(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 102859523 A
(43)申请公布日 2013.01.02

(21)申请号 201180018778.X

(22)申请日 2011.03.31

(30)优先权数据
12/760128 2010.04.14 US

(85)PCT申请进入国家阶段日
2012.10.12

(86)PCT申请的申请数据
PCT/US2011/030621 2011.03.31

(87)PCT申请的公布数据
WO2011/130008 EN 2011.10.20

(71)申请人 微软公司
地址 美国华盛顿州

(72)发明人 陆建平 张东晖 H.S.K. 万

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 李静岚 汪扬

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

权利要求书 2 页 说明书 12 页 附图 6 页

(54)发明名称

利用子查询自动生成查询建议

(57)摘要

可以通过识别所希望的子查询生成查询建议。可以累积搜索引擎数据来为各种查询确定使用特征。可以根据使用数据生成和排名潜在子查询。在对潜在子查询进行排名之后，当接收到搜索请求时，可以使用排名来选择子查询。可以将所选的子查询直接用作查询建议，或可以将子查询用作另一个查询建议引擎的输入。

原始查询	子查询	排名值
免费下载歌	下载歌	22120000
免费下载歌	费下	20157000
免费下载歌	费下载	17554000
免费下载歌	免费下	10398000

CN 102859523 A

1. 一种生成查询建议的方法，其包括：

获取查询日志文件；

识别含有至少 4 个查询元素的包含在查询日志文件中的查询；

为每个识别的查询确定子查询；

将确定的子查询与查询日志文件中的查询匹配；

为每个匹配的子查询计算排名，该排名基于不同用户的数量、页面视图数据、子查询中的查询元素的数量和子查询的母查询的数量；

接收搜索查询；

为接收的搜索查询确定搜索子查询，所述搜索子查询的至少一个对应于具有所计算排名的匹配子查询；

根据所选一个或多个搜索子查询的相应计算排名选择一个或多个搜索子查询；以及

根据所选一个或多个搜索子查询提供一个或多个建议查询。

2. 如权利要求 1 所述的方法，其中为每个识别的查询确定子查询包括为每个识别的查询确定 n 元组。

3. 如上面权利要求的任何一项所述的方法，其中为每个识别的查询确定子查询包括为每个识别的查询确定位置无关子查询，每个位置无关子查询含有比相应所识别的查询更少的查询元素。

4. 如上面权利要求的任何一项所述的方法，其中所述识别包含在查询日志文件中的查询，所述为每个识别的查询确定子查询，所述匹配确定的子查询，以及所述为每个匹配的子查询计算排名都是自动进行的。

5. 如上面权利要求的任何一项所述的方法，进一步包括过滤查询日志文件以排除一个或多个查询，其中识别包含在查询日志文件中的查询包括根据过滤的查询识别查询。

6. 如上面权利要求的任何一项所述的方法，其中为每个识别的查询确定子查询包括确定具有查询项的阈数或更少的子查询。

7. 如上面权利要求的任何一项所述的方法，其中根据所选一个或多个搜索子查询提供一个或多个建议查询包括将所选的一个或多个搜索子查询用作查询建议引擎的输入并提供所述查询建议引擎生成的至少一个查询。

8. 如上面权利要求的任何一项所述的方法，其中所识别的查询包含 4 到大约 60 个查询元素。

9. 如上面权利要求的任何一项所述的方法，其中查询元素对应于基于字的书面语言的字。

10. 如权利要求 9 所述的方法，其中基于字的书面语言是汉语、日语或朝鲜语。

11. 如上面权利要求的任何一项所述的方法，其中为每个匹配的子查询计算排名包括：

计算每个子查询的母查询的数量；

根据不同用户的数量和页面视图信息为每个子查询计算频率；

根据子查询中的查询元素的数量、母查询中的查询元素的数量、查询日志文件中的查询的数量、以及子查询的母查询的数量为每个子查询计算一个或多个归一化加权频率值，其中为子查询计算的归一化加权频率值的数量对应于子查询的母查询的数量；以及

根据子查询的一个或多个归一化加权频率值和子查询的母查询的数量为所述子查询计算平均归一化加权频率值。

12. 如权利要求 11 所述的方法，其中平均归一化加权频率值进一步基于子查询中的查询项的数量。

13. 如权利要求 11 或 12 所述的方法，其中该方法进一步包括将生成的排名清单提供给查询建议引擎。

利用子查询自动生成查询建议

背景技术

[0001] 像可在网络上获得的文档那样的庞大文档集合的关键词或查询搜索现在是常见的活动。随着搜索引擎越来越唾手可得，使用搜索技术的用户的数量增加了，并且这些用户搜索越来越广的主题。因此，许多用户在用户不熟悉的主题领域中进行许多搜索。这可能导致用户难以构想搜索查询。

[0002] 在努力帮助用户的搜索技术中，有时提供查询建议作为对搜索查询的响应的一部分。查询建议向用户提供用户可以选择的可替代查询。这可以帮助用户识别可能更好地适用于找到感兴趣信息的其他搜索查询。

发明内容

[0003] 在各种实施例中，可以通过识别所希望的子查询生成查询建议。可以累积搜索引擎数据来为各种查询确定使用特征。可以根据使用数据生成和排名潜在子查询。在对潜在子查询进行排名之后，当接收到搜索请求时，可以使用排名来选择子查询。可以将所选的子查询直接用作查询建议，或可以将子查询用作另一个查询建议引擎的输入。

[0004] 提供本发明内容来以简化形式介绍下面在具体实施方式中进一步描述的概念的选择。本发明内容不旨在识别要求保护主题的关键特征或必要特征，也不是旨在用于孤立地帮助确定要求保护主题的范围。

附图说明

[0005] 下面参考附图详细描述本发明，在附图中：

图 1 是用在实现本发明的实施例中的示范性计算环境的框图；

图 2 示意性地示出了适合执行本发明的实施例的系统；

图 3 描绘了按照本发明的实施例的方法的流程图；

图 4 描绘了按照本发明的实施例的方法的流程图；

图 5 描绘了按照本发明的实施例的方法的流程图；以及

图 6 和 7 描绘了根据使用汉语书面语言查询元素的本发明实施例的应用所得的结果。

具体实施方式

[0006] 概况

在各种实施例中，提供了生成查询建议的系统和方法。查询建议的生成可以基于首先识别具有高排名的一个或多个子查询。可以将一个或多个高排名子查询用作查询建议，或可以将一个或多个子查询用作传统查询建议方法的输入。在一些实施例中，这些系统和方法可以用于基于像包含 4 到大约 60 个查询元素的查询那样的较长查询的查询建议。在其他实施例中，可以使用不需要人为干预的系统和方法自动生成查询建议。这些系统和方法也可以与用于语言的查询元素的性质无关地应用于各种语言。因此，这些系统和方法可以有效地应用于查询元素是单词的查询(像英语的查询那样)，以及查询元素是字的查询(像

汉语、日语或朝鲜语那样的查询)。

[0007] 尽管向用户提供查询建议是传统做法,但提供高质量建议仍然存在许多障碍。一种这样的障碍是提供基于含有大量查询项的查询的查询建议。越来越多的搜索查询是包括4个或更多个关键词或查询元素的查询。项数增加的的一部分是使用“自然语言”查询的增加,其中查询是部分或甚至整个句子而不是关键词的集合。经验不足的用户更易于构想这样的长查询。长查询也可以用于进一步指定所希望搜索目标。当搜索庞大文档集合时,较长的查询可以有助于生成更相关排序的搜索结果。

[0008] 虽然较长查询可以为搜索者带来好处,但提供所建议查询的传统方法对于长查询可能不那么有效。许多查询建议方法基于流行项的附加或相关项的替代。对于只有两个或三个查询元素的搜索查询,每个查询元素可以用作改变查询的基础而无需生成从中选择的选项的过大清单。但是,随着一个查询越来越长,变体的数量可以成指数增加,导致为了确定查询建议而评估的大量排列。

