[19] 中华人民共和国国家知识产权局

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

[12] 发明专利申请公布说明书

[21] 申请号 200480038818.7

[43] 公开日 2007 年 1 月 17 日

[11] 公开号 CN 1898670A

[22] 申请日 2004.12.29
[21] 申请号 200480038818.7
[30] 优先权
[32] 2003.12.30 [33] US [31] 10/749,730
[86] 国际申请 PCT/US2004/043918 2004.12.29
[87] 国际公布 WO2005/066847 英 2005.7.21
[85] 进入国家阶段日期 2006.6.23
[71] 申请人 GOOGLE 公司
地址 美国加利福尼亚州
[72] 发明人 亚历山大·M·弗朗茨
莫妮卡·亨青格尔
[74] 专利代理机构 北京康信知识产权代理有限责任公司
代理人 余 刚

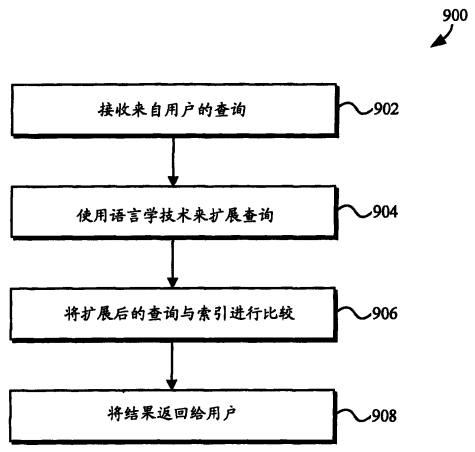
权利要求书 5 页 说明书 13 页 附图 10 页

[54] 发明名称

提高搜索质量的系统和方法

[57] 摘要

公开了用于改善搜索质量的系统和方法。 使用多种语言学技术扩展搜索查询。 例如，可以用从复合字、字尾变化形式、和/或正字法变化的数据库中获得的相关字来补充查询中的字。 扩展后的查询可以用来执行对相应文档的搜索。 可以用类似的技术扩展文档索引。



1. 一种方法，包括:

接收包括至少一个查询项的查询;

执行以下步骤中的至少一个:

（A）确定所述查询是否包括一个或多个复合查询项，如果是，则自动扩展所述查询，以包括所述一个或者多个复合查询项的一个或多个可选表示;

（B）确定一个或多个查询项是否包括在一组字尾变化形式中，如果是，则自动扩展所述查询，以包括来自所述组字尾变化形式的一个或多个相应的字尾变化形式;以及

（C）确定一个或者多个查询项是否包括在一组可选拼写中，如果是，则自动扩展所述查询，以包括来自所述组可选拼写的一个或多个相应的可选拼写;

使用所扩展的查询来搜索数据库;以及

返回结果给用户。

2. 根据权利要求1所述的方法，其中，所述方法包括确定所述查询是否包括一个或者多个复合查询项，如果是，则自动扩展所述查询，以包括所述一个或多个复合查询项的一个或多个可选表示。

3. 根据权利要求1所述的方法，其中，所述方法包括确定一个或多个查询项是否包括在一组字尾变化形式中，如果是，则自动扩展所述查询，以包括来自所述组字尾变化形式的一个或多个相应的字尾变化形式。

4. 根据权利要求 1 所述的方法，其中，所述方法包括确定一个或多个查询项是否包括在一组可选拼写中，如果是，则自动扩展所述查询，以包括来自所述组可选拼写的一个或多个相应的可选拼写。

5. 根据权利要求 4 所述的方法，其中，所述方法还包括执行(B)，以及其中，在自动扩展所述查询以包括来自所述组字尾变化形式的一个或多个相应的字尾变化形式的步骤之前，执行自动扩展所述查询以包括来自所述组可选拼写的一个或多个相应的可选拼写的步骤。

6. 根据权利要求 1 所述的方法，其中，所述方法包括执行所述步骤（A）、（B）、和（C）中的至少两个步骤。

7. 根据权利要求 1 所述的方法，其中，确定所述查询是否包括一个或多个复合查询项的步骤包括将查询项与复合项表相比较。

8. 根据权利要求 7 所述的方法，其中，所述一个或多个复合查询项的所述一个或多个可选表示从所述复合项表中获得。

9. 根据权利要求 1 所述的方法，其中，所述查询用德文书写。

10. 根据权利要求 1 所述的方法，其中，以排列的顺序执行所述操作。

11. 一种方法，包括:

识别一组与文档相关联的项目;

通过以下内容中进一步与所述文档相关联的一个或多个内容来扩展与所述文档相关联的所述组项目:

与所述文档相关联的所述组项目中的至少一个项目的一个或多个可选拼写；

与所述文档相关联的所述组项目中的至少一个复合项的一个或多个可选表示；以及

与所述文档相关联的所述组项目中的至少一个项目的一个或多个附加的字尾变化形式；

使用所扩展的组项来对所述文档建索引。

12. 根据权利要求 11 所述的方法，还包括：

从用户接收查询，所述查询包括一个或多个所述可选拼写、可选表示、或附加的字尾变化形式；以及

作为对所述查询的响应，将所述文档识别给所述用户。

13. 根据权利要求 11 所述的方法，其中，所述文档包括网页。

14. 一种方法，包括：

在第一组文档中搜索连字符连接字；

在所述第一组文档中搜索与所述连字符连接字相对应的非连字符连接字；以及

在所述连字符连接字与所述对应的非连字符连接字之间建立一组联系。

15. 根据权利要求 14 所述的方法，还包括：

在所述第一组文档中搜索与所述非连字符连接字和相应的连字符连接字相对应的分离字对；

进一步将所述分离字对与在所述连字符连接字与所述对应的非连字符连接字之间的所述组联系相关联。

16. 根据权利要求 14 所述的方法，还包括:

接收来自用户的查询，所述查询包括第一查询项;

在连字符连接字与对应的非连字符连接字之间的所述组联系中定位所述第一查询项; 以及

扩展所述查询，以包括第二查询项，所述第二查询项在连字符连接字与对应的非连字符连接字之间的所述组联系中与所述第一个查询项相关联。

17. 根据权利要求 16 所述的方法，还包括:

使用所扩展的查询来执行搜索;

发送给用户响应于所述查询的一个或多个文档的表。

18. 根据权利要求 14 所述的方法，还包括:

在文档中定位连字符连接字;

在连字符连接字与对应的非连字符连接字之间的所述组联系中搜索所述连字符连接字;

如果在连字符连接字与对应的非连字符连接字之间的所述组联系中没有找到所述连字符连接字，则去掉所述连字符连接字的连字符; 以及

使用所述去连字符字对所述文档建索引。

19. 一种计算机程序包，其驻留于计算机可读介质上，所述计算机程序包包括指令，当处理器执行所述指令时，所述指令使所述处理器执行选自包括以下步骤的组中的操作:

通过包括至少一个查询项的一个或多个可选拼写来扩展从用户接收的查询;

