3 = 0.5×0.8，所以排序后的查询结果为 RF(q) = {doc1，doc2，doc3}。

[0087] 也可以根据相似程度对获取的查询结果进行排序，此时，假设 similarity(doc 1，doc2) = 0.5，similarity(doc 1，doc3) = 0.1，similarity(doc2，doc3) = 0.2，则排序后的查询结果为：RF(q) = {doc1，doc3，doc2}。

[0088] 本发明实施例还相应提供一种查询结果多样化装置，如图 9 所示，包括：

[0089] 关键字确定单元 901，用于根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集；

[0090] 查询单元 902，用于根据相关关键字组合集中的各个相关关键字组合进行搜索，获得查询结果集；

[0091] 查询结果获取单元 903，用于从查询结果集中获取相应个数的查询结果；

[0092] 排序单元 904，用于对获取的查询结果进行排序，获得多样化查询结果。

[0093] 其中，关键字确定单元 901 具体用于：

[0094] 根据给定查询每个关键字，确定该关键字在领域本体中的相关关键字；

[0095] 根据各个相关关键字，确定相关关键字组合集。

[0096] 关键字确定单元 901 根据各个相关关键字，确定相关关键字组合集，具体包括：

[0097] 确定相关关键字组合集为：$S(Q) = \{(c_1, c_2, \ldots, c_m) | c_1 \in C_1 \&\& c_2 \in C_2 \&\& \ldots c_m \in C_m\}$，其中，$C_i$ 为给定查询中 m 个关键字的第 i 个关键字的相关关键字集合。

[0098] 其中，关键字确定单元 901 还用于：

[0099] 在根据给定查询中的每个关键字，确定该关键字在领域本体中的相关关键字后：

[0100] 在根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集后：

[0101] 对于相关关键字组合集中的每个相关关键字组合，从领域本体抽取连接各个关键字的最小子图，其中，最小子图为实现连接各关键字的领域本体子图中，边数最少的子图；

[0102] 查询单元 902 具体用于：

[0103] 对于每个最小子图，确定由该最小子图中包括的关键字及其它节点构成子查询；

[0104] 根据每个子查询中包括的关键字及其它节点进行搜索，获得与最小子图数量相同的子查询结果集；

[0105] 确定查询结果集为各个子查询结果集构成的集合。

[0106] 查询结果获取单元 903 具体用于：

[0107] 根据每个子查询给定查询的相关程度，从每个子查询结果集中获取相应个数的查询结果；

[0108] 合并从各个子查询结果集中获取的查询结果。

[0109] 进一步，查询结果获取单元 903 具体用于：

[0110] 确定每个最小子图的子图权重为：$w=\sum_{i=1}^{m} r_i \Big/ m\times|E|$，其中 m 为查询关键字的数量，ri 为根据领域本体确定的相关关键字与相应的关键字的匹配值，E 为该子图包括的边的数量；

[0111] 根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果；

[0112] 合并从各个子查询结果集中获取的查询结果。

[0113] 具体的，查询结果获取单元 903 根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果，具体包括：

[0114] 从该最小子图对应的子查询结果集中获取的查询结果为与该最小子图关联程度最大的前 a 个查询结果，a 为不大于当前最小子图的子图权重与所有最小子图的子图权重和的比值的最大整数。

[0115] 排序单元 904 具体用于：

[0116] 对于每个查询结果，确定该查询结果与对应的最小子图的关联程度值；

[0117] 对于每个查询结果，根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重；

[0118] 根据查询结果的权重，对获取的查询结果进行排序，获得多样化查询结果。

[0119] 具体的，排序单元 904 根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重，具体包括：

[0120] 确定该查询结果的权重为该查询结果与对应的最小子图的关联程度值与该最小子图的子图权重的乘积。

[0121] 排序单元 904 根据查询结果的权重，对获取的查询结果进行排序，具体包括：

[0122] 直接按照查询结果的权重大小，对获取的查询结果进行排序；或者

[0123] 确定权重最大的查询结果为排在第一位的查询结果，并确定每两个查询结果之间的相似程度值；对于其它查询结果，确定每个查询结果的相似权重为：$s\times\prod_{d'\in D}(1-similarity(d,d'))$，其中，s 为查询结果的权重，d 为当前查询结果，D 为已排序的查询结果构成的集合，similarity(d，d′) 为 d 和 d’的相似程度值；按照相似权重的大小，对除排在第一位的查询结果外的查询结果进行递归排序。