[0009] 提供查询建议的另一个难题可能与跨越各种语言地提供查询建议有关。例如,查询建议算法使用自然语言查询的语法以便把重点放在最相关查询元素上。不幸的是,这种手段需要为使用的每种不同语言修改查询建议算法。由于像汉语那样的基于字书面语言的语法差异很大,所以这样的修改可能相当大。另外,即使在像英语那样的单一语言内,对于每个讲英语区域,语法的变体也可能需要不同的算法。

[0010] 相关问题是对于查询建议需要人为干预或训练的任何搜索引擎所面临的难题。人为训练可以包括提供特殊方式对待的单词的词典,例如作出建议时可以忽略的单词,或应该关联的单词。人为训练还可以包括提供用于开发关联性的一组训练文档。不管训练的类型为何,对人为干预的需要将意味着对查询建议系统的更新将是不频繁和耗时的。这可能导致来自查询建议系统的建议是过时的。

[0011] 在一些实施例中,提供了不依靠查询建议系统的人为训练地自动提供查询建议的系统和方法。该系统和方法可以独立于语言语法,使得该系统和方法稍作修改或不修改就可以用于各种语言。另外,该系统和方法可以有效地根据具有4个到大约60个查询项的查询作出查询建议。另外,在一些实施例中,该系统和方法可以与现有查询建议系统结合在一起使用。

[0012] 查询和子查询

查询可以包括一个或多个查询元素。查询元素是查询的独立部分。对于英语的查询,查询项通常是单词。注意,“单词”在这里表示搜索者可以用作和理解为单个查询项的一组字母、数字和/或其他符号。例如,寻找有关丙烷的附加信息的搜索者输入“C3H8”作为查询的一部分。在这种状况下,应该理解为“C3H8”构成查询项。可选地,对于允许搜索例如用引号或括号将一系列单词放入查询内的短语的搜索引擎,这样的短语可以被认为是单个查询项。相反,在有关 chocolate cake 的查询中,不认为字母“ch”是查询项,因为这不是所提交的查询内的完整“单词”。在像汉语、日语或朝鲜语那样的基于字书面语言中,查询元素可以是字。

[0013] 将查询中的查询长度定义成该查询中的查询元素的数量。在一些实施例中,可以为任何查询长度的所有查询提供查询建议。可替代的是,可以为查询长度为至少4个查询元素到大约60个查询元素的查询提供查询建议。查询长度可以是至少4个查询元素,至少

5 个查询元素，或至少 6 个查询元素。查询长度可以是大约 75 个或更少个查询元素，大约 60 个或更少个查询元素，大约 50 个或更少个查询元素，或大约 40 个或更少个查询元素。

[0014] 子查询是由母查询的一个或多个查询元素形成的查询。识别查询的可能子查询的一种方式是形成 n 元组（n-gram）。形成 n 元组的一种方式是在保留查询元素的次序的同时形成导致较短查询的查询元素的任何可能组合。换句话说，可以不改变其余查询元素的次序地从查询的头部、中部或尾部开始从查询中移除查询元素。这样的 n 元组可以称为位置相关 n 元组。对于四元素查询，可能子查询可以对应于四个 1 元素 n 元组、六个 2- 元素 n 元组、和四个 3- 元素 n 元组。可替代的是，可以形成允许查询元素在子查询中改变位置的位置无关子查询，对于包含四个查询元素的查询，存在四个 1 元素位置无关子查询、12 个 2 元素位置无关子查询、和 24 个 3- 元素位置无关子查询。

[0015] 在又一个实施例中，可以从母查询中使用查询项的连续串形成子查询。在这样的实施例中，可以从母查询的头部或尾部开始舍弃查询项，但如果查询项在保留在子查询中的其他查询项之间，则不舍弃该查询项。对于包含四个查询元素的查询，这种类型的实施例可以产生四个 1 元素子查询、三个 2 元素子查询、和两个 3 元素子查询。

[0016] 在可选实施例中，查询或子查询可以包括查询中的任何查询元素的显而易见变体。例如，一些文字处理程序现在都包括拼写检查功能，其中如果预定单词应该是什么是比较明确的，则可以自动纠正未出现在拼写检查词典中的单词。在这样的可选实施例中，可以在例如通过形成 n 元组来形成子查询的过程之前纠正拼写错误。可替代的是，当试图匹配查询时，可以顾及这样的拼写差异。

[0017] 在另一个可选实施例中，从母查询中形成的 n 元组（或其他子查询）可以局限于小于查询项的阈数的 n 元组或子查询。例如，根据母查询构建的 n 元组可以局限于含有 3 个或更少个查询项的 n 元组。在这样的例子中，尽管含有 5 个查询项的母查询潜在地含有包含 4 个查询项的子查询，但 4 查询项 n 元组可以因大于阈值 3 而被忽略掉。在实施例中，从母查询中形成的子查询可以局限于 2 个或更少个查询项，3 个或更少个查询项，4 个或更少个查询项，5 个或更少个查询项，或任何其他方便阈数。

[0018] 因为子查询小于相应母查询，所以可以从不止一个母查询中构建给定的子查询。例如，2 元素查询“chocolate cake”是 6 元素查询“how to make a chocolate cake”和 3 元素查询“chocolate cake ingredients”两者的子查询。将子查询的母计数定义成可以用于形成子查询的母查询的数量。在一些实施例中，母计数可以局限于只包括具有适当查询长度的母查询。例如，母计数可以基于具有 4 到大约 60 个查询元素的查询长度的母查询。

[0019] 查询日志文件

查询通常被提交给搜索引擎，搜索引擎根据相关性得分将该搜索查询与文档匹配。匹配文档可以以任何方便方式提供给用户。返回搜索结果的一种典型方式是提供搜索引擎认为与该查询有关的相关性得分最高的 10 个文档的清单。还可以与相关查询的建议一起提供链接，以便查看相关性得分较低文档的清单。搜索引擎可以通过任何方便方法确定与查询有关的文档的相关性得分。在初始结果页面上返回的文档的数量也可以是像 1，2，5，10，20，50 或其他数量那样任何方便的数量。

[0020] 当用户将搜索查询提交给搜索引擎时，可以在日志文件中跟踪和记录各种类型的信息。可以记录的一种信息是查询本身。在日志文件中，可以记录用户提交的查询。可选

地，这可以包括记录在查询项中可能存在拼写错误的查询。可选地，也可以跟踪查询被提交的总次数。

[0021] 可以跟踪的另一种信息是提交查询的不同用户的数量的计数。提交查询的不同用户的数量的计数可以提供查询普及性的指示。如上所述，可以跟踪查询被提交的总次数。不幸的是，当用户决定一个查询有用时，该用户可能多次提交该查询。这可能是由于，例如，在第一浏览器仍然在显示通过搜索识别的文档之一时希望打开第二浏览器再次观看查询结果导致的。确定查询普及性的一种潜在改进是跟踪不同用户的数量。不同用户的数量可以以多种方式确定。计数不同用户的一种方法是对于提交搜索查询的每种身份只增加不同用户计数一次。根据这种选项，一旦给定用户身份提交了一个搜索查询，则无论该用户身份提交多少次该查询，不同用户计数再也不会增加。计数不同用户的另一种方法是对于每种身份在给定时间段内只增加不同用户计数一次。例如，如果用户在 20 分钟的时段内提交了一个查询 5 次，则不同用户计数只增加一次。但是，如果该用户 10 天之后提交该查询，则不同用户计数将再次增加。任何方便时间段都可以用作允许另一次增加不同用户计数的时间段。例如，该时间段可以是一个小时，24 个小时，一个星期，一个月，或任何其他方便时间段。更一般地说，可以使用对提交了给定搜索查询的不同用户的数量进行计数的任何其他方便方法。