用至少一个复合查询项的一个或多个可选表示来扩展所述查询；以及

用至少一个查询项的一个或多个字尾变化形式来扩展所述查询。

20. 根据权利要求 19 所述的计算机程序包，还包括指令，当处理器执行所述指令时，使所述处理器执行包括以下步骤的操作：

使用所扩展的查询来搜索文档数据库；

识别一个或多个响应于所扩展的查询的文档；以及

准备所述一个或多个文档的表，用于向用户传送。

21. 根据权利要求 19 所述的计算机程序包，还包括指令，当处理器执行所述指令时，使所述处理器执行包括以下步骤的操作：

发送所扩展的查询到另一计算机系统；以及

接收来自所述另一计算机系统的响应于所扩展的查询的一个或多个文档的表。

22. 一种信息检索系统，所述系统包括：

文档数据库，所述文档数据库包括一组文档；以及

查询处理逻辑电路，可操作用于接收查询，使用一种或多种语言学技术来扩展所述查询，以及响应于所述查询在所述文档数据库中的文档中搜索信息。

23. 根据权利要求 22 所述的系统，其中，所述一种或多种语言学技术包括一个或多个复合项扩展、字尾变化形式集合扩展、或正字法扩展。

提高搜索质量的系统和方法

技术领域

本发明一般涉及信息搜索和检索。更具体地，披露了用于提高搜索质量的系统和方法。

背景技术

在信息检索系统中，用户通常输入查询，然后收到包含查询项的一列文档。不包含查询项的文档被忽略。因此这种系统鼓励正确的查询公式化。

需要用于改善查询的系统和方法，使得它们更多可能地产生有用的搜索结果。

发明内容

本发明提供了用于提高搜索质量的系统和方法。应该明白，本发明可以用很多方式来实现，包括作为过程、设备、系统、装置、方法、或者计算机可读介质，例如计算机可读存储介质或在其上通过光或电的通信线路发送程序指令的计算机网络。下面描述本发明的几个具体实施例。

在一个实施例中，一种方法，总体上可包括：接收包括至少一个查询项的查询；确定查询是否包括复合查询项、包括在一组字尾变化形式中的查询项、和/或包括在一组可选拼写中的查询项，如果是，则自动扩展查询，以包括复合查询项的可选表示、来自该组字

尾变化形式的相应字尾变化形式、和/或来自该组可选拼写的相应可选拼写；使用扩展的查询来搜索数据库；以及返回结果给用户。

在另一个实施例中，一种方法，总体上可包括：识别（identify，标识）一组与文档相关的项；通过进一步将文档与一个或多个可选拼写、该组项中至少一项的附加字尾变化形式、和/或该组项中至少一个复合项的一个或多个可选表示相联合，来扩展该组项；以及使用该扩展组项来对文档建索引。

在另一个实施例中，一种方法，总体上包括：用连字符连接字搜索第一组文档；用与连字符连接字相对应的非连字符连接字搜索第一组文档；以及在连字符连接字与相应非连字符连接字之间产生一组联系。在一个实例中，该方法可进一步包括：接收来自用户的包括第一查询项的查询；在连字符连接字与相应非连字符连接字之间的该组联系中定位第一查询项；以及扩展该查询，以包括在连字符连接字与相应非连字符连接字之间的该组联系中与第一查询项相关联的第二查询项。

根据另一个实施例，一种计算机程序包，其驻留于计算机可读介质上，计算机程序包包括指令，当处理器执行指令时，指令使处理器执行如下操作：通过包括至少一个查询项的一个或多个可选拼写来扩展从用户接收的查询；用至少一个复合查询项的一个或多个可选表示来扩展查询；和/或用至少一个查询项的一个或多个字尾变化形式来扩展查询。

根据另一个实施例，一种信息检索系统，总体上包括：文档数据库，文档数据库包括一组文档；以及查询处理逻辑电路，可操作用于接收查询，使用一种或多种语言学技术扩展查询，以及响应于查询在文档数据库中的文档中搜索信息。这些语言学技术可以包括复合项扩展、字尾变化形式集合扩展、和/或正字法扩展。

本发明的这些和其他特征和优点将在后面的详细描述和附图中呈现，其以实例形式阐述本发明的原则。

附图说明

通过以下结合附图的详细说明可以容易地理解本发明，其中，相同的标号表示相似结构的部件。

图 1 是信息检索系统的示意图。

图 2 是用于实施本发明的实施例的示例性计算装置的示意图。

图 3 示出了可对其执行搜索的一组文档。

图 4 示出了图 3 中所示的文档的索引。

图 5 是用于搜索例如图 3 中所示的一组文档的方法的流程图。

图 6A 示出了用于产生一列复合字（compound word）的方法。

图 6B 是使用一列复合字搜索一组文档的方法的流程图。

图 7A 示出了用于产生关于一组字（word）的字尾变化（inflection）集合的方法。

图 7B 是使用字尾变化信息搜索一组文档的方法的流程图。

图 8 是使用正字法信息搜索一组文档的方法的流程图。

图 9 是使用一种或多种语言学技术扩展搜索查询来搜索一组文档的方法的流程图。

图 10 是图 3 所示的文档的扩展索引。

图 11 是使用诸如图 10 中所示的索引来搜索一组文档的方法的流程图。

具体实施方式

披露了用于提高搜索质量的系统和方法。给出下面的描述，使本领域任何技术人员都能够制造和使用本发明。提供的具体实施例和应用的描述仅作为示例，对于本领域技术人员来说，显然很容易做出各种修改。例如，虽然是以德语搜索引擎的上下文环境列举了多个实例，但应该明白，在不脱离本发明的精神和范围的前提下，此处所描述的一般原则可以应用到其他语言、实施例、和应用中。类似地，尽管下面给出的许多例子描述为使用互联网网页作为要搜索的文档，但应该明白，脱机文档，例如，书、报纸、杂志、或其他扫描成电子格式的纸质文档，同样可以被搜索。因此，给予本发明最大范围，包含与本文所披露的原则和特征相一致的各种可选物、修改、和同等物。为了清楚起见，没有详细描述涉及本发明的领域内所公知的技术资料的相关细节，以避免使得本发明不必要地不清楚。

在信息检索系统中，用户通常通过检索接口输入查询，以找到相应文档。返回的结果通常只限于以某种方式匹配该查询的那些文档。系统和方法描述为通过一种或多种语言学技术的应用来扩展用户查询。在一个实施例中，使用复合字、字尾变化形式（inflectional form）、和/或正字法变化（orthographic variation）的数据库来扩展用户的原始查询。扩展后的查询随后被用来执行搜索相应文档。

图 1 示出了系统 **100**，其中，可以实施符合本发明的方法和装置。系统 **100** 可以包括多个客户设备 **102**，其通过网络 **106** 连接到多个服务器 **104**、**105**。客户设备 **102** 可以包括浏览器 **110**，用于接收用户输入，并用于显示通过网络 **106** 从其他系统 **102**、**104**、**105**

