(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 101576888 B
(45)授权公告日 2011.05.11

(21)申请号 200810095689.5

(22)申请日 2008.05.07

(73)专利权人 香港理工大学
地址 中国香港九龙红磡

(72)发明人 陆永邦

(74)专利代理机构 隆天国际知识产权代理有限公司 72003
代理人 郭晓东

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

(56)对比文件

Christopher C. Yang,et al..《Combination and Boundary Detection Approaches on Chinese Indexing》.《JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE》.2000,340-351.

GERARD SALTON,et al..《TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL》.《Information Processing & Management》.1988,第24卷(第5期),513-523.

Robert W.P. LUK.《Different Retrieval Models and Hybrid Term Indexing》.《Proceedings of the Third NTCIR Workshop》.2003,1-8.

ROBERT W.P. LUK,et al..《A Comparison of Chinese Document Indexing Strategies and Retrieval Models》.《ACM Transactions on Asian Language Information Processing》.2002,第1卷(第3期),225-268.

审查员 马晓宇

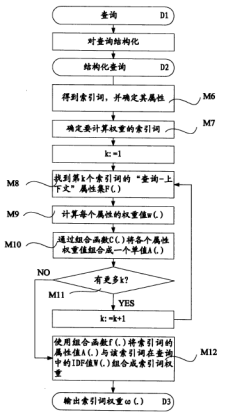
权利要求书 1 页 说明书 8 页 附图 3 页

(54)发明名称

中文信息检索中基于结构约束的索引词权重计算方法

(57)摘要

本发明是有关于一种中文信息检索中基于结构约束的特征权重计算方法,包括以下步骤:a、对查询进行结构化处理,得到结构化查询结果;结构化处理包括:分词、对切分出的词进行词性标注、对查询进行浅层句法分析或对查询进行句法分析中一个或几个;b、根据述结构化查询结果确定索引词,然后根据与所述索引词相邻并位于词列表中的结构化查询的结果,确定所述索引词的查询一上下文属性集;c、计算查询一上下文属性集中每个属性的权重值;d、通过第一组合函数将各个属性的权重值组合成所述索引词的属性值;e、使用第二组合函数对所述索引词的属性值组合,得到所述索引词权重。无论索引词是否在词列表中,本发明的方法都能准确计算出其权重。



CN 101576888 B

1. 一种中文信息检索中基于结构约束的特征权重计算方法，其特征在于，包括以下步骤：

a、对查询进行结构化处理，得到结构化查询结果；

所述结构化处理包括：

分词、对切分出的词进行词性标注或对查询进行句法分析中一个或几个；

b、根据所述结构化查询结果确定索引词，然后根据与所述索引词相邻并位于词列表中的所述结构化查询的结果，确定所述索引词的查询－上下文属性集，其中所述词列表由字典中标题词或简单的子字符串组成；

c、计算所述查询－上下文属性集中每个属性的权重值；

d、通过第一组合函数将所述各个属性的权重值组合成所述索引词的属性值；

f、使用第二组合函数对所述索引词的属性值组合，得到所述索引词权重；

其中，所述索引词的查询－上下文属性集中每个属性都是一个数组，所述数组中的元素为所述结构化查询的结果中的子字符串、子字符串组、子字符串的词性、子字符串的语义特征或子字符串组的句法成分，且所述数组中的元素在所述索引词所在的位置、紧邻所述索引词所在的位置、和所述索引词所在的位置间隔一个字、和所述索引词所在的位置间隔两个字、和所述索引词所在的位置间隔三个字或括入所述索引词所在的位置。

2. 根据权利要求 1 所述的中文信息检索中基于结构约束的特征权重计算方法，其特征在于，其中，所述第一组合函数是 IDF(x) 函数；所述 IDF(x) 函数为模糊合取、模糊析取、模糊否定或是模糊聚集函数，所述模糊聚集函数为扩展布尔合取函数或扩展布尔析取函数。

3. 根据权利要求 1 所述的中文信息检索中基于结构约束的特征权重计算方法，其特征在于，所述第二组合函数将所述索引词的属性值与所述索引词在查询中的逆文本频率指数 IDF 值组合成索引词权重。

4. 根据权利要求 1 所述的中文信息检索中基于结构约束的特征权重计算方法，其特征在于，在进行语音查询和跨语言检索时，步骤 a 和步骤 b 之间进一步包括：使用 N-best 算法列出 N 个结构化查询文本；

所述第二组合函数将同一查询的所述 N 个结构化查询文本的同一个索引词的属性值组合成索引词权重。

5. 根据权利要求 4 所述的中文信息检索中基于结构约束的特征权重计算方法，其特征在于，所述第二组合函数是模糊聚集函数，模糊连接函数或扩展布尔连接函数，其中所述模糊聚集函数为扩展布尔合取函数或扩展布尔析取函数。

中文信息检索中基于结构约束的索引词权重计算方法

技术领域

[0001] 本发明涉及一种中文信息检索技术，特别涉及一种中文信息检索中基于结构约束的索引词权重计算方法。

背景技术

[0002] 由于因特网的普及，大量的信息迅速积累并广泛地被使用。因此，时空距离远近不再是人们存取与使用信息的最大障碍，取而代之的问题是缺乏有效率的方式在浩瀚的因特网海量信息中寻找想要的信息。信息检索技术 (information retrieval technologies) 因为能够提供使用者便捷的方式去存取与使用想要的信息，因此在近几年来格外地受到重视。