[0022] 可以记录的又一种信息是用户从查询结果中进入的文档链接的数量。一个选项可以是对用户基于搜索查询选择的文档的总数进行计数。因此，对于用户从查询的结果页面中选择的每个文档链接，都使计数增加。另一个选项可以是对用户选择的被认为是“高相关性”文档的文档的数量进行计数。认为文档是“高相关性”的一种方便代理是文档是否处在响应于搜索查询所显示的结果的第一页上。可替代的是，高相关性得分文档可以对应于与查询有关的相关性得分最高的文档的阈数，例如前 1，2，5，10，20，50 个或其他方便的阈数。

[0023] 在实施例中，可以选择对高相关性得分文档的定义以帮助确定与查询匹配的文档是否是用户感兴趣的。例如，用户可能提交了高分文档不是用户感兴趣的查询。相反，用户只查看未显示在第一页上和 / 或相关性得分低于高相关性得分截止分数的文档。在这种状况下，尽管搜索查询给出了用户感兴趣的文档，但这些感兴趣文档未作为高相关性得分文档出现。这往往指示该搜索查询可能没有其他一些搜索查询那么有价值，因为用户希望的结果未对应于高相关性得分结果。跟踪页面视图的总数和高相关性得分页面视图的数量两者可以有助于识别这样可能不那么有价值的查询。

[0024] 通过跟踪例如在限定地理区域内使用特定搜索引擎的所有用户的一群用户的各种数量，可以形成提供有关搜索查询的信息的查询日志文件。该查询日志文件可以包括查询的清单，以及每个查询的不同用户的数量、高相关性得分页面视图的数量、和页面视图的总数。如果需要的话，该查询日志文件还可以包括其他数据。该查询日志文件可以代表在例如一天或多天、一个或多个星期、一个或多个月、或一年或多年的任何方便时间段内对一群用户累积的数据。可选地，查询日志文件的大小可以局限于大约 6 个或更少个月，大约 10 个或更少个月，大约 12 个或更少个月，大约 18 个或更少个月，或大约 24 个或更少个月。限制查询日志文件的大小可以在处理查询日志文件数据时使计算时间更短。

[0025] 确定高排名子查询的考虑

查询日志文件中的数据可以用于帮助识别高排名子查询。高排名子查询可以通过多种

方法确定。在一些实施例中，用于识别高排名子查询的系统或方法可以基于一些或所有如下考虑。

[0026] 一种考虑可以是选择频繁使用的子查询。例如，未出现在查询日志文件中的子查询是搜索用户未提交过的子查询。这样的子查询不可能是相关的。更一般地说，不同用户的数量可以提供子查询的普及性以及相关性的指示。

[0027] 另一种考虑可以是选择尽可能多地保留包含在原始查询中的信息的子查询。一般说来，与母查询共有较多查询元素的子查询可以保留更多母查询的原始含义。因此，含有更多查询元素的查询和 / 或出现在母查询的查询元素的较高百分比可以是更相关查询的指示。

[0028] 又一种考虑可以是选择大部分页面视图是高相关性得分文档的子查询。如上所述，像显示在第一页上的文档那样的由搜索引擎返回的最高相关性文档可以提供给定搜索查询是否给出与用户的意图匹配的结果的指示。相对于页面视图的总数高相关性页面视图占大部分的搜索查询可以被认为是有效的搜索查询。

[0029] 再一种考虑可以是子查询的母查询的数量。生成查询建议的目标之一是向用户提供搜索相似主题的可替代方式。如果子查询具有相对较少的母查询，则子查询有较大的可能保留原始查询的本意。相反，如果子查询具有过多的母查询，则子查询有一定可能包括更通用的项，降低了子查询保留用户原意的可能性。因此，具有较大量母查询的子查询可以被认为是不太有效的搜索查询。

[0030] 处理查询日志文件

查询日志文件可以通过任何方便方法获得。可以如上所述生成查询日志文件，可以从另一个实体接收查询日志文件，或可以通过组合从两个或更多个实体收集的信息组装查询日志文件。在各种实施例中，在获得查询日志文件之后，可以通过识别高排名的一个或多个子查询发起生成查询建议的方法。识别高排名子查询的初步步骤是可以是生成潜在母查询的清单。在实施例中，只有查询元素的数量在最小值与最大值之间的查询，例如只有长度从 4 个到大约 60 个查询元素的查询可以用作母查询。可以抽取查询日志文件中具有适当长度的查询来形成母查询清单或文件。母查询清单提供可以用于生成排名子查询的查询。

[0031] 另一个可选初步步骤可以是过滤查询日志文件以排除一个或多个查询。由于多种原因，一些查询可以被认为是非所希望的。例如，可能希望排除与搜索成人内容或暴力内容有关的任何查询。另一个选项可以是排除普及性低的任何查询。例如包含拼错单词的查询那样的普及性低的查询可以代表查询数据中的“噪声”。考虑到这一点，可以排除不同用户和 / 或总页面视图的数量低于阈值的查询。在一些实施例中，可以排除少于大约 10 个不同用户的查询，少于大约 25 个不同用户的查询，或少于大约 100 个不同用户的查询。在其他实施例中，如果查询导致大约 10 个或更少的页面视图，大约 25 个或更少的页面视图，或大约 100 个或更少的页面视图，则可以排除该查询。可以通过任何方便方法排除查询，例如通过创建不包含被排除的查询的第二文件或清单，或通过在查询日志文件中标记被排除的查询。可替代的是，每当认为要对查询加以处理时，就可以进行查询日志文件中查询的过滤。注意，母查询清单可以在过滤查询日志之后，在过滤查询日志之前，或在进行了一些过滤之后形成。可选地，也可以对母查询文件或清单上进行过滤。

[0032] 如上所述的考虑可以用于确定查询日志文件中的高排名子查询。该方法可以从过

滤查询日志文件以移除非所希望的查询开始。然后可以使这个经过滤的查询清单对应于查询文件或查询清单。然后可以通过抽取具有所希望长度的所有查询构建母查询文件或清单。查询日志文件也可以用于确定每个查询的“频率”。在实施例中，可以根据查询的不同用户和页面视图的数量计算查询的频率，可能包括对页面访问的总数相对于高相关性页面视图的数量的单独考虑。在另一个实施例中，可以根据具有与方程(1)相似的特征的方程计算频率：

(1) 频率 =(# 不同用户)*(# 高相关性页面视图)/[1+(# 总页面视图)]

在方程(1)中，频率与不同用户的数量成正比。频率也与高相关性页面视图与总页面视图的比率成正比。针对以上方程格式的变体是可能的。首先注意到，在方程中使用了“1+# 总页面视图”。包括“1”是为了防止该表达式变成未定义的。在那个位置中使用非零值对于避免计算错误是有价值的。但是，本领域的普通技术人员应该认识到，包括这个常数是为了便于计算。在其他实施例中，如果适当地管理对查询日志文件的处理，则可以在计算频率之前过滤掉查询日志文件中在频率计算时导致未定义值的任何查询。这种便于计算的类型可以用在下面所示的其他方程中以避免未定义值的潜在性。

[0033] 可以对方程(1)作出的另一种修改是包括像对数项那样的一些项。在一些情况下，查询日志文件可以代表若干个月或甚至几年累积的数据。在这样的情况下，查询日志文件中的许多数值，例如页面视图的数量或不同用户的数量从绝对意义上来讲可能很大。为了便于管理大值，方程(1)中的一些或所有项可以作为对数值包含。例如，方程(1)中的不同用户部分可以取而代之地表达成“log[1+(# 不同用户)]”。例如底数 2，底数 10，或底数 20 的任何方便底数都可以用于对数。还要注意到，包括 1 作为非零值是为了便于避免导致未定义值的计算。

[0034] 也可以处理母查询文件来识别要考虑的潜在子查询。如上所述，潜在子查询可以通过形成母查询的 n 元组来形成。可替代的是，可以将任何其他方便的方法用于形成子查询，例如形成查询元素比母查询少的所有位置无关的变体。在一些实施例中，潜在子查询可以局限于查询项少于阈数的子查询。