接收的信息。服务器**104**、**105**可以包括搜索引擎**112**，用于接收通过网络**106**传送的用户查询，搜索文档数据库，并将结果返回给用户。网络**106**可以包括局域网（LAN）、广域网（WAN）、虚拟专用网络（VPN）、电话网，诸如公共电话交换网（PSTN），内联网，互联网，或多种网络的组合。为了方便图示，图1示出了连接到网络**106**的三个客户设备**102**和两个服务器**104**、**105**；然而，应该明白，实际当中，可以有更多或更少的客户设备、服务器、和/或网络，并且一些客户设备也可以执行服务器的功能，一些服务器可以执行客户端的功能。

图2示出了更详细的系统**200**实例，诸如图1中所示的客户端**102**或服务器**104**、**105**。在一个实施例中，系统**200**包括计算装置，诸如个人计算机、便携式电脑、大型机、个人数字助理、移动电话、和/或相似的设备。系统**200**通常将包括处理器**202**、存储器**204**、用户接口**206**、用于接受可移动存储介质**208**的输入/输出端口**207**、网络接口**210**、以及连接上述元件的总线**212**。

系统**200**的操作将通常由处理器**202**在存储于存储器**204**中的程序指导下操作所控制。存储器**204**将通常包括计算机可读介质的一些组合，诸如高速随机存取存储器（RAM）和非易失性存储器（诸如只读存储器（ROM））、磁盘、磁盘阵列、和/或磁带阵列。端口**207**可以包括用于接受例如软盘、CD-ROM、DVD、存储卡、磁带等计算机可读介质的磁盘驱动器或存储器插槽。例如，用户接口**206**可以包括用于输入信息的键盘、鼠标、笔、或语音识别装置，以及一个或多个用于向用户呈现信息诸如显示器、打印机、扬声器、和/或类似机构。网络接口**210**通常可操作用于通过有线、无线、光的、和/或其他连接在系统**200**与其他系统（和/或网络**220**）之间提供连接。

下面将更详细地描述，系统**200**可以执行各种搜索和检索操作。这些操作将通常响应于处理器**202**执行计算机可读介质（例如存储器**204**）中所包含的软件指令而被执行。软件指令可以从另一计算机可读介质(例如数据存储装置**208**),或者通过通信接口**210**或I/O端口**207**从另一装置读取到存储器**204**中。如图2所示，存储器**204**可以包括各种程序或模块，用于控制系统**200**的操作和执行下面将更详细描述的搜索和检索技术。例如，如果系统**200**是服务器（例如图1中所示的服务器**105**），则存储器**204**可以包括文档数据库**229**和相应索引。存储器**204**还可以包括：搜索引擎**230**，用于使用从用户接口**206**接收到的和/或通过网络**220**远程地从用户接收到的查询来搜索数据库**229**。如图2所示，存储器**204**还可以包括一个或多个程序，用于使用下面更详细描述的技术来扩展查询和/或文档；以及用户接口应用程序**232**，用于操作用户接口**206**和/或用于通过网络**220**提供用户界面网页给远程用户。尽管图2示出了一个主要是基于软件的系统，但应该明白，在其他实施例中，可以使用专用电路替代软件指令或与之结合使用来执行与本发明一致的处理。因此，本发明不限制于任何特定的硬件和软件结合。

应该明白，本发明的系统和方法可以用缺少如图1和图2中所示的一些部件的和/或具有未示出的其他部件的设备和/或结构来实现。因此，应该明白，图1和图2是用作说明的目的，不是对本发明的范围的限制。例如，应该明白，为了图示的目的，系统**200**被描述为单一的、通用的计算装置，例如个人计算机或网络服务器，在其他实施例中，系统**200**可以包括一个或多个采用分布式计算技术一起操作的这种系统。在这种实施例中，图2中描述的一些或全部组件和功能可以扩展到多地点的和/或多方操作的多个系统中。例如，查询扩展应用程序**231**可以在与其上装有数据库**229**的系统分离的系统上实现（例如，一些实施例中，查询扩展可以在客户端上

而不是在服务器上执行)。很明显，在不背离本发明的原则的前提下，可以对图 1 和图 2 中所示说明做出很多相似的变化。

如之前指出的，图 1 和图 2 所示的系统可以用来响应于用户的查询帮助检索文档（例如，网页）。图 3 示出了一组德语文档 **302**、**304**、**306**、**308**，可以对其执行这种搜索。例如，文档 **302**、**304**、**306**、**308** 可以被存储在一个或多个如图 1 所示的服务器 **104**、**105** 上。如图 3 所示，第一文档 **302** 包含字（word）“abendzeitung”、“autotelefon”、“abirrungen”、和 “betttuch”。第二文档 **304** 包含字 “abend-zeitung”、“abirrung”、“autotelephon”、和 “abisolieren”。第三文档 **306** 包括字 “bettuch”、“bahnwagon”、“abisolierten”、和 “abendzeitung”。以及第四文档 **308** 包含字 “autotelefon”、“bahnwaggon”、“abisolierte”、和 “abirrung”。文档 **302**、**304**、**306**、**308** 还可以包括一个或多个到其他文档的链接（或引用）**310**。尽管为了方便图示，图 3 示出用德文书写的文档，但应该明白，这些文档可以是用任何一种语言或多种语言的组合来书写的。

图 4 示出了基于图 3 所示的文档的索引 **400**。索引的第一列包括一列项目（term），第二列包含一列与这些项目相对应的文档。一些项目（例如 “bahnwaggon”）只对应于（例如，出现在）一个文档（即，文档 **308**）。其他项目（诸如 “autotelefon”）对应于多个文档（即，文档 **302** 和 **308**）。

图 5 示出了过程 **500**，搜索引擎（例如图 1 所示的搜索引擎 **112**）通过该过程可以利用图 4 中示出的索引 **400**，响应于查询来提供搜索结果。搜索引擎 **112** 接收查询（方框 **502**），然后使用索引（例如索引 **400**）来确定哪些文档对应于该查询（方框 **504**）。例如，可以使用布尔逻辑来使查询与文档匹配，或可以使用基于信息检索积分的项目频率反演文档频率（term frequency-inverse document frequency，缩写为 tf-idf），使查询中的字与每个文档中的字结合。

因此，例如，如果查询是“abendzeitung”，搜索引擎 **112** 可以使用索引 **400** 来确定“abendzeitung”出现在文档 **302** 和 **306** 中。然后返回这些文档，和/或这些文档的引用给用户（方框 **506**）。

如在前述例子中看到的，搜索可能不能够识别出不包括精确查询项的文档。例如，结合图 5 描述的例子中，查询“abendzeitung”不能够定位包含项目“abend-zeitung”的文档 **304**。

改善查询结果的一条途径是扩展查询，使其包括查询项可能的变化，从而确保包含这些变化的相应文档不被漏掉。在优选实施例中，使用诸如复合字、字尾变化、和正字法（例如，拼写）变化的各种语言特征来实现这个目的。