[0003] 搜索引擎 (Search Engine) 是基于信息检索技术来实现的，搜索引擎的重要功能就是对文本信息提供检索，中文信息检索技术中至关重要的环节是索引构建，而索引的构建离不开索引词 (index term) 的权重计算。

[0004] 在进行索引词的权重计算之前，需要对中文查询进行结构化处理。请参阅图 1 所示，其为现有技术中对查询进行结构化的流程图。其中，D1 为要进行的查询，例如该查询是一个句子，步骤 M1 对该查询进行分词；步骤 M2 将切分出的词进行词性标注；步骤 M3 对该查询进行浅层句法分析；步骤 M4 对该查询进行进一步的句法分析，最后得到结构化查询 D2，D2 中包含上述结构化处理 M1、M2、M3 及 M4 的结果，例如为句子的各句法成分、切分出的各个词、对各个词的词性标注、各个词的语义特征等。

[0005] 上述处理步骤 M1、M2、M3 及 M4 是形成结构化查询的现有方法。对中文查询进行结构化处理，可选择其中一个或几个步骤，但至少需要其中一个步骤。

[0006] 下面详细描述对查询进行结构化的过程。令 q 为一个中文查询（即图 1 中的 D1）。在步骤 M1 中，q 被分词算法（例如正向最大匹配法，逆向最大匹配法，正向 - 逆向最大匹配法）分为 m 个字符串组，$q = p_1, p_2 \cdots p_i \cdots p_m$，其中这些字符串组是连续的。$p_i$ 是 q 中的子字符串组。令 $p_i = [(q_{i,1}, t_{i,1}), \ldots (q_{i,j}, t_{i,j}) \ldots, (q_{i,n}, t_{i,n}) ; T_i]$，其中 $q_{i,j}$ 是已识别的查询子字符串，其与给定的词列表 W（例如词典）中的一些词条匹配，$t_{i,j}$ 例如为 $q_{i,j}$ 的词性标注（或一些语义特征）（由图 1 中步骤 M2 处理），Ti 是该子字符串组的句法成分（由图 1 中的步骤 M3 和步骤 M4 处理），句法分析所得到的句法成分例如为名词短语或者为介词短语。因此，查询 q 可被处理成 (1) 式格式，图 1 中的 D2 便可为此 (1) 式格式：

[0007] $[(q_{1,1}, t_{1,1}), (q_{1,2}, t_{1,2}), \ldots, (q_{1,n}, t_{1,n}), T_1], \ldots [(q_{m,1}, t_{m,1}), (q_{m,2}, t_{m,2}), \ldots, (q_{m,n}, t_{m,p}), T_m]$ (1)

[0008] 其中 q 是由下列子字符串组成 $q_{1,1}, q_{1,2} \cdots q_{1,n} \cdots q_{m,1}, q_{m,2}, \cdots q_{m,n}$。一个字符串组也可被嵌入另一个字符串组形成一个字符串组的嵌套结构。例如，下面就是嵌套字符串组结构（图 1 中 D2 可能的格式）：

[0009] $[(q_{1,1}, t_{1,1}), [(q_{1,2}, t_{1,2}), (q_{1,n}, t_{1,n}), T_2], T_1]$

[0010] 其中，子字符串组 T2 嵌入子字符串组 T1，这些子字符串组无需连续。这些非连续

组可应用于中文分词中（特别是采用正向最大匹配算法时）。目前的表示法是通过增加一个数字后缀指示一对括号来表示非连续组。例如 $[_1(q_{1,1}, t_{1,1}), [_2(q_{1,2}, t_{1,2}), T_1]_1, (q_{1,n}, t_{1,n})T_2]_2$ 有如下嵌套组 $(q_{1,1}, t_{1,1}), (q_{1,2}, t_{1,2})T_1$ 和 $(q_{1,2}, t_{1,2}), (q_{1,n}, t_{1,n})T_2$。注意在上述表达中，T1 与最近的括号，即“$]_1$”绑定。

[0011] 查询的结构不需要完全被句法结构（由图 1 中的步骤 M4 处理）限制，也不要求查询一定是名词短语、动词短语等。查询可以是由浅层句法分析（即图 1 中的步骤 M3）识别出来的一些句法成分。这些子字符串组也可以是基于语义特征的，例如表 1 中 L1 的语义特征是 +Loc（位置），语义特征例如可以通过语义知识库“知网”(HowNet) 进行语义分析来获得。

[0012]

<table>
<tr><td rowspan="6">结构约束</td><td rowspan="4">子字符串组的句法成分</td><td colspan="8">NP_3</td></tr>
<tr><td colspan="7">PP_1</td><td rowspan="3"></td></tr>
<tr><td colspan="5">NP_2 (+Loc)</td><td rowspan="2"></td></tr>
<tr><td colspan="3">NP_1 (+Loc)</td></tr>
<tr><td>子字符串（词）
子字符串的词性
子字符串的语义特征</td><td>L1,
名词
+Loc</td><td>L2,
名词
+Loc</td><td>L3,
名词
+Loc</td><td>L4,
介词
None</td><td>L5,
名词
+Ani</td></tr>
<tr><td colspan="2">字</td><td>香</td><td>港</td><td>理</td><td>工</td><td>大</td><td>学</td><td>之</td><td>友</td></tr>
<tr><td colspan="2">位置</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr>
<tr><td rowspan="2">重叠二元索引词</td><td>在词列表 W 中</td><td colspan="2">香港</td><td colspan="2">理工</td><td colspan="2">大学</td><td>之</td><td>友</td></tr>
<tr><td>不在词列表 W 中</td><td></td><td colspan="2">港理</td><td colspan="2">工大</td><td colspan="2">学之</td><td></td></tr>
<tr><td>非连续二元索引词</td><td>在词列表 W 中</td><td></td><td></td><td colspan="3">理……大</td><td colspan="3">学……友</td></tr>
</table>