[0035] 在形成潜在子查询之后，可以将潜在子查询与查询日志文件中的查询匹配。在一个实施例中，可以将潜在子查询只与查询日志文件中的精确匹配进行匹配。可以舍弃在查询日志文件中没有匹配项的任何潜在子查询。也可以为每个匹配子查询计算母查询的数量。母查询数量的计算可以发生在匹配过程之前，期间或之后。可以将给定子查询的母查询的总数称为那个子查询的母计数。

[0036] 此刻，可以为每个查询计算若干数值。首先，可以相对于子查询的相应母查询地为每个子查询计算加权频率。在一个实施例中，可以将加权频率计算成：

(2) 加权频率 =(# 子查询中的元素)* 频率 /(# 母查询中的元素)

加权频率顾及了子查询与母查询相比的相对项数。这可以针对可以得出该子查询的每个母查询来计算，因此，取决于得到考虑的特定母查询，具有不止一个母查询的子查询可以具有多个不同加权频率值。接着，通过计及子查询的母查询的数量，可以将加权频率值用于计算归一化加权频率值。一种归一化方法可以使用过滤查询文件(或如果未进行过滤，则为查询日志文件)中的查询的总数与可以产生子查询的母查询的数量，或母计数的比值。在搜索背景下，这种归一化加权频率类似于 TFIDF (词频逆文档频率)值。归一化加权频率的一

种可能格式是：

（3）归一化加权频率 =log[加权频率]*(查询清单的大小)/(母计数)

方程(3)可以用于得出子查询的归一化加权频率值。该对数(log)可以具有像底数 2，底数 10 或底数 20 那样的任何方便底数。

[0037] 对于具有不止一个母查询的子查询，可以存在多个归一化加权频率值。为了得出用作排名值的单个值，可以对归一化加权频率值求平均，例如通过简单求和归一化加权频率值并除以求和项数。可选地，可以将这个平均频率值用作排名值。在另一个实施例中，可以将平均频率乘以子查询中的元素数量来调整平均频率值。这个调整频率值也可以用作排名值。为了简化下面的讨论，将调整频率值用作排名值。但是，此刻可以作出其他调整以便进一步修改每个子查询的排名值。

[0038] 在确定了子查询的排名值之后，可以创建包含所有子查询和相应排名值的排名清单。这个子查询和排名值的清单可以用于生成查询建议。这个清单可以称为排名清单。

[0039] 生成查询建议

当接收到查询时，可以使用子查询和排名值的清单生成查询建议。当接收到查询时，可以识别可能的子查询。可能的子查询可以使用上述方法之一识别，例如从查询中创建 n 元组或通过创建查询元素少于原始查询的子查询的所有可能位置无关的变体。一旦识别出可能的子查询，就从排名清单中确定每个可能子查询的排名。可以选择最高排名子查询，或可以选择若干最高排名子查询，例如前三个子查询。

[0040] 一个或多个所选子查询可以作为查询建议直接提供。可替代的是，可以将一个或多个所选子查询用作使用生成查询建议的其他方法的基础。例如，可以将一个或多个所选子查询用作对一个查询被补充附加项以形成所建议查询的方法的输入。可替代的是，可以将一个或多个所选子查询用作查询建议引擎的输入，并且可以将查询建议引擎生成的一个或多个查询作为查询建议来提供。由于所选子查询短于初始查询，所以可以使用所选子查询更好地执行这样生成查询建议的传统方法。

[0041] 例 1—使用英语查询项的例子

为了演示按照本发明实施例的操作，提供了如下预言性例子。下面提供的排名值旨在例示本发明的操作。

[0042] 在如下的例子中，考虑了可以向其提供查询建议的两个查询。第一个查询是“chocolate cake nutrition facts”，而第二个查询是“recipe for baking chocolate cake)。在如下的例子中，例示了子查询的查询项数局限于两个查询项的本发明实施例。

[0043] 首先，查询“chocolate cake nutrition facts”可以用于演示作为潜在子查询的 n 元组的构建。对于这个查询，存在包含 1 个查询元素的四个 n 元组 :chocolate ;cake ; nutrition ;以及 facts。存在包含 2 个查询元素的六个 n 元组 :chocolate cake ;chocolate nutrition ;chocolate facts ;cake nutrition ;cake facts ;以及 nutrition facts。存在包含 3 个查询元素的四个 n 元组 :chocolate cake nutrition ;chocolate cake facts ; chocolate nutrition facts ;以及 cake nutrition facts。但是，因为这个实施例只使用含有 2 个或更少个查询元素的子查询，所以不再考虑具有 3 个查询元素的四个 3 元组。因为在本例中使用 n 元组，所以查询中的词序在子查询中未变更。

[0044] 在确定了潜在 n 元组(在这种情况下，1 个查询元素和 2 个查询元素 n 元组)之后，

可以将 n 元组与排名清单相比较，以确定最高排名子查询。表 1 示出了几个子查询的排名值。表 1 中的排名值代表按照本发明实施例生成的排名值。

[0045] 表 1

子询问	排名值
chocolate cake	2,847,686
nutrition facts	2,315,702
chocolate	1,910,153
nutrition	1,522,711
cake	669,997
facts	486,333

[0046] 对于显示在表 1 中的子查询，子查询“chocolate cake”具有最高排名值。假设其他 2 元素子查询具有较低排名值，则“chocolate cake”将是选作查询建议，或另一种查询建议算法输入的第一子查询。在选择不止一个子查询的实施例中，“nutrition facts”将是第二所选子查询，而“chocolate”将第三选择。

[0047] 可以将相同过程应用于查询“recipe for baking chocolate cake”。表 2 示出了几个子查询的可能排名值。注意，对于表 2 中的例子，一些排名值代表只用于例示目的的样本值。

[0048] 表 2

子询问	排名值
recipe chocolate cake	3,122,456
chocolate cake	2,847,686
baking chocolate	2,222,222
recipe cake	1,854,321
cake	669,997

[0049] 在表 2 中，子查询“recipe chocolate cake”具有最高排名值。注意，子查询“chocolate cake”和“chocolate”在表 1 和表 2 两者中具有相同排名值。在各种实施例中，子查询在排名清单中具有单个排名值。一旦形成排名清单，就使用排名值来选择所有子查询，因此，特定子查询的排名值不会随用户提交的查询而变。

[0050] 例 2—使用汉语查询项的例子

查询日志文件通过将根据用汉字提交的搜索所得的信息存储到搜索引擎中生成。按照本发明的实施例分析查询日志文件以产生搜索查询的排名表。图 6 和 7 示出了包括搜索查询和来自基于搜索查询的各种子查询的排名表的排名的清单的表格。与像英语那样查询元素是由来自字母表的字母组成的单词的语言相对照，图 6 和 7 表明可以容易地将本发明的实施例应用于像汉语或日语那样查询元素是字的语言。

[0051] 附加实施例

上面已经简要描述了本发明的各种实施例的概况，现在描述适合执行本发明的示范性操作环境。一般性地参数附图，尤其首先参照图 1，实现本发明实施例的示范性操作环境被一般地显示和指定成计算设备 100。计算设备 100 只不过是适当计算环境的一个例子，而无意暗示对本发明的使用或功能的范围的任何限制。计算设备 100 也不应该被解释为具有与所例示的任何一个部件或部件的组合有关的任何依赖或要求。

[0052] 本发明的实施例可以在被计算机或像个人数据助理或其他手持设备那样的其他机器执行的计算机代码或包括像程序模块那样的计算机可执行指令的机器可用指令的一般背景下来描述。一般说来，包括例程、程序、对象、部件、数据结构等的程序模块指的是执行特定任务或实现特定抽象数据类型的代码。本发明可以在包括手持设备、消费类电子产品、通用计算机、更专用计算设备等的多种系统配置中实施。本发明也可以在由通过通信网络链接的远程处理设备执行任务的分布式计算环境下实施。