复合字

在许多语言中，某些字对（word pair）可以分开来书写、复合书写、或者以连字符连接书写。例如，在德语中，可以将许多名词连接起来形成较长的名词复合字。在许多情形中，没有书写这些字的标准方式（例如，连接的，连字符连接的，或者分开的），因此不同的形式可以用在不同的文档中。例如，项目“frensehprogramm”（意思是电视节目）既可以写成“frensehprogramm”也可以写成“frenseh-programm”。因此，使用这个字的一种形式而不是另一种形式的查询可能导致不能定位相应文档。

在一个实施例中，可以通过建立一个可能的复合字的表，然后利用该表来扩展查询，使之包含该表中的一个或多个复合字来解决或改善这个问题。可以通过各种方式建立字对（或三字词组、等等）表。例如，可以使用字典，或通过对文档文集（例如，因特网网页）动态搜索然后生成复合项的表，来形成该表。

图 6A 示出了这种方法 **600** 的实例。如图 6A 所示，通过在文档集合中搜索连字符连接字，生成可能的字对的表（方框 **602**），然后在文档中搜索每个字相应的非连字符连接的形式（方框 **604**）。然后可以生成识别出的每个字对（例如，“AB”或“A-B”）的表（方框 **606**）。在一些实施例中，接着可以通过例如去除文档集合中以较低频率出现的字对来缩短所得的表（方框 **608**）。例如，可以检查“AB”在文集中出现的次数，“A-B”出现的次数等。应该明白，可以对图 6A 中所示的基本过程做出多种改变。例如，在一些实施例中，可以在文档集合中搜索其中“复合”字表现为单独的、非连字符连接的字的字对（或三字词组，等等）（例如，“AB”）的实例。

如图 6B 所示，所得的复合字表接着可以用来扩展包含该表中一个或多个字的查询。例如，当接收到查询时（方框 **652**），可以检查该查询，以确定是否该查询包含字对表中的任何字。如果该查询包含为复合字对中的一部分的字，则可以补充该查询使之包含字对中的其他部分（方框 **654**）。例如，这个字可以由该字的两种形式的逻辑和（disjunction，“或”）替代。例如，“AB”可以由“AB OR A-B”替代；“A-B”可以由“A-B OR AB”替代；等等。因此，例如，前面结合图 5 讨论的查询“abendzeitung”，可以扩展为“abendzeitung OR abend-zeitung”，并且在与索引相比较时，就会得出文档 **302**、**304**、和 **306**（而不只是文档 **302** 和 **306**）。

在一些实施例中，上述复合字表还可以用来在其他方面改善搜索结果。例如，用诸如 Postscript（PS）或 Adobe’s Document Format（PDF）的格式书写的文档通常包含用于在行尾断字的连字符。这些字可能被当作连字符连接字而被不适当地索引。因此，在一个实施例中，在文档建索引（或语法分析）时可以使用上述复合字表。当遇到连字符连接字时，使其与复合字表进行比较，并且如果没有被定位，则当这个字被建索引时可以去掉连字符。

字尾变化

类似地，许多字具有多种字尾变化形式来表达语法上的联系，诸如格、性、数、人称、时态、或语气。英语字尾变化的例子包括在名词字尾加“s”构成复数，或在动词字尾加“ed”表示过去时态。其他字尾变化包括改变基本字本身，例如字尾变化集合“speak”、“spoke”、“spoken”。

德语同样有很多种字尾变化形式。例如，“abirrung”和“abirrungen”是相同字根的不同字尾变化形式，像“spiel”、“spiele”、“spielen”、“spieles”、和“spiels”。因此，采用一种字尾变化形式而不是其他形式的查询可能会使建立查询的用户找不到感兴趣的文档。

因此，在一个实施例中，汇集了多组字尾变化形式，然后用来扩展查询。可以通过各种方式获得字尾变化集合，例如通过查字典或通过使用自动工具。例如，如果德语是查询语言，则可以使用具有字根形式的较大字典的语言分析或产生工具生成字尾变化集合，例如使用任何合适的字形分析器。

如图 7A 所示，在一个实施例中，可以通过从文档文集（例如，网页）中收集一组字来产生一组字尾变化形式（方框 **702**）。然后可对该组字应用字形分析器，产生在字尾变化字与字根之间的一组映射（方框 **704**）。在一些实施例中，可以通过使用在文档中只出现适当的次数或百分比的那些字（例如，在至少 100 个文档中出现的那些字）来过滤该组映射（方框 **706**）。然后可以反转（invert）该表，形成在字根与字尾变化形式之间的一组映射（方框 **708**）。

图 7B 示出了使用诸如图 7A 中所示的方法建立的字尾变化集合实现查询扩展的方法。如图 7B 所示，如果查询包含属于字尾变

化集合中的字（方框 **752**），则通过将字尾变化集合（或某个合适的子集）中的所有成分的逻辑和包含在内来扩大该查询（方框 **754**）。例如，查询“auto spiel”可以变为“(auto OR autos)(spiel OR spiele OR spiel OR spiele OR spielen OR spieles OR spiels)”。然后扩展后的查询被用来搜索文档数据库（例如，通过将该搜索与数据库索引进行比较）（方框 **756**），并且将搜索结果呈现给用户（方框 **758**）。因此，例如，如果用户提交包括字“abisolieren”的查询，那么可以将它扩展为“abisolieren OR abisolierten OR abisolierte”，因此能够使图 3 中所示的文档的搜索识别出文档 **306** 和 **308** 以及文档 **304**。

应该明白，可以对在图 7A 和图 7B 中示出的基本概念作出多种变化。例如，可以将查询项的字根形式的其他变化包括在该扩展中，不管这些变化严格来说是否是查询项的字尾变化。作为另一示例，在一些实施例中，用来执行查询扩展的字尾变化集合可以通过查阅字典或其他资源而不是以结合图 7A 描述的方式应用字形分析器的方法来建立。

正字法变化

许多语言包括大量可以用不同方法拼写的字。例如，许多德语字由于方言变化和/或近代拼写变革而有不同的拼写。常见的德语拼写变化的例子包括“ph”和“f”（例如，“telefon”或“telephon”），“ß”和“ss”（例如，“maße”或“masse”）的可互换性，各种重复字母顺序的可互换性（例如，“wagon”或“waggon”、“bettuch”或“betttucn”、等等），以及撇号的使用（例如，“kantsch”或“kant'sch”）。

因此，在一个实施例中，产生了正字法变化表。举例来说，这可以通过查阅字典或其他资源来完成。例如，许多德语拼写中的变化可以通过检查关于德语拼写变革（例如，使用任何合适的字形分

析器）的数据等而得到。举例来说，由 Institut fuer Deutsche Sprache（德语语言学院）（发表了关于德语的大量信息的基金会）在 http://www.ids-mannheim.de/org/上提供关于德语拼写变革的信息。如图 8 所示，该表格可以用来扩展用户查询（方框 **802-804**），然后扩展后的查询可以用来搜索相应的文档（方框 **806-808**）。