[0013] 表 1

[0014] 一个查询示例 q_1 = 香港理工大学之友，对其进行结构化处理，形成结构化查询，结构化查询的结果如表 1 所示，该结构化处理可以用已有的算法来实现，该查询可利用正向或者逆向最大匹配法以及这两种方法的组合来进行分词。词性标注可以通过隐式马尔科夫模型 (hidden Markov model) 或者错误驱动转换 (error-driven transformation) 方法来确定。实词可以通过分类法识别。句法分析可以通过 CYK 分析法，图表分析法等实现，在表 1 中，句法成分 NP 表示名词短语，句法成分 PP 表示介词短语；句法分析可以推广至分析属

性语法，其中一些属性是基于语义的。在表 1 中，语义特征例如为 +Loc（位置）。

[0015] 这里，有两种类型的索引词，即重叠二元词 (overlapping characterbigrams) 和非连续二元词 (non-contiguous bigrams)。在本查询示例 q_1 中，重叠二元词即“香港”、“港理”、“理工”、“工大”、“大学”、“学之”和“之友”，其中，在词列表 W 中有的索引词例如为“香港”、“理工”，不在词列表 W 中的索引词例如为“港理”、“工大”；而非连续二元词例如为“理……大”、“学……友”，即非连续二元索引词是三个字的子字符串中的第一个字和最后一个字。

[0016] 接下来，就可以将该结构化查询所得到的各字符串组和字符串做为索引词来进行权重计算。索引词的权重计算需要依赖词列表，因此词列表的规模对于权重计算的准确程度有着很大的制约作用。中文中，新词会频繁的出现，词列表也需要频繁的更新，词列表更新后，在先使用的旧列表就过期了，索引词的权重便需要重新计算，索引也需要重新构建，而搜索引擎的这种频繁更新是很难实现的，权重计算的准确程度也就无法得到保证。

[0017] 在这种情况下，一些不在词列表中的索引词的权重计算便显得尤为重要，现有技术中对这种索引词权重的计算有如下方法：如果是基于词的索引，其权重通过单个字的权重得到，例如“港理”这个词，通过单字“港”和“理”来计算其权重，完全没有考虑该词和上下文，即“香港”和“理工”的关系，因此这种权重的计算结果是不准确的。如果不是基于词的索引，通常会通过 n-gram 这种统计方法来分词并计算其权重，计算使用这种方法分出的索引词的权重时，并不会考虑该索引词是不是词列表中某个词的一部分，也不会考虑该索引词是不是在词汇表中某两个词的边界，或者该索引词本身就是一个词，这样计算出来的权重同样是不准确的。

[0018] 有鉴于上述现有技术存在的缺陷，本发明人提出一种基于结构约束的索引词权重计算方法，其能够改进现有技术的权重计算方法，使索引词可以得到更准确的权重。

发明内容

[0019] 本发明的主要目的在于，提供一种中文信息检索中基于结构约束的词权重计算方法，所要解决的技术问题是无论该索引词是否在词列表中，都能准确计算出其权重，从而实现词列表升级而无需完全重新计算索引词的权重。

[0020] 本发明的目的及解决其技术问题是采用以下技术方案来实现的。依据本发明提出的一种中文信息检索中基于结构约束的特征权重计算方法，包括以下步骤：a、对查询进行结构化处理，得到结构化查询结果；所述结构化处理包括：分词、对切分出的词进行词性标注、对查询进行浅层句法分析或对查询进行句法分析中一个或几个；b、根据所述结构化查询结果确定索引词，然后根据与所述索引词相邻并位于词列表中的所述结构化查询的结果，确定所述索引词的查询－上下文属性集；c、计算所述查询－上下文属性集中每个属性的权重值；d、通过第一组合函数将所述各个属性的权重值组合成所述索引词的属性值；f、使用第二组合函数对所述索引词的属性值组合，得到所述索引词权重。

[0021] 本发明的目的及解决其技术问题还可采用以下技术措施进一步实现。

[0022] 前述的中文信息检索中基于结构约束的特征权重计算方法，所述索引词的查询－上下文属性集中每个属性都是一个数组，所述数组中的元素为所述结构化查询的结果中的子字符串、子字符串组、子字符串的词性、子字符串的语义特征或子字符串组的句法成

分，且所述数组中的元素在所述索引词所在的位置、紧邻所述索引词所在的位置、和所述索引词所在的位置间隔一个字、和所述索引词所在的位置间隔两个字、和所述索引词所在的位置间隔三个字或括入所述索引词所在的位置。

[0023] 前述的中文信息检索中基于结构约束的特征权重计算方法，其中，所述第一组合函数是 IDF(x) 函数；所述 IDF(x) 函数为模糊合取、模糊析取、模糊否定或是模糊聚集函数，所述模糊聚集为扩展布尔合取或扩展布尔析取。