[0053] 继续参考图 1，计算设备 100 包括直接或间接耦合如下设备的总线 110 :存储器 112、一个或多个处理器 114、一个或多个呈现部件 116、输入 / 输出（I/O）端口 118、I/O 部件 120、和例示性电源 122。图 1 进一步示出了按照本发明实施例的查询建议生成部件 117。总线 110 代表可以是一条或多条总线（像地址总线、数据总线、或其组合）。尽管图 1 的各种框为了清楚起见用直线示出，但实际上，不用那么清楚描绘各种部件，比方说，这些直线是灰色的和模糊的更准确。例如，可以认为像显示设备那样的呈现部件是 I/O 部件。另外，许多处理器都具有存储器。发明人认识到技术的本质就是这样，并重申图 1 的示图仅仅例示了可以与本发明的一个或多个实施例结合在一起使用的示范性计算设备。在像“工作站”、“服务器”、“膝上型电脑”、“手持设备”等那样的类别之间不加区分，因为所有这些都被设想在图 1 的范围之内并统称为“计算设备”。

[0054] 计算设备 100 通常包括多种计算机可读介质。计算机可读介质可以是可通过计算设备 100 访问的任意可用介质，并且包括易失性和非易失性介质两者，或可移除和非可移除介质两者。举例来说，但非限制性地，计算机可读介质可以包含计算机存储介质和通信介质。计算机存储介质包括以用于存储信息的任何方法或技术实现的易失性和非易失性介质，或可移除和非可移除介质， 该信息例如是计算机可读指令、数据结构、程序模块或其他数据。计算机存储介质包括但不限于随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪存储器或其他存储器技术、CD-ROM、数字多功能盘（DVD）或其他全息存储器、磁盒、磁带、磁盘存储体或其他磁存储设备、载波、或可以用来编码所希望信息并且可以被计算设备 100 访问的任何其他媒介。在另一个实施例中，计算机存储介质可以是有形计算机存储介质。在又一个实施例中，计算机存储介质可以是非瞬时计算机存储介质。

[0055] 存储器 112 包括易失性和 / 或非易失性存储器形式的计算机存储介质。该存储器可以是可移除的，非可移除的，或其组合。示范性硬件设备包括固态存储器、硬盘驱动器、光盘驱动器等。计算设备 100 包括从像存储器 112 或 I/O 部件 120 那样的各种实体中读取数据的一个或多个处理器。（多个）呈现部件 116 向用户或其他设备呈现数据指示。示范性呈现部件包括显示设备、扬声器、打印部件、振动部件等。

[0056] I/O 端口 118 使计算设备 100 可以与一些可以内置的包括 I/O 部件 120 的其他设备逻辑耦合。例示性部件包括麦克风、操纵杆、游戏垫、卫星天线、扫描仪、打印机、无线设备

等。

[0057] 本发明的实施例涉及生成搜索查询建议的系统和方法。现在转到图2，其中例示了示出依照本发明实施例的示范性计算系统200的框图。本领域的普通技术人员应该明白和懂得，显示在图2中的计算系统200仅仅是一种适当计算系统环境的例子，而无意暗示对本发明实施例的使用或功能的范围的任何限制。计算系统200也不应该被解释为具有与本文例示的任何单个部件或部件的组合体有关的任何依赖或要求。并且，计算系统200可以作为独立产品，作为软件开发环境的一部分，或它们的任何组合来提供。

[0058] 计算系统200包括查询和结果分析器206、查询过滤器218、搜索引擎214、查询建议引擎210、排名生成器212和子查询生成器208，所有这些都经由网络204和/或经由公用设备上的地点相互通信。这些元件的一个或多个可以取决于实施例地可选的。虽然查询和结果分析器206、查询过滤器218、搜索引擎214、查询建议引擎210、排名生成器212和子查询生成器208在图2中被显示成分立元件，但在一些实施例中可以组合这些元件的一个或多个。网络可以非限制性地包括一个或多个局域网(LAN)和/或广域网(WAN)。这样的联网环境在办公室、企业范围内计算机网络、内联网和互联网中是司空见惯的。于是，这里不再对网络204作进一步描述。

[0059] 搜索引擎214可以是接收搜索查询和生成作为结果返回的匹配文档的清单的任何适当搜索引擎。可选地，查询和结果分析器206可以是搜索引擎214的一部分。查询和结果分析器206可以分析用户与搜索引擎之间的交互的各个方面。可以将这种分析存储在查询日志文件中。查询和结果分析器206可以跟踪搜索引擎214接收的查询；跟踪提交查询的不同用户；跟踪用户查看的响应于查询而提供的文档；以及跟踪用户查看的响应于查询而提供的高相关性文档。

[0060] 查询过滤器218可选地可以是查询和结果分析器206和/或搜索引擎214的一部分。查询过滤器218可以根据像成人或暴力内容那样的查询的性质来拒绝考虑一些查询。查询过滤器218也可以根据查询的普及性或频率排除一些查询。

[0061] 子查询生成器208可以生成与给定母查询相对应的子查询。子查询生成器208也可以确定查询内查询项的数量和/或子查询的母查询的数量。

[0062] 排名生成器212可以生成和提供子查询的排名清单。排名生成器212可以没有进一步人为干预地根据来自查询日志文件的信息自动计算排名。可选地，查询和结果分析器206和/或子查询生成器208可以是排名生成器212的一部分。

[0063] 查询建议引擎210可以根据输入查询提供查询建议。当使用根据来自排名生成器212的排名选择的子查询时，查询建议引擎210可以通过像根据与现有查询项的相似性添加附加项或添加和/或替换一些项的任何方便方法来生成所建议的查询。在一些实施例中，查询建议引擎210可以是可以根据所选子查询的使用而不是提交给搜索引擎214的查询提供改进结果的传统查询建议引擎。

[0064] 图3描绘了示出按照本发明的实施例的方法的流程图。在显示在图3中的实施例中，获取(310)查询日志文件。识别(320)像查询日志文件中的查询那样的含有至少4个查询元素的查询。为识别的子查询确定(330)子查询。将确定的子查询与来自查询日志文件的查询匹配(340)。为匹配的子查询计算(350)排名。然后接收(360)搜索查询。为接收的搜索查询确定(370)搜索子查询。根据为搜索子查询计算的相应排名选择(380)一个或

多个搜索子查询。根据所选的搜索子查询提供(390)建议查询。

[0065] 图4描绘了示出按照本发明的另一个实施例的方法的流程图。在显示在图4中的实施例中,获取(410)查询日志文件。该查询日志文件包括含有基于字的查询元素的查询。识别(420)像查询日志文件中的查询那样的含有至少4个查询元素的查询。为识别的子查询确定(430)子查询。将确定的子查询与来自查询日志文件的查询匹配(440)。为匹配的子查询计算(450)排名。然后接收(460)搜索查询。为接收的搜索查询确定(470)搜索子查询。根据为搜索子查询计算的相应排名选择(480)一个或多个搜索子查询。根据所选的搜索子查询提供(490)建议查询。

[0066] 图5描绘了示出按照本发明的又一个实施例的方法的流程图。在显示在图5中的实施例中,获取(510)查询日志文件。识别(520)像查询日志文件中的查询那样的含有至少4个查询元素的查询。为识别的子查询确定(530)子查询。将确定的子查询与来自查询日志文件的查询匹配(540)。为匹配的子查询计算排名。排名的计算包括为每个子查询计算(550)母查询的数量。为每个子查询计算(560)频率。然后为每个子查询计算(570)归一化加权频率。接着,为每个子查询计算(580)平均归一化加权频率。根据子查询的平均归一化加权频率生成(590)排名清单。