因此，描述了多种改进搜索结果的技术。应该明白，这些技术可以单独应用，或彼此相互配合和/或与其他技术结合来使用。图 9 示出了应用诸如前述那些语言学技术来执行对文档的索引或数据库的搜索的一般过程。如图 9 所示，当从用户接收到查询时（方框 **902**），通过应用前述一种或几种技术将查询扩展（方框 **904**）。然后，将扩展后的查询与数据库索引相比较，以定位相应的文档（方框 **906**），然后将相应的文档返回或识别给用户（方框 **908**）。

应该明白，根据本发明的实施例可以对上述系统和方法做出多种改变。例如，上述技术可以与例如拼写校正、同义字和/或相关字扩展、语言翻译、非索要信息（spam）缩减、和/或相似的其他技术结合起来应用，以进一步改善搜索结果。作为另一例子，在一些实施例中，可以响应于用户的查询执行多个搜索。例如，可以首先使用用户的原始查询执行搜索，随后使用该查询扩展后或重写的版本执行一个或多个搜索。可以对这些搜索结果作评价（例如，使用关于用户的喜好和搜索历史的信息），然后可以返回确定最可能有用的结果。例如，如果扩展后的查询的那些结果被确定具有较高或相当的质量，那么可以用它们对原始查询的最高质量的结果进行补充。可选地，或另外地，扩展查询中的项目可以被不同地加权。例如，可以给予原始查询项较高的加权，而给予通过扩展附加的项目较轻的加权。

另外，尽管上述示例涉及用户查询的扩展，但在其他实施例中，也可以代替（或附加）扩展文档索引本身。图 10 示出了如图 3 所

示的文档的扩展索引的示例。如图 10 所示，不同的复合项、字尾变化集合、和正字法变化在索引的左侧栏中被一起分组，并且包含组中任何项目的文档在右侧栏中列出。如图 11 所示，一旦产生扩展的索引（方框 **1102**），则可以不执行查询扩展就将用户查询（方框 **1104**）直接与索引（方框 **1106**）相比较。可选地，可以使用索引扩展和查询扩展的一些结合。

此外，虽然以上提供的示例是应用在德语环境中，但应该明白，所述技术也可以很容易地应用到其他语言中。每一种语言都有自己形成搜索问题的语言特征。因此，为了设计针对给定语言地搜索引擎、和/或通用搜索引擎，可以努力识别出这些问题并解决它们。例如，可以执行随机搜索来查看是什么搜索项引起问题。随后可以改变这些搜索项来查看是否有了改善。也可以分析用户对话来发现用户搜索行为的方式。例如，用户可能使用某些变化来补偿语言上有问题的方面。一旦识别出一组问题区域，就可以做工作来找出解决方案。通过对可能的解决办案进行测试或仿真来确定它们的有效性和实现它们所需要付出的工作量。

虽然在这里描述和说明了本发明的优选实施例，但应该明白，它们仅仅是说明性的，并且在不背离本发明的精神和范围的前提下，可以对这些实施例进行修改。因此，仅根据权利要求来对本发明进行限定。