[0124] 本发明实施例提供一种查询结果多样化方法及装置，通过领域本体确定给定查询的关键字集合的相关关键字组合集，并使用这些相关关键字组合进行查询，避免使用不可靠的查询日志确定子查询关键字，从而使得多样化查询结果更加准确。

[0125] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

[0126] 本发明是参照根据本发明实施例的方法、设备（系统）、和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机程序指令实现流程图和 / 或方框图中的每一

流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0127] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0128] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0129] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0130] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

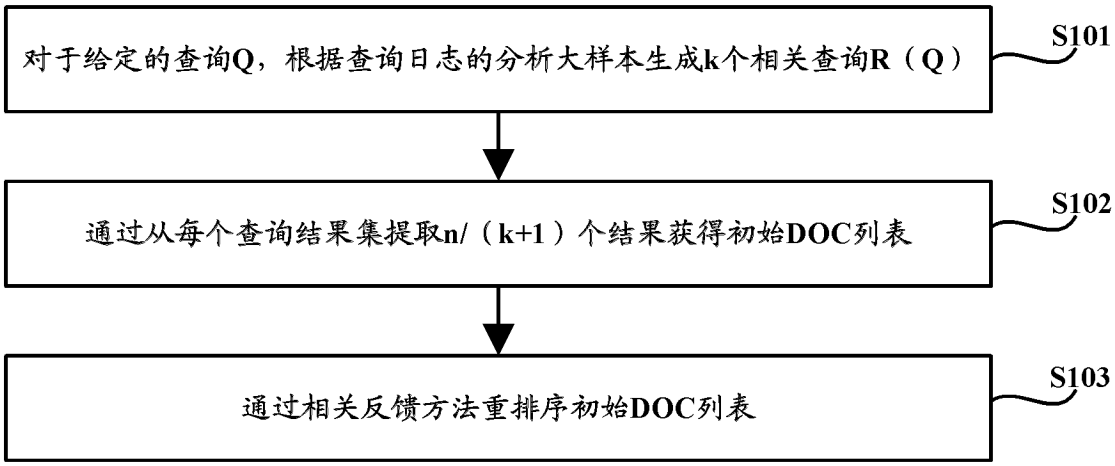

对于给定的查询Q，根据查询日志的分析大样本生成k个相关查询R（Q）
S101
通过从每个查询结果集提取n/（k+1）个结果获得初始DOC列表
S102
通过相关反馈方法重排序初始DOC列表
S103