[0067] 在另一个实施例中,可以提供生成查询建议的方法。可选地,该方法可以以一个或多个计算机可读介质的形式提供,该计算机可读介质包含当被执行,提供生成查询建议的方法的计算机可执行指令。该方法包括获取查询日志文件。可选地,查询日志文件中的查询可以是含有与像汉语、日语或朝鲜语那样的基于字的书面语言相对应的查询元素的查询。可以识别含有至少4个查询元素的包含在查询日志文件中的查询。可以为每个识别的查询确定子查询。可以将确定的子查询与查询日志文件中的查询相匹配。可以为每个匹配的子查询计算排名,该排名基于不同用户的数量、页面视图数据、子查询中的查询元素的数量和子查询的母查询的数量。然后可以接收搜索查询。可以为接收的搜索查询确定搜索子查询,其中搜索子查询的至少一个对应于具有所计算排名的匹配子查询。可以根据所选一个或多个搜索子查询的相应计算排名选择一个或多个搜索子查询。然后可以根据所选一个或多个搜索子查询提供一个或多个建议查询。

[0068] 在又一个实施例中,可以提供生成查询建议的方法。可选地,该方法可以以一个或多个计算机可读介质的形式提供,该计算机可读介质包含当被执行,提供生成查询建议的方法的计算机可执行指令。该方法包括获取查询日志文件。可选地,查询日志文件中的查询可以是含有与像汉语、日语或朝鲜语那样的基于字的书面语言相对应的查询元素的查询。可以识别含有至少4个查询元素的包含在查询日志文件中的查询。可以为每个识别的查询确定子查询。可以将确定的子查询与查询日志文件中的查询相匹配。可以为每个匹配的子查询计算排名。该计算可以包括计算每个子查询的母查询的数量。可以根据不同用户的数量和页面视图信息为每个子查询计算频率。可以根据子查询中的查询元素的数量;母查询中的查询元素的数量;查询日志文件中的查询的数量;以及子查询的母查询的数量为每个子查询计算一个或多个归一化加权频率值。为子查询计算的归一化加权频率值的数量对应于子查询的母查询的数量。然后可以根据子查询的一个或多个归一化加权频率值和子查询的母查询的数量计算平均归一化加权频率值。可以根据子查询的平均归一化加权频率值生成子查询的排名清单。

[0069] 上面结合无论从哪个方面来看都旨在例示而非限制的特定实施例描述了本发明的实施例。可替代实施例在不偏离本发明的范围的情况下对于本发明所属领域的普通技术人员来说是显而易见的。

[0070] 从上文可以看出，本发明十分适用于达到上文所述的所有目标和目的，以及显而易见的和结构固有的其他优点。

[0071] 应该理解，某些特征和子组合是实用的，并且可以无需涉及其他特征和子组合地采用。这可以通过权利要求来设想并在权利要求的范围之内。

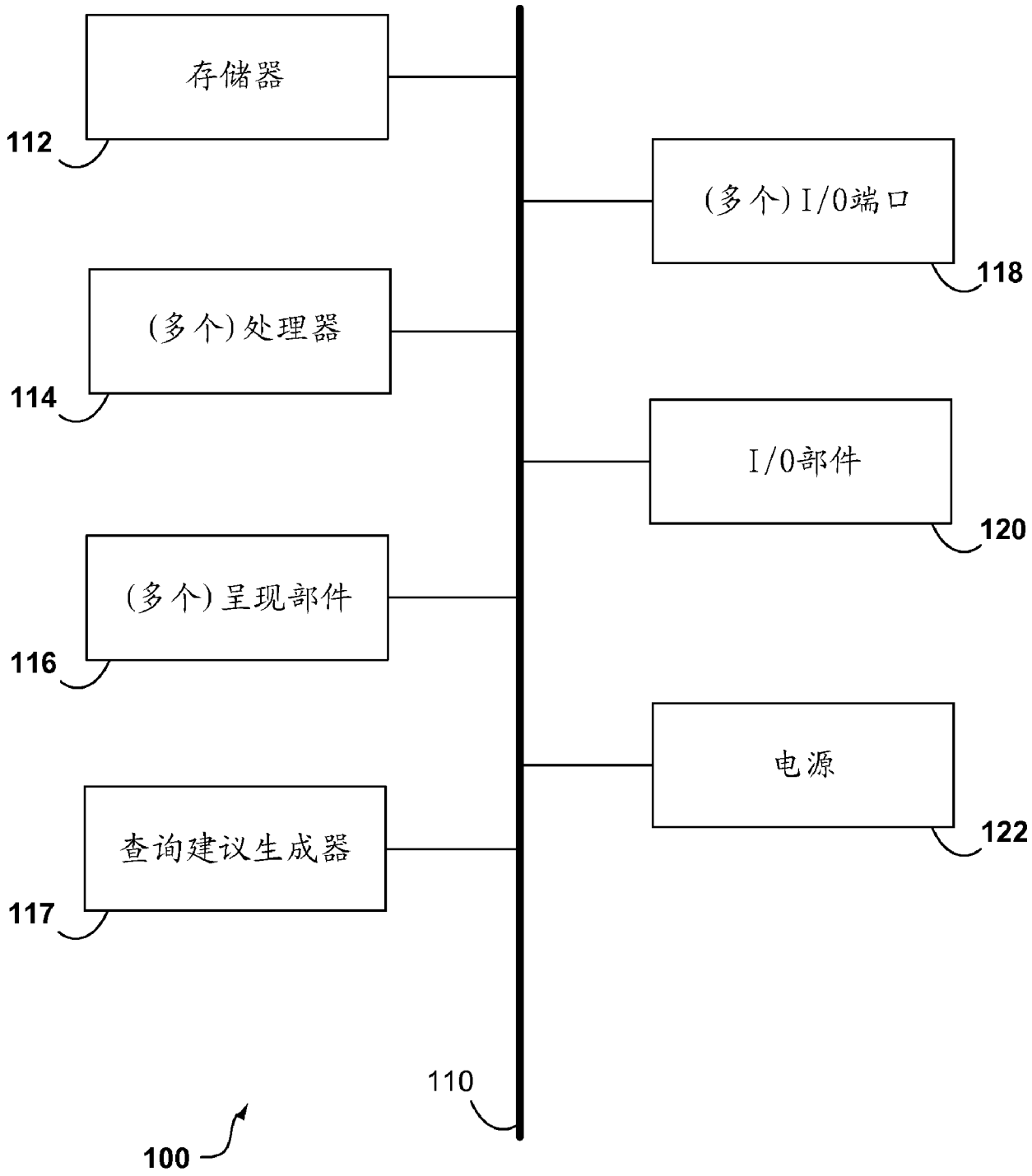
存储器
112
(多个)处理器
114
(多个)呈现部件
116
查询建议生成器
117
(多个)I/O端口
118
I/O部件
120
电源
122
110
100