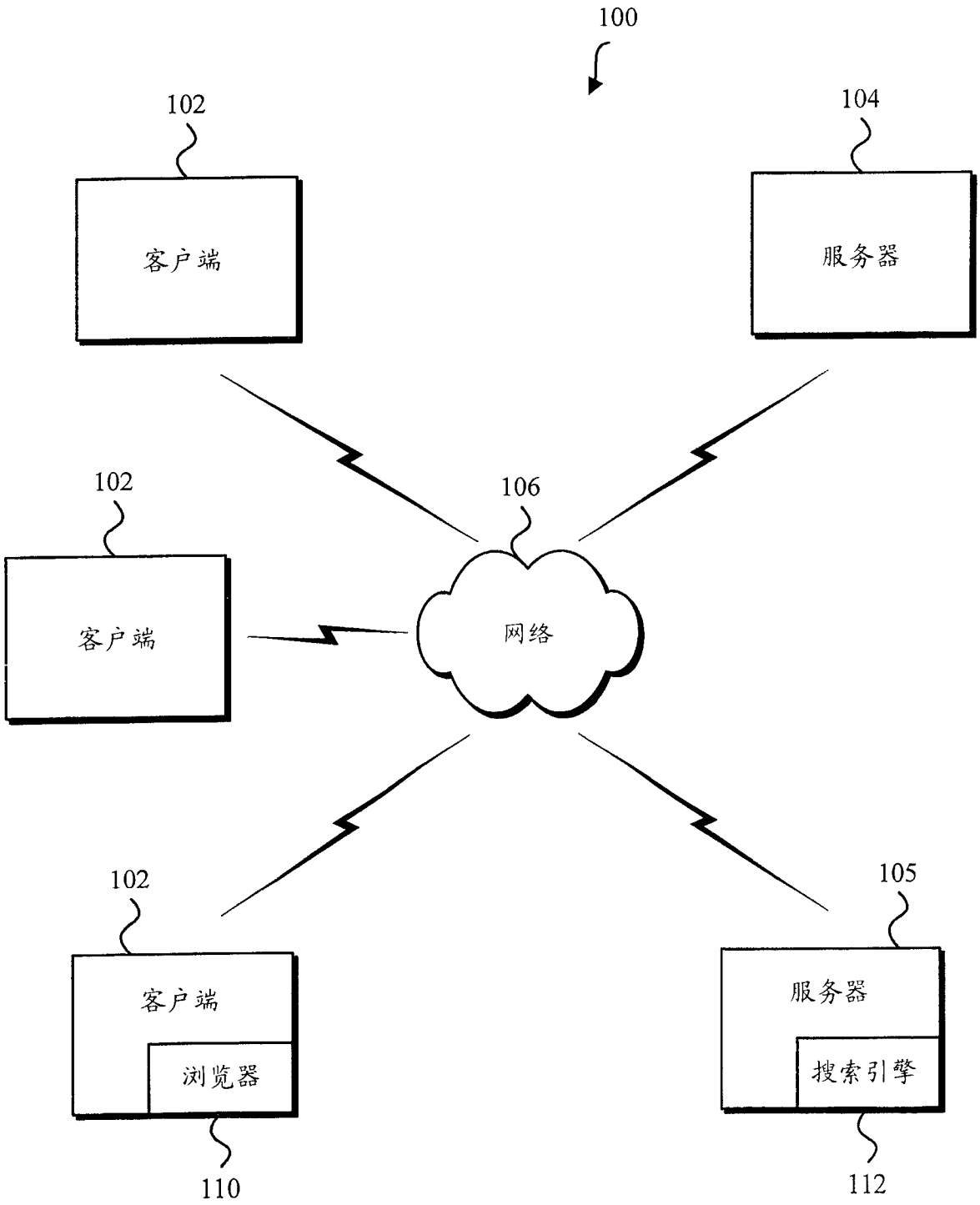

100
102
客户端
104
服务器
102
客户端
106
网络
102
客户端
浏览器
110
105
服务器
搜索引擎
112

图 1

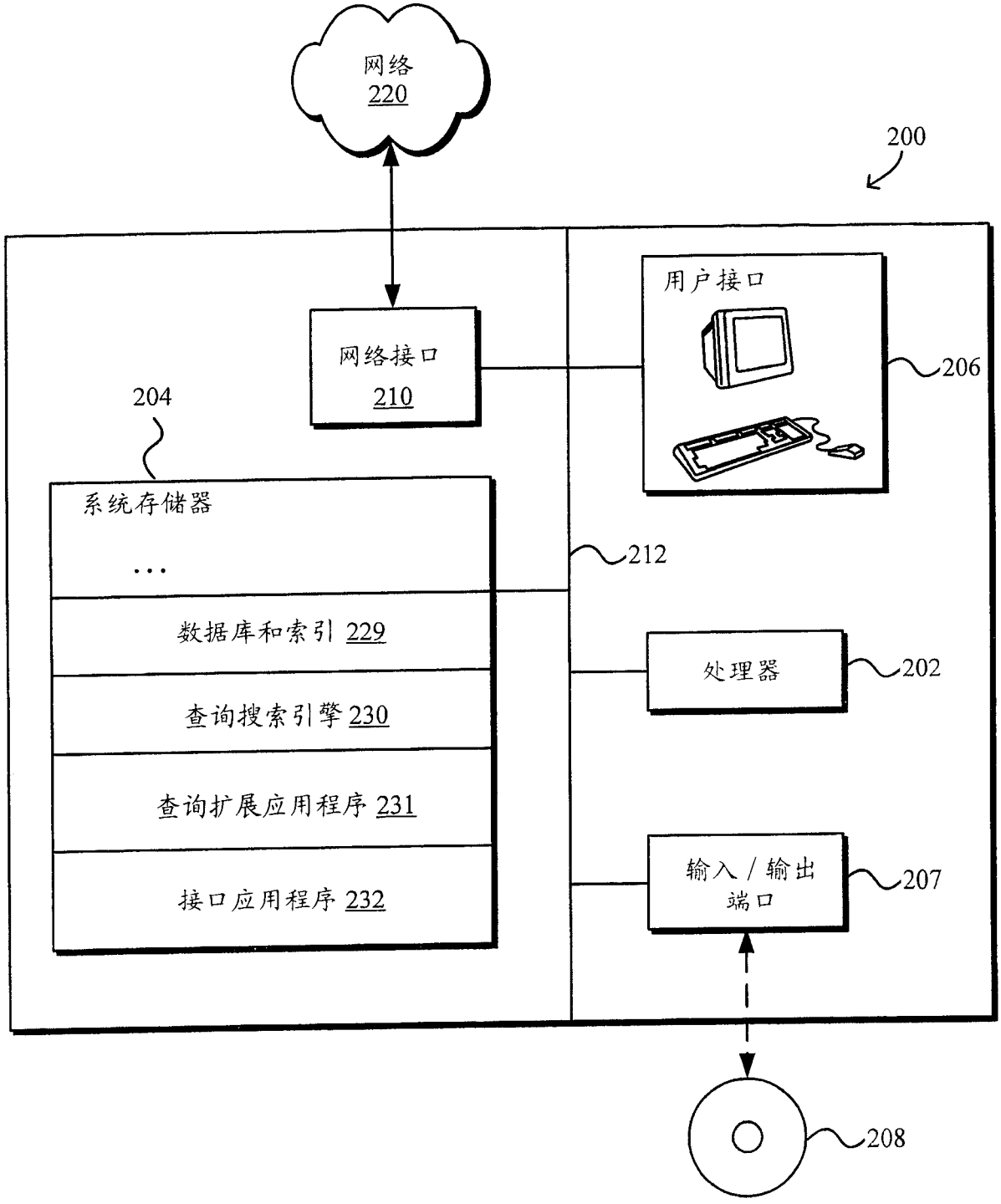

网络
220
200
用户接口
网络接口
210
206
204
系统存储器
...
212
数据库和索引 229
处理器
202
查询搜索引擎 230
查询扩展应用程序 231
输入/输出
端口
207
接口应用程序 232
208