图 1

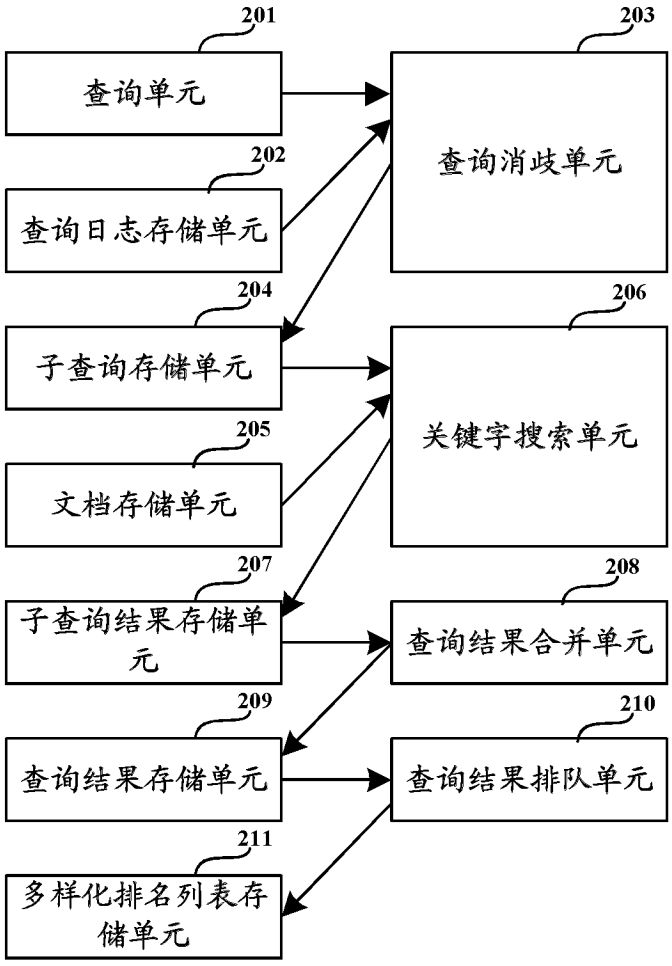

201
查询单元
202
查询日志存储单元
203
查询消歧单元
204
子查询存储单元
205
文档存储单元
206
关键字搜索单元
207
子查询结果存储单元
208
查询结果合并单元
209
查询结果存储单元
210
查询结果排队单元
211
多样化排名列表存储单元

图 2

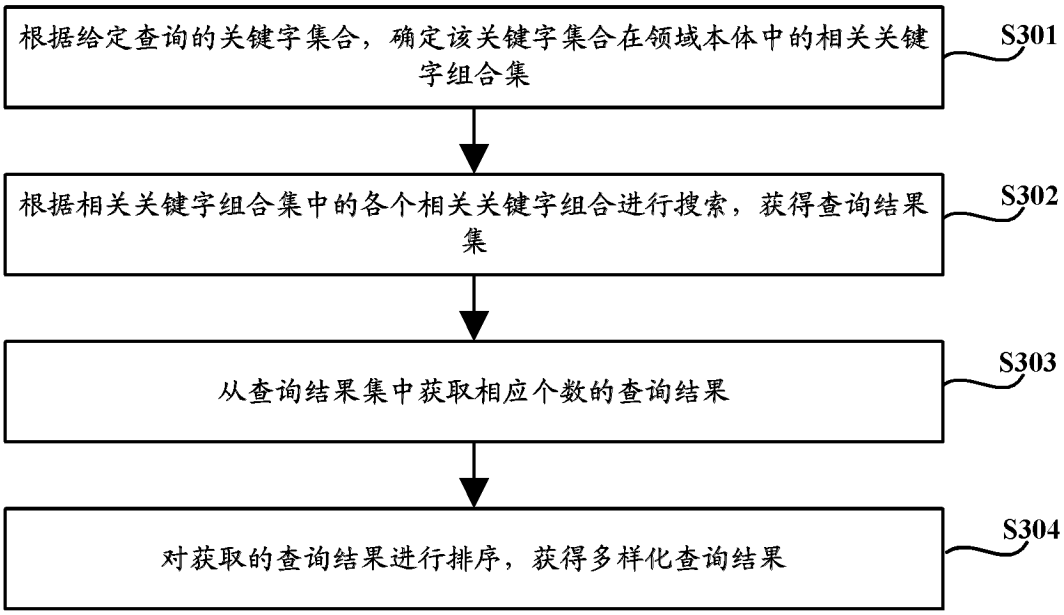

根据给定查询的关键字集合，确定该关键字集合在领域本体中的相关关键字组合集
S301
根据相关关键字组合集中的各个相关关键字组合进行搜索，获得查询结果集
S302
从查询结果集中获取相应个数的查询结果
S303
对获取的查询结果进行排序，获得多样化查询结果
S304