[0024] 前述的中文信息检索中基于结构约束的特征权重计算方法，所述第二组合函数将所述索引词的属性值与所述索引词在查询中的 IDF 值组合成索引词权重。

[0025] 前述的中文信息检索中基于结构约束的特征权重计算方法，在进行语音查询和跨语言检索时，步骤 a 和步骤 b 之间进一步包括：使用 N-best 算法列出 N 个结构化查询文本。

[0026] 前述的中文信息检索中基于结构约束的特征权重计算方法，所述第二组合函数将同一查询的所述 N 个结构化查询文本的同一个索引词的属性值组合成索引词权重。

[0027] 前述的中文信息检索中基于结构约束的特征权重计算方法，所述第二组合函数是模糊聚集，模糊连接或扩展布尔连接。

[0028] 由上述技术方案可知，本发明具有以下有益效果：

[0029] 1、本发明利用第一组合函数将不在词列表中的索引词的“查询 - 上下文”属性集中属性的权重组合成属性值，从而可以为词列表中没有的索引词通过上下文关系更为准确的赋权重值，进而实现搜索引擎的词列表升级而不影响原有索引词的权重，其对于中文搜索引擎的动态调整适应，动态更新词列表是很重要的。

[0030] 2、本发明提出了用位置区间来对属性进行处理，该处理方法支持非连接的索引词，跨越索引词边界，在索引词内和邻近索引词的词，因此本发明能为检索中的非连接的查询索引词赋予权重。

[0031] 3、本发明使用第二组合函数（例如，模糊聚集，模糊连接或扩展布尔连接）来组合同一查询（例如语音查询和跨语言查询）的不同结构化查询文本中同一索引词的权重，形成该索引词的权重，因此在进行语音查询和跨语言检索时，也可较准确的计算其中未在词列表中出现的索引词和非连接索引词的权重。

[0032] 通过以下参照附图对优选实施例的说明，本发明的上述以及其它目的、特征和优点将更加明显。

附图说明

[0033] 图 1 为现有技术中对查询进行结构化的流程图。

[0034] 图 2 为本发明基于结构约束的索引词权重值计算方法流程图。

[0035] 图 3 为本发明对同一查询的不同结构化查询文本中同一索引词权重的计算方法流程图。

具体实施方式

[0036] 下面将详细描述本发明的具体实施例。应当注意，这里描述的实施例只用于举例说明，并不用于限制本发明。

[0037] 请参阅图 2 所示，其为本发明基于结构约束的词权重计算方法的流程图。对于一

个查询 D1，可以按照图 1 所示的步骤对其进行结构化，形成结构化查询 D2。令 q 为一个中文查询，其被结构化为 (1) 式，如果没有字符串组，则 (1) 式可简化为：

[0038] $[(q_{1,1}, t_{1,1})T_1], \ldots\ldots, [(q_{m,1}, t_{m,1}), T_m] \ldots\ldots$ (2)

[0039] 其中若没有类型标识符，即没有 T_i，则 (2) 式进一步可简化为：

[0040] $(q_{1,1}, t_{1,1}), \ldots\ldots, (q_{m,1}, t_{m,1}) \ldots\ldots$ (3)

[0041] 若没有词性标注，即没有 $t_{i,j}$，则在这种特殊情况下，结构化查询可被简化为：

[0042] $q_{1,1}, \ldots\ldots q_{m,1} \ldots\ldots$ (4)

[0043] 本发明可以使用上述各式的算法对查询进行结构化，形成结构化查询，如步骤 D2 所示。本发明的词权重计算方法可以应用到非连接 n-gram 索引词和连接的 n-gram 索引词。

[0044] 子字符串 $q_{i,j}$ 为单词列表 W 中条目。这个列表可以是某些字典中标题词，也可以是某些简单的子字符串（例如常规表达）。诸如，时间词“10 月”就是通过识别一个或多个数字，之后识别时间词（即月）而识别出来的。此外，列表中也包括从一些文档资料中选取出来的子字符串。诸如两个字符的子字符串可以从基于点点交互方式的信息中选取。其也可以推广到选取 n 个特征字 ($n > 2$)。

[0045] 假设查询 q_1 ＝香港理工大学之友，结构化为非连接组 $(_1$ 香港 $(_2$ 理工 $)_1$ 大学 $)_2(_3$ 之友 $)_3$，该算法生成的子字符串组和许多自然语言的句法分析生成的子字符串组是不同的，因为该算法没有考虑上下文的语法。这些非连接子字符串组按照第一个字出现的顺序排列。在本例中，q_1 ＝ $(_1$ 香港 $(_2$ 理工 $)_1$ 大学 $)_2(_3$ 之友 $)_3$，其中，第一个子字符串是香港理工，第二个是理工大学，最后一个是之友。

[0046] 在本实施例中，搜索引擎使用 n-gram 算法进行分词。但本发明并不限定于此，本发明可以使用至其他算法的词索引，或一些复合词的词索引等（例如，组合字，组合词和二元词）。

[0047] 步骤 M6 通过查询的区间得到索引词，并（例如以深度优先的方式）得到这些索引词的属性。例如，在表 1 中，NP_2 的表示的位置区间是 1 到 6，所以 NP_2 这个索引词使用两个位置（即 1 和 6），为查找在某个给定位置的重叠的各个位置区间的索引词，可以使用线段树 (segment tree)、区间空指令表 (interval skip list) 和 R-tree。这样的索引结构也支持快速查询处理。