图 1

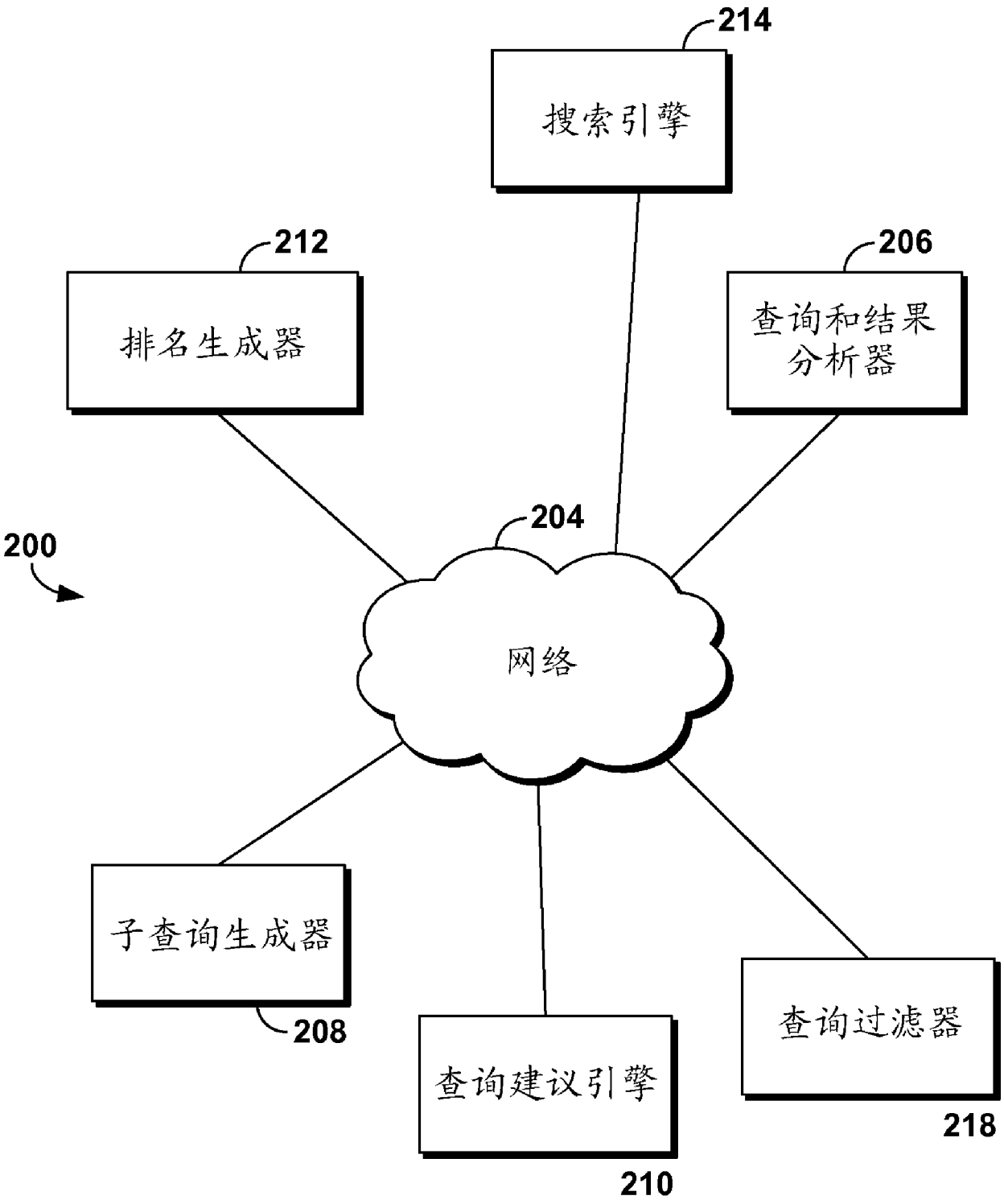
214
搜索引擎
212
排名生成器
206
查询和结果分析器
200
204
网络
子查询生成器
208
查询建议引擎
210
查询过滤器
218

图 2

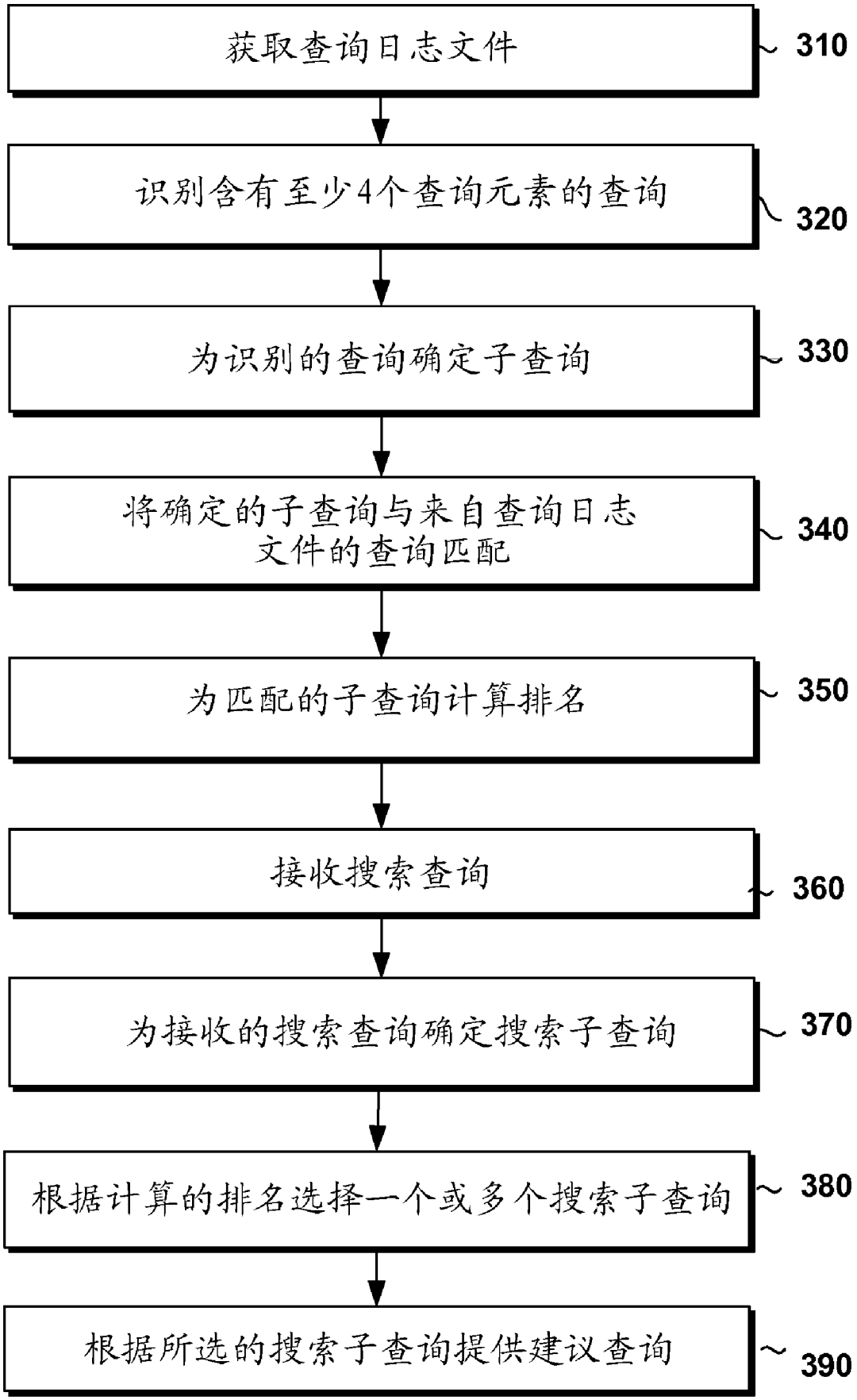
获取查询日志文件
310
识别含有至少4个查询元素的查询
320
为识别的查询确定子查询
330
将确定的子查询与来自查询日志文件的查询匹配
340
为匹配的子查询计算排名
350
接收搜索查询
360
为接收的搜索查询确定搜索子查询
370
根据计算的排名选择一个或多个搜索子查询
380
根据所选的搜索子查询提供建议查询
390