图 3

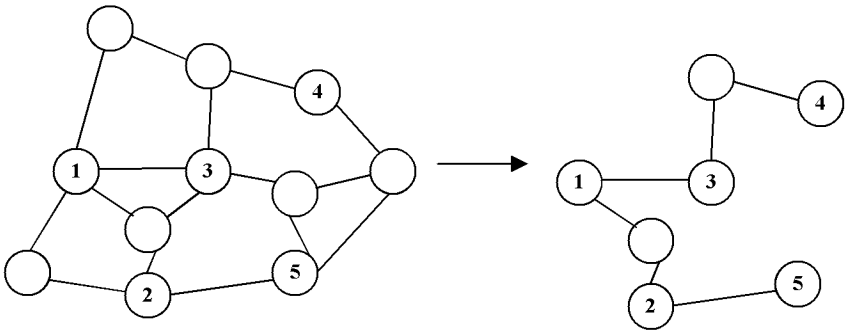

1
2
3
4
5
1
2
3
4
5

图 4

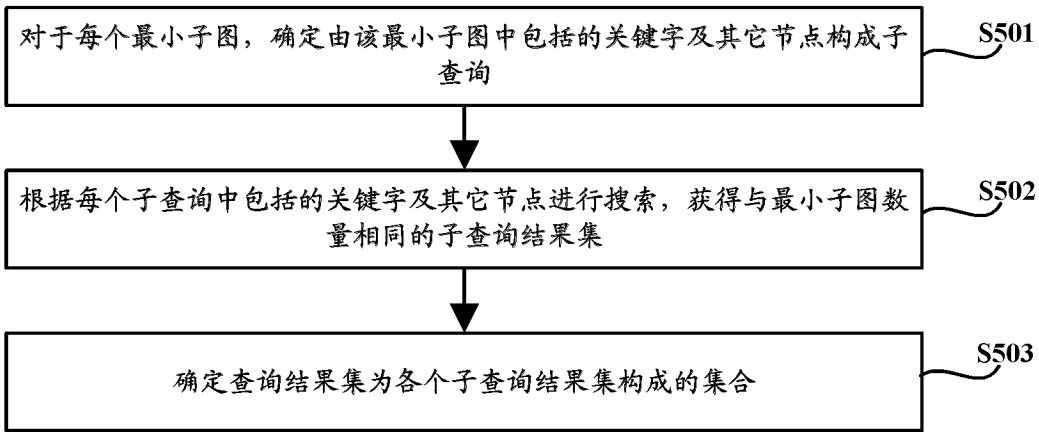

对于每个最小子图，确定由该最小子图中包括的关键字及其它节点构成子查询
S501
根据每个子查询中包括的关键字及其它节点进行搜索，获得与最小子图数量相同的子查询结果集
S502
确定查询结果集为各个子查询结果集构成的集合
S503