[0048] 步骤 M7 用来确定位置 p 第 k 个索引词 $t_{k,p}$。

[0049] 每个索引词都有一系列的“查询 - 上下文”属性集，该“查询 - 上下文”属性集已经由步骤 M6 得到了，步骤 M8 的函数 F(.) 执行的就是找到索引词 $t_{k,p}$ 的“查询 - 上下文”属性集，即 $F(q, p, t_{k,p}) = \{a_j\}_j$。如表 1 中位置 2 的不在词列表 W 中的第一个索引词“港理”的“查询 - 上下文”属性集为：

[0050] $F(q_1, 2, t_{1,2}$ ＝港理 $) = \{a_1 = ($word_boundary, L1, 名词, L2, 名词 $)$,

[0051] $a_2 = ($word_group_boundary, L1, 名词, $NP_1)$,

[0052] $a_3 = ($in_word_group, $NP_2)$,

[0053] $a_4 = ($in_word_group, $NP_3)\}$

[0054] “查询 - 上下文”属性集中每个属性都是一个数组，数组中又包含若干个元素，这些元素为结构化查询的结果，例如是子字符串（子字符串例如为一个词）、子字符串组、子字符串的词性、子字符串的语义特征或子字符串组的句法成分，且每个属性数组的元素都是

与索引词 $t_{k,p}$ 相邻且在词列表中的元素，这里所说的相邻可以为该元素在该索引词 $t_{k,p}$ 所在位置 p、该元素紧邻该索引词 $t_{k,p}$ 所在位置 p（例如，词末在 p-1）、该元素接近该索引词 $t_{k,p}$ 所在位置 p（例如，对于接近的三个词，词末分别在位置 p-1，p-2 或 p-3）或该元素括入该索引词 $t_{k,p}$ 所在位置 p（例如在位置 p-1 的四字词）。为表述更为清晰，本实施例中没有考虑语义特征。否则，由于它的所有成分都具有相同的语义特征 +LOC，进而我们要增加特征来计算权重。

[0055] 再如，在表 1 中的非连接二元词“理……大”所占的位置是 3 和 5，可以将这个非连接的 n-gram 索引词看成从位置 3 到 5 的一个字符串，其相关的特征可以根据 3 到 5 的位置区间来确定。

[0056] 有些在词列表中的索引词，则可以在属性集中加入另外一些特征，另外加入的特征将会给索引词增加更多的权重。

[0057] 步骤 M9 通过函数 w(.) 来计算每个“查询 - 上下文”属性的权重值，这里我们假设各属性的权重值为：

[0058] w(word_boundary，L1，名词，L2，名词）= 0.5

[0059] $w(word\_group\_boundary, L1, 名词, NP_1) = 0.5 \times 1/|NP_1|_1 \times IDF(NP_1)$

[0060] $w(in\_word\_group, NP_2) = 1/[|NP_1|_1+2] \times IDF(NP_1)$

[0061] $w(in\_word\_group, NP_3) = 1/[|NP_3|_1+2] \times IDF(NP_3)$

[0062] 其中，$|.|_1$ 是该字符串组的 city-block 长度，IDF(x) 是字符串组 x 中所有索引词的逆文本频率指数的平均值。通常，IDF(x) 可以视为 P- 范数 (P-norm) 扩展布尔合取值 (Extended Boolean conjunction)：

[0063] $$IDF(x) = 1 - \sqrt[P]{\frac{1}{\#(x)} \sum_{t \in x} (1 - IDF(t))^P}$$

[0064] 其中 t 是索引词，x 是包含多个索引词的字符串组，#(x) 是 x 中索引词的数量（包括重复的），IDF(t) 是索引词 t 的 IDF 值；

[0065] 或者，IDF(x) 是 P- 范数 (P-norm) 扩展布尔析取值 (Extended Booleandisjunction)：

[0066] $$IDF(x) = \sqrt[P]{\frac{1}{\#(x)} \sum_{t \in x} IDF(t)^P}$$

[0067] 其中词 IDF(t) 可能为：

[0068] $$IDF(t) = \log\frac{N}{df(t)}, \log\frac{N+0.5}{df(t)+0.5}, \log\frac{N-df(t)+0.5}{df(t)+0.5}, \log\left[\frac{N+0.5}{df(t)+0.5}+1\right],$$

[0069] 或 $$0.5 + \log\left[\frac{N+0.5}{df(t)+0.5}+1\right]$$

[0070] 其中 N 是检索中的总文档数，df(t) 是出现索引词 t 的文档频率。

[0071] 因此，通过函数 $w(a_j)$ 计算，所有的属性都成为了数字值。

[0072] 步骤 M10 通过第一组合函数 C(.) 将各个属性权重值组合成一个单值，即属性值 A(.)：

[0073] $A(q,p,t_{k,p}) = C(\{w(a_j)\}_j)$.