图 2

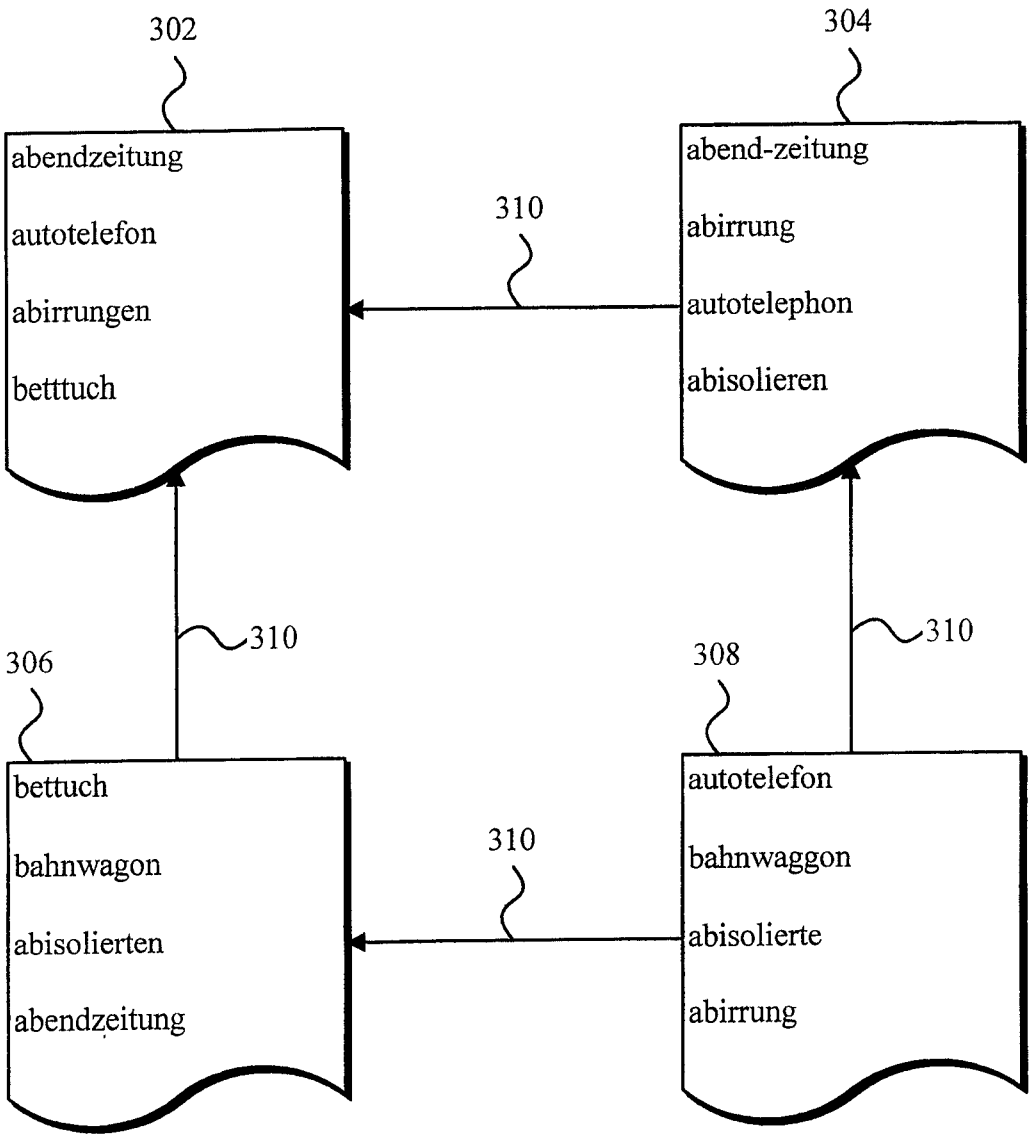

302
304
abendzeitung
autotelefon
abirrungen
betttuch
310
abend-zeitung
abirrung
autotelephon
abisolieren
310
310
306
308
bettuch
bahnwagon
abisolierten
abendzeitung
310
autotelefon
bahnwaggon
abisolierte
abirrung

图 3

400

项目	位置
abendzeitung	文档302 和 306
abend-zeitung	文档304
abirrung	文档304 和308
abirrungen	文档302
abisolieren	文档304
abisolierte	文档308
abisolierten	文档306
autotelefon	文档302 和308
autotelephon	文档304
bahnwaggon	文档308
bahnwagon	文档306
betttuch	文档302
bettuch	文档306
...	...