图 5

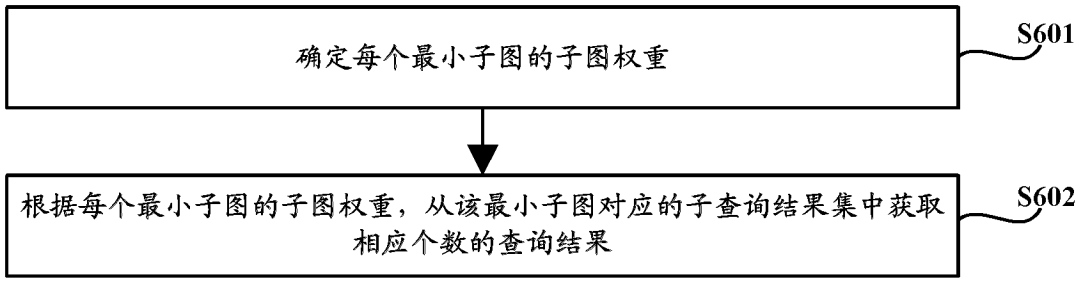

确定每个最小子图的子图权重
S601
根据每个最小子图的子图权重，从该最小子图对应的子查询结果集中获取相应个数的查询结果
S602

图 6

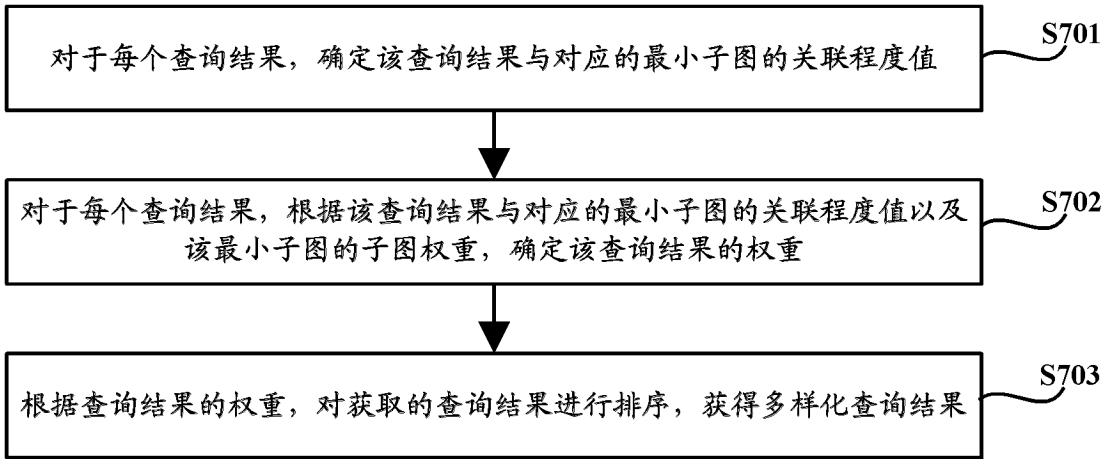

对于每个查询结果，确定该查询结果与对应的最小子图的关联程度值
S701
对于每个查询结果，根据该查询结果与对应的最小子图的关联程度值以及该最小子图的子图权重，确定该查询结果的权重
S702
根据查询结果的权重，对获取的查询结果进行排序，获得多样化查询结果
S703

图 7

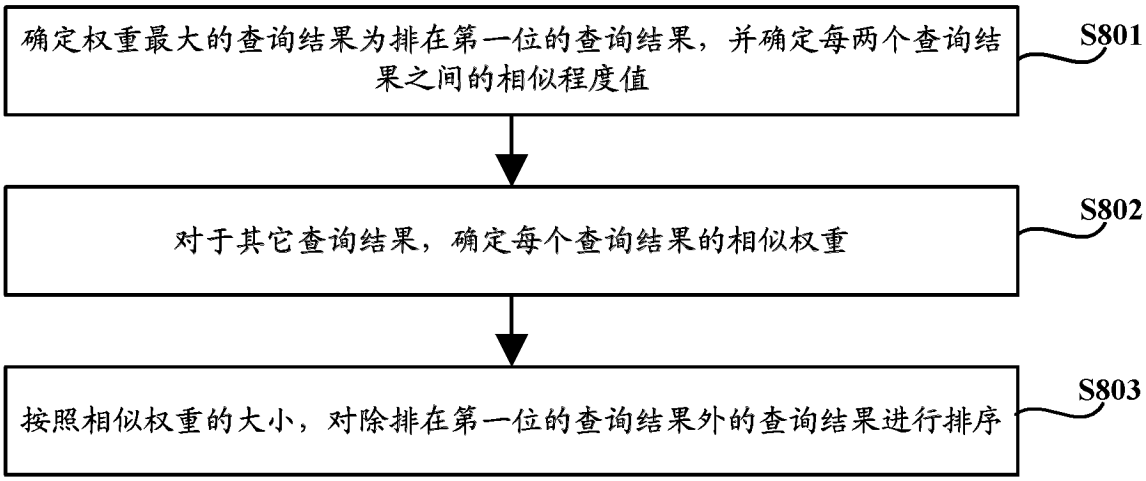

确定权重最大的查询结果为排在第一位的查询结果，并确定每两个查询结果之间的相似程度值
S801
对于其它查询结果，确定每个查询结果的相似权重
S802
按照相似权重的大小，对除排在第一位的查询结果外的查询结果进行排序
S803

图 8

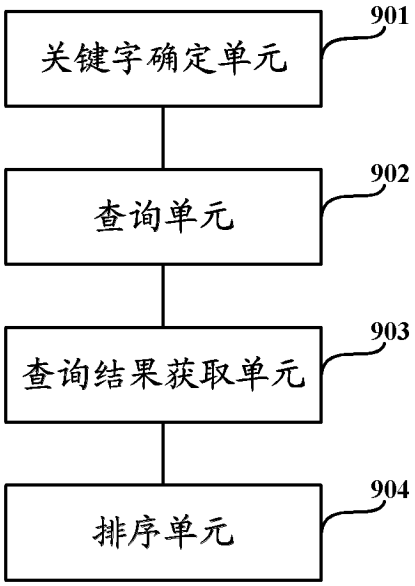


图 9