[0074] 如果对于所有的 j，$w(a_j)_j$ 的值都在 [0,1] 区间，则 C(.) 可能是模糊连接 (fuzzy connective) 的某种组合（诸如：模糊合取 (fuzzy conjuction)，模糊析取 (fuzzy disjunction) 和模糊否定 (fuzzy negation)）或者是模糊聚集 (fuzzyaggregation)（例如，扩展布尔合取或者扩展布尔析取），若 $w(a_j)_j$ 的值不在 [0,1] 区间，也可使用某种函数将其映射到 [0,1] 区间，这些都是可以通过现有技术实现的。对于我们在表 1 的例子，假定使用 P- 范数扩展布尔析取，则“港理”的属件值为：

[0075]

$$A(q_1, 2, \text{港理}) = \sqrt[P]{\frac{1}{4}\left[0.5^P + \left(\frac{0.5 \times IDF(NP_1)}{|NP_1|_1}\right)^P + \left(\frac{IDF(NP_1)}{|NP_1|_1 + 2}\right)^P + \left(\frac{IDF(NP_3)}{|NP_3|_1 + 2}\right)^P\right]}$$

[0076] 接下来，判断同一位置是否还有索引词（即步骤 M11），如果有，则计算下一个索引词 $t_{k+1,p}$ 的属性值 $A(q,p,t_{k+1,p})$，直到没有更多的索引词。

[0077] 步骤 M12 使用组合函数 f(.) 将 A(.) 的值与该索引词在查询中的 IDF（或其变式）值 W(.) 组合成索引词权重 ω(.)，即：$\omega(q,t_{k,p}) = f(A(q,p,t_{k.p}), W(t_{k,p}))$。

[0078] 例如，函数 f(x,y) 的可能实现为：

[0079] $1-\sqrt[p]{(1-x)^P+(1-y)^P}$ 或 $\alpha.x+(1-\alpha).y$，

[0080] 其中 P 例如为一个规定的参数，α 例如是一个在 [0,1] 区间的参数，步骤 D3 将索引词权重 ω(.) 输出。这样，通过上述步骤，可以得到在位置 p 的查询索引词的权重。

[0081] 本发明可以延伸至基于音节的索引构建，以用来进行基于语音查询的检索，由语音查询产生的不同的文本查询的各音节的权重组合成音节的权重。

[0082] 请参阅图 3 所示，其为本发明对同一查询的不同结构化查询文本中同一索引词权重的计算方法流程图。对于一个语音查询 s，可能产生一系列的候选文本查询，对候选文本查询结构化后，形成如图 2 中步骤 D2 所示的结构化查询；接下来执行步骤 M5，使用 N-best 算法列出 N 个最有可能的结构化查询文本 q_h(h = 1……N) 例如对于 s_1，产生两个文本查询：

[0083] q_1 = 香港理工大学之友；和 q_2 = 香港理工大学只有；

[0084] 其中 s_1 =“xiang gang li gong da xue zhi you”（这里使用的是汉语识别）；接下来执行步骤 M6，得到索引词，并确定其属性；假设索引词是重叠二元词，可以结合两个查询文本中的二元索引词的权重，例如，索引词“港理”均出现在 q_1 和 q_2 中，则可以组合索引词“港理”在 q_1 和 q_2 中的属性值计算出该索引词的权重。

[0085] 步骤 M13 确定第 h 个结构化查询，步骤 M14（即图 2 中步骤 M7-M11）得到位置 p 的各索引词的属性值 $A(q,p,t_{k,p})$，并继续计算位置 p+1 的各索引词的属性值 $A(q,p+1,t_{k,p+1})$，将所有位置的各索引词的属性值都计算出来后，开始确定第 h+1 个结构化查询，并计算相应的各索引词的属性值 $A(q,p,t_{k,p})$。

[0086] 步骤 15 使用组合函数 c(.) 组合不同结构化查询的索引词的属性值 A(.) 组合成索引词权重，例如均出现在 q_1 和 q_2 中的索引词“港理”，其权重为：

[0087] $\omega(s_1, \text{港理}) = c(A(q_1, 2, \text{港理}), A(q_2, 2, \text{港理}))$ (5)

[0088] 其中，c(.) 为组合函数，例如为模糊聚集，模糊连接、扩展布尔连接或 p- 范数扩展布尔析取。虽然这里以两个查询（即 h = 2）为例，事实上，本发明提出的方法可以用至任

意多个查询文本。

[0089] 本发明提出的权重计算方法也可以用于跨语言检索。例如，英文查询 e_1 是“Friends of Hong Kong Polytechnic university”，使用翻译软件翻译该查询可能翻译成查询文本：

[0090] q_1 ＝香港理工大学之友；和 q_3 ＝香港理工大学的朋友；

[0091] 索引词“港理”均出现在 q_1 和 q_3 中，根据图 3 所示的权重计算方法，在这两个查询中，索引词“港理”的权重为：

[0092] $\omega(e_1,港理) = c(A(q_1,2,港理),A(q_3,2,港理))$ (6)

[0093] 此公式 (6) 和公式 (5) 相同，但 q_3 中的结构约束和 q_2 中的结构约束是不同的，因此，查询 s_1 中索引词“港理”的属性权重可能与 e_1 中索引词“港理”的属性权重不同。

[0094] 当倒排索引没有位置信息来支持 (5)、(6) 式这种接近性查询时，本方法计算不同位置的同一个索引词的权重 $\omega(.)$ 有两个变量；而有位置信息来支持 (5)、(6) 式这种接近性查询时，词权重 $\omega(.)$ 有三个变量，即 $\omega(q,p,t_k)$，则 (6) 式中不同查询文本中同一个位置的同一个索引词的权重为：

[0095] $\omega(e_1,2,港理) = c(A(q_1,2,港理),A(q_3,2,港理))$。

[0096] 对于每个查询，其被发送至不同的搜索引擎时（例如在元搜索中或全文搜索或跨库搜索中），会得到来自不同搜索引擎的查询索引词权重，然后，来自不同搜索引擎的排序列表组合形成最后的排序列表。