图 3

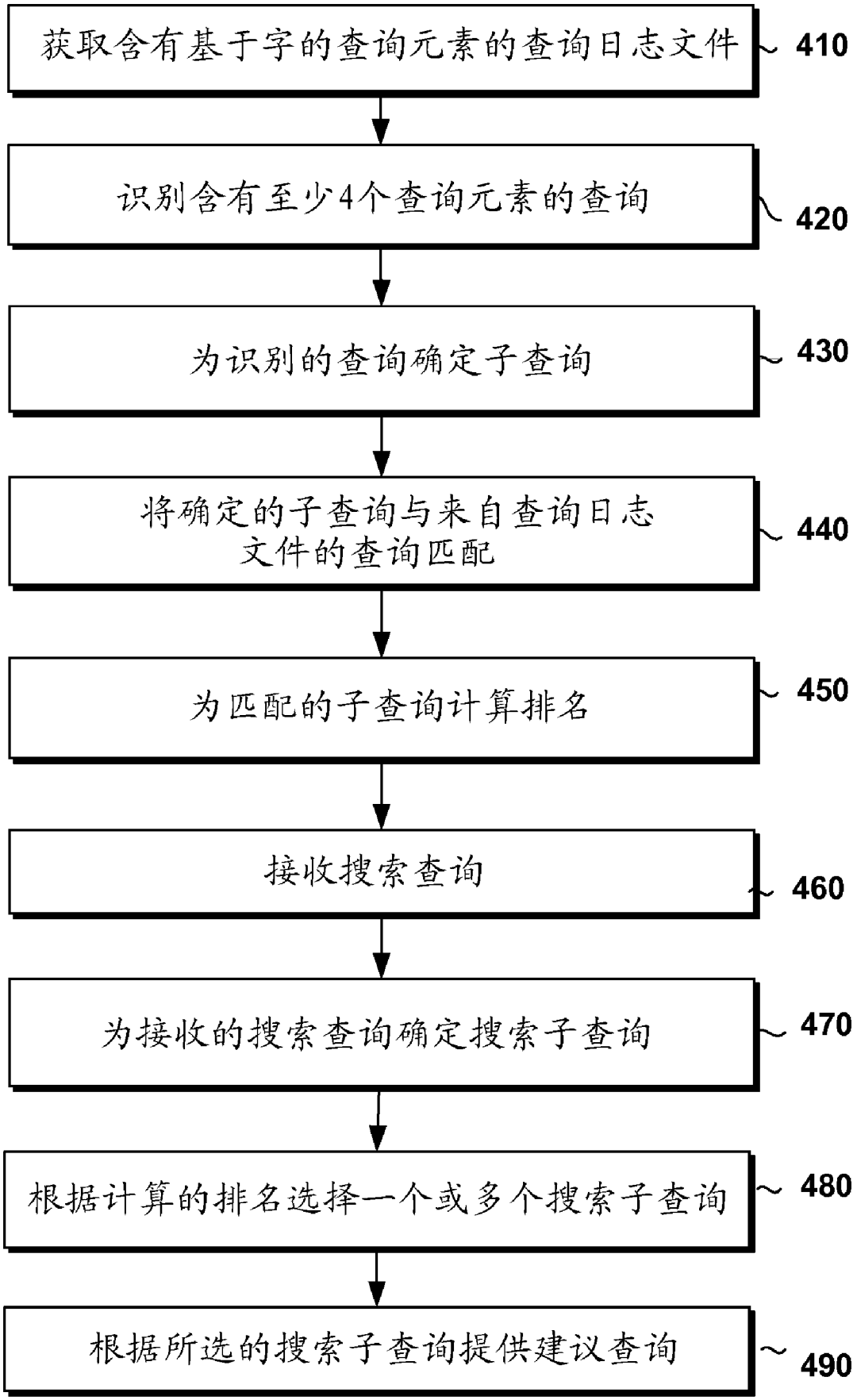
获取含有基于字的查询元素的查询日志文件
410
识别含有至少4个查询元素的查询
420
为识别的查询确定子查询
430
将确定的子查询与来自查询日志文件的查询匹配
440
为匹配的子查询计算排名
450
接收搜索查询
460
为接收的搜索查询确定搜索子查询
470
根据计算的排名选择一个或多个搜索子查询
480
根据所选的搜索子查询提供建议查询
490

图 4

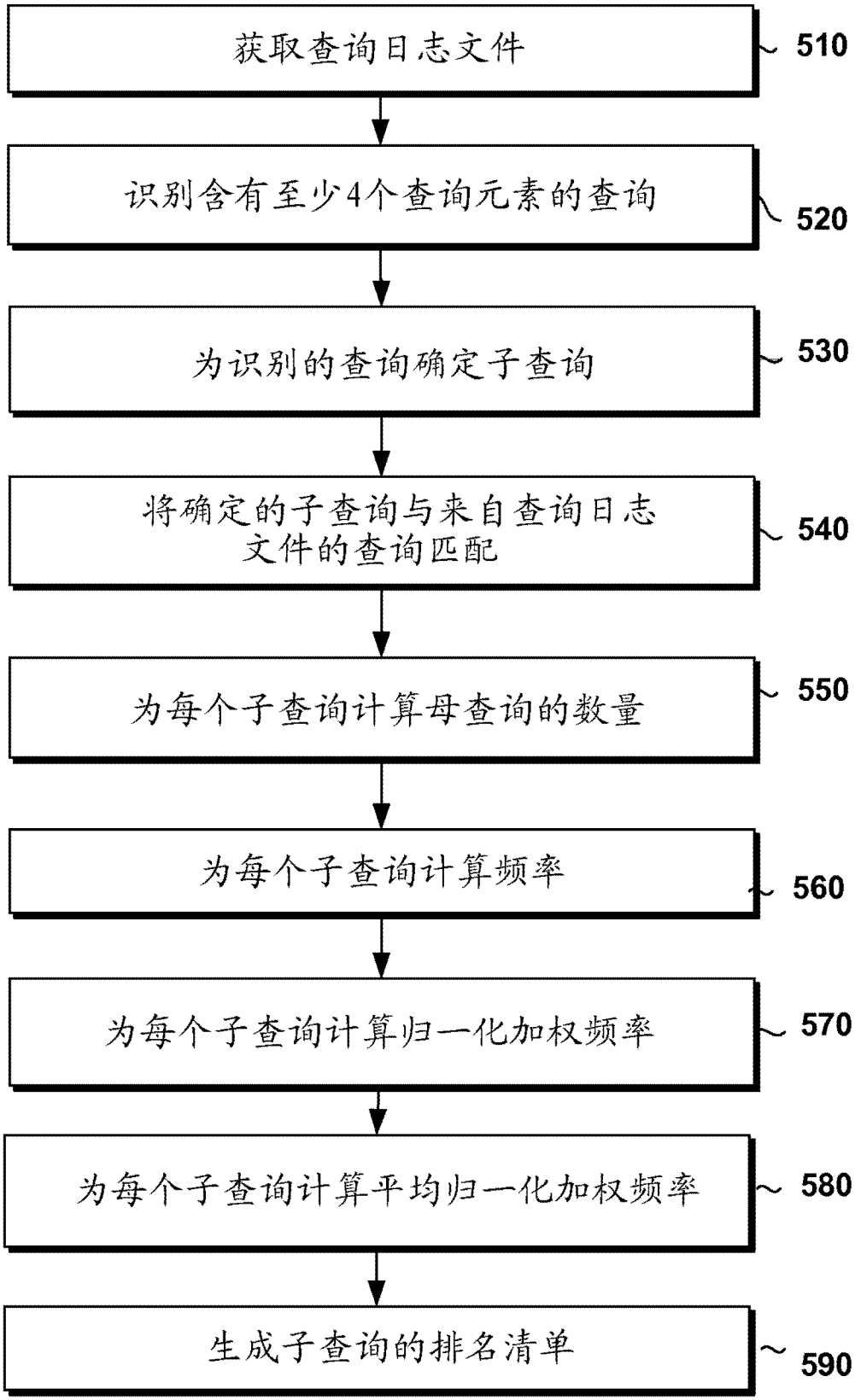
获取查询日志文件
510
识别含有至少4个查询元素的查询
520
为识别的查询确定子查询
530
将确定的子查询与来自查询日志文件的查询匹配
540
为每个子查询计算母查询的数量
550
为每个子查询计算频率
560
为每个子查询计算归一化加权频率
570
为每个子查询计算平均归一化加权频率
580
生成子查询的排名清单
590

图 5

原始查询	子查询	排名值
免费下载歌	下载歌	22120000
免费下载歌	费下	20157000
免费下载歌	费下载	17554000
免费下载歌	免费下	10398000

图 6

原始查询	子查询	排名值
我们每一天做工	每一天	360483
我们每一天做工	我们	252751
我们每一天做工	一天	164423
我们每一天做工	做工	55826
我们每一天做工	每一	54112

图 7