[0097] 本发明的方法可以准确地为非连接的 n-gram 索引词、连接的 n-gram 索引词和在词列表中的词汇赋权重，这些索引词可以是字符串组、字符串、词。本发明的方法也可被用于具有结构约束的英文查询。

[0098] 虽然已参照几个典型实施例描述了本发明，但应当理解，所用的术语是说明和示例性、而非限制性的术语。由于本发明能够以多种形式具体实施而不脱离发明的精神或实质，所以应当理解，上述实施例不限于任何前述的细节，而应在随附权利要求所限定的精神和范围内广泛地解释，因此落入权利要求或其等效范围内的全部变化和改型都应为随附权利要求所涵盖。

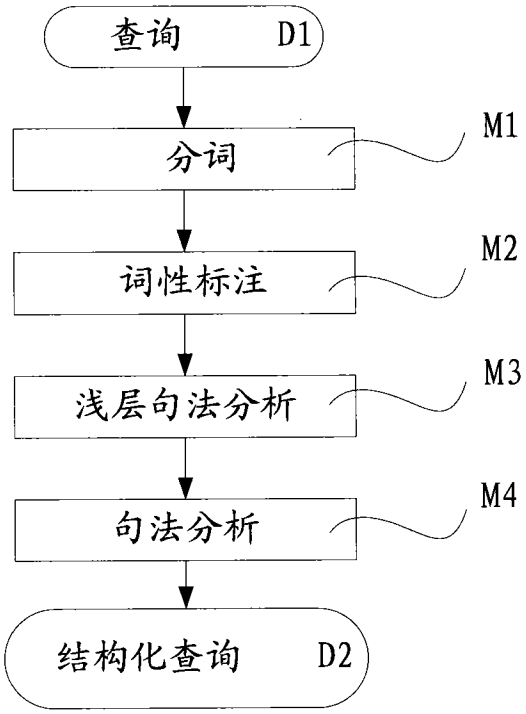


图 1

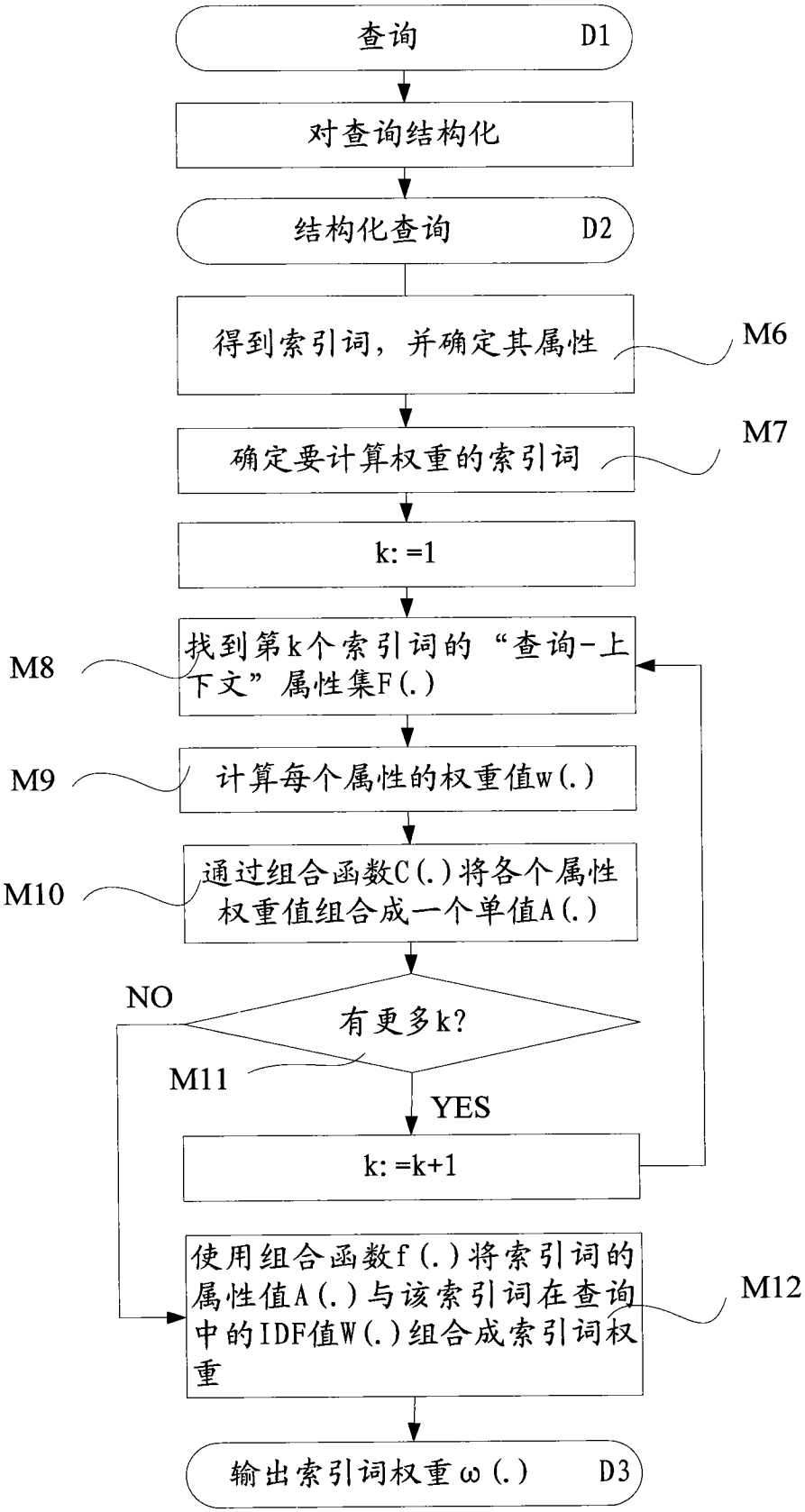

查询 D1
对查询结构化
结构化查询 D2
得到索引词，并确定其属性
M6
确定要计算权重的索引词
M7
k: =1
找到第k个索引词的“查询-上下文”属性集F(.)
M8
计算每个属性的权重值w(.)
M9
通过组合函数C(.)将各个属性权重值组合成一个单值A(.)
M10
有更多k?
NO
M11
YES
k: =k+1
使用组合函数f(.)将索引词的属性值A(.)与该索引词在查询中的IDF值W(.)组合成索引词权重
M12
输出索引词权重ω(.) D3

图 2

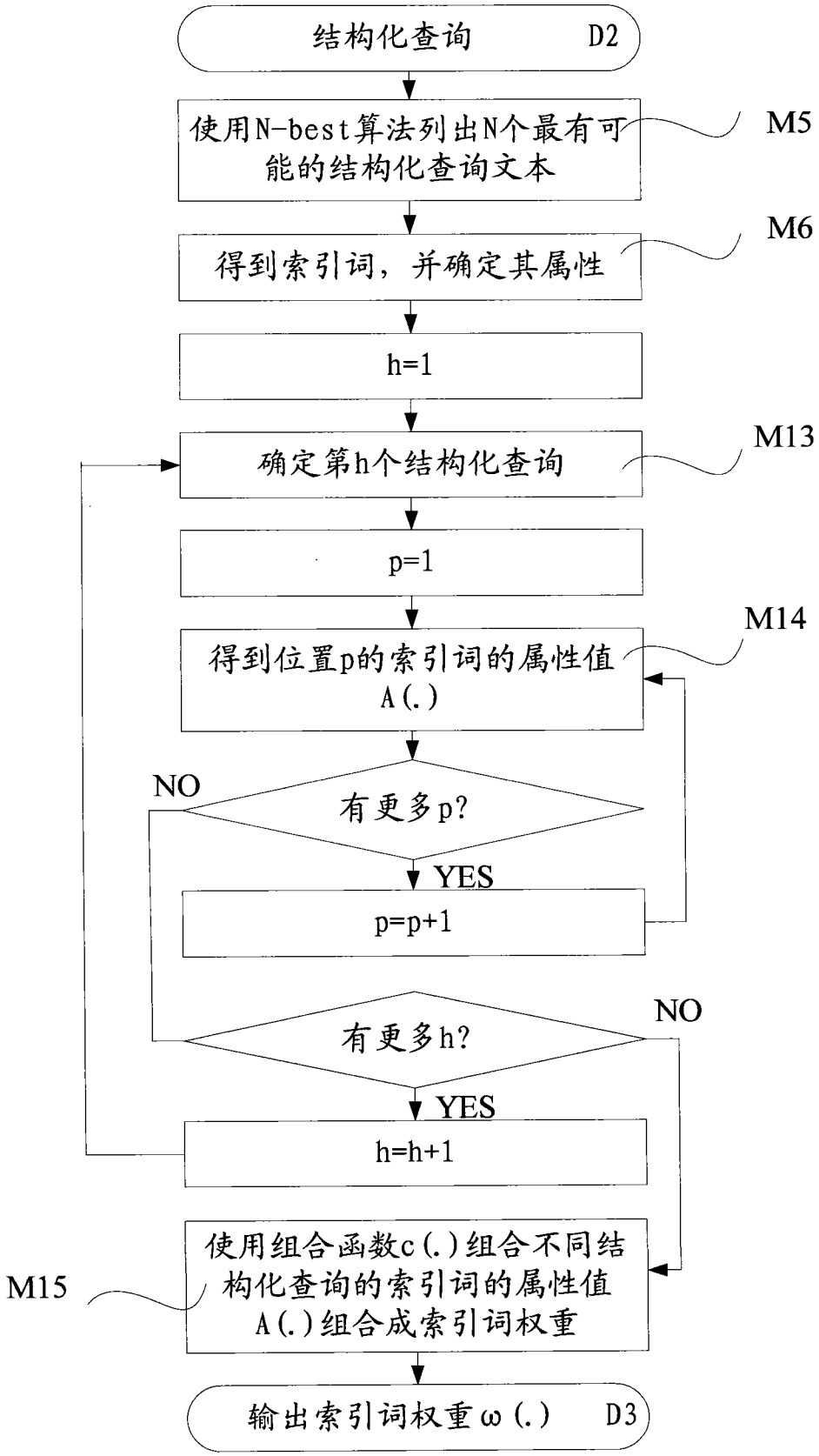

结构化查询 D2
使用N-best算法列出N个最有可能的结构化查询文本
M5
得到索引词，并确定其属性
M6
h=1
确定第h个结构化查询
M13
p=1
得到位置p的索引词的属性值A(.)
M14
NO
有更多p?
YES
p=p+1
NO
有更多h?
YES
h=h+1
使用组合函数c(.)组合不同结构化查询的索引词的属性值A(.)组合成索引词权重
M15
输出索引词权重ω(.) D3

图 3