图 4

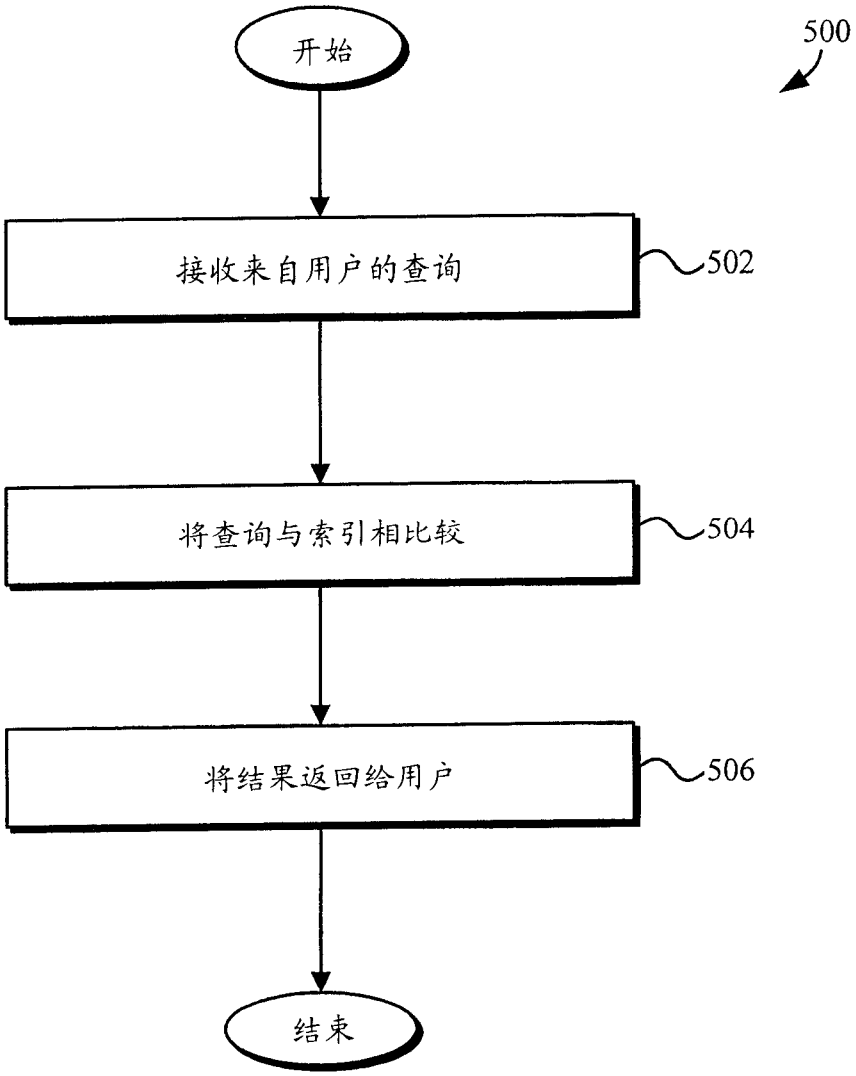


图 5

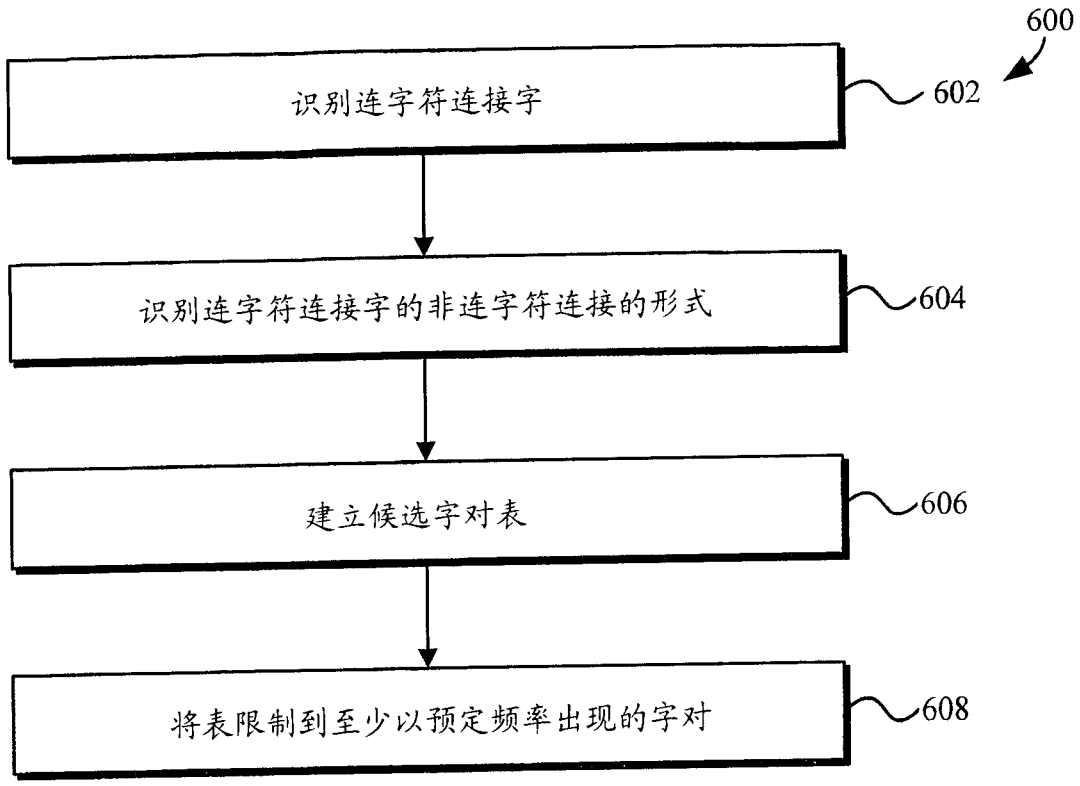

600
识别连字符连接字
602
识别连字符连接字的非连字符连接的形式
604
建立候选字对表
606
将表限制到至少以预定频率出现的字对
608

图 6A

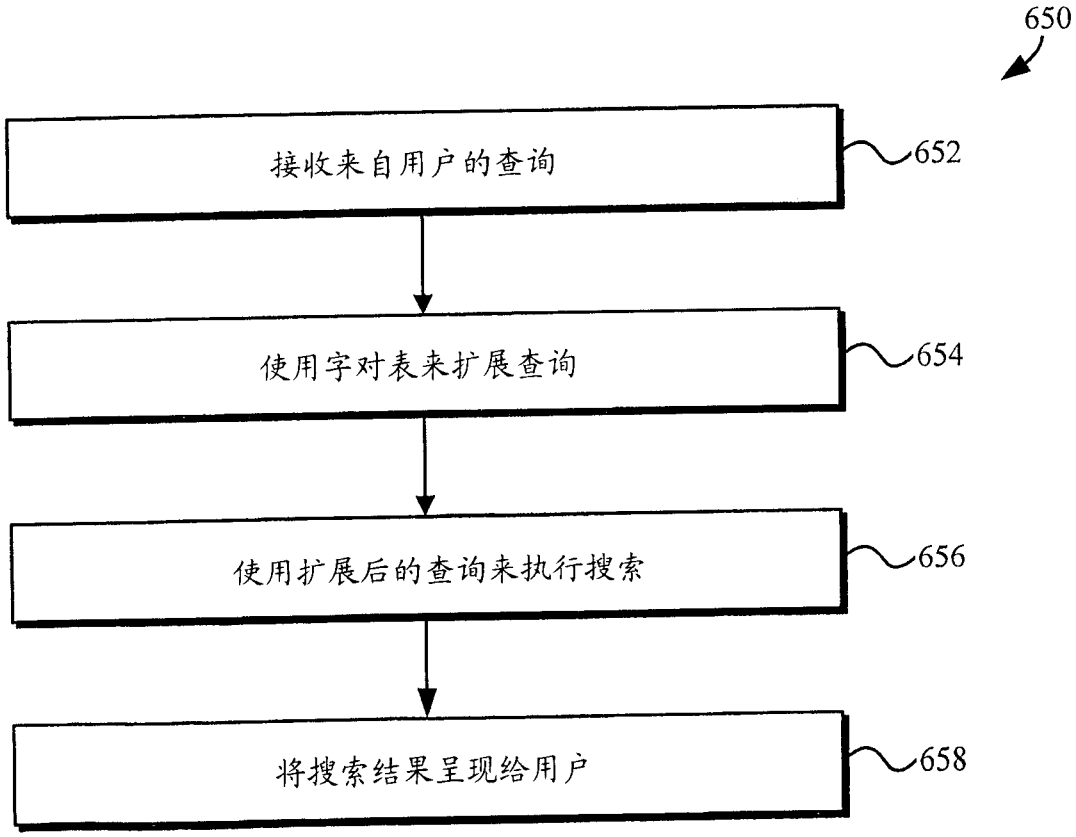

650
接收来自用户的查询
652
使用字对表来扩展查询
654
使用扩展后的查询来执行搜索
656
将搜索结果呈现给用户
658

图 6B

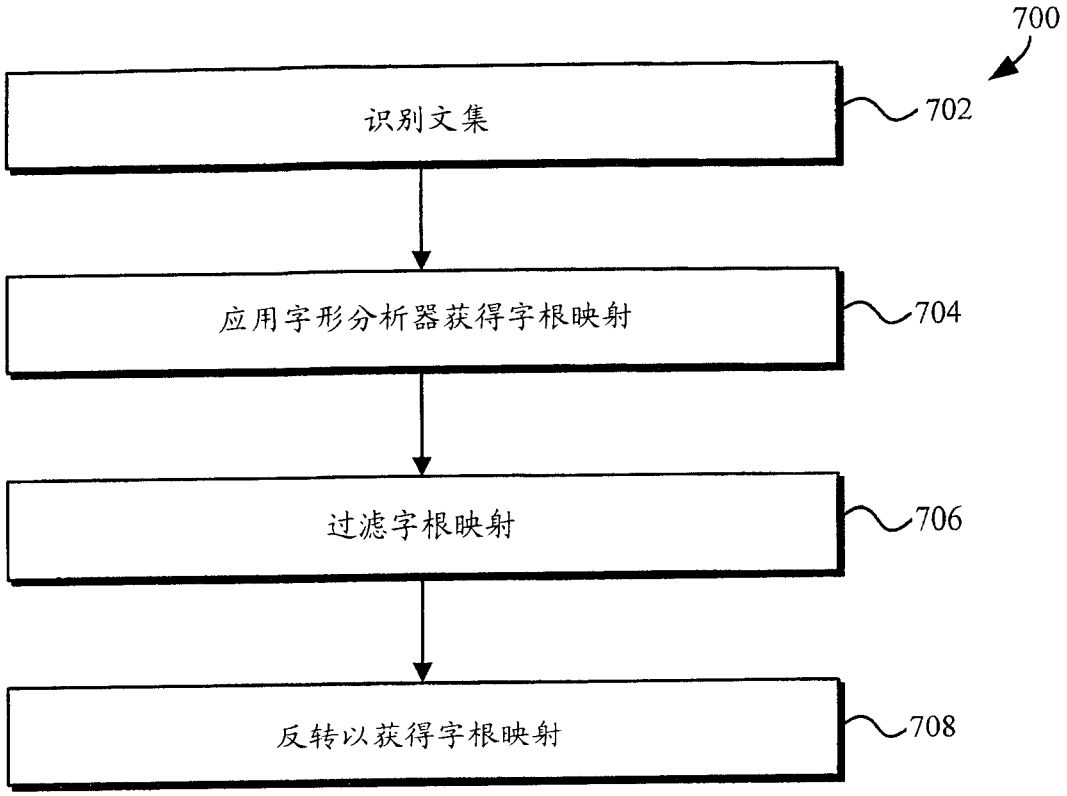

700
识别文集
702
应用字形分析器获得字根映射
704
过滤字根映射
706
反转以获得字根映射
708

图 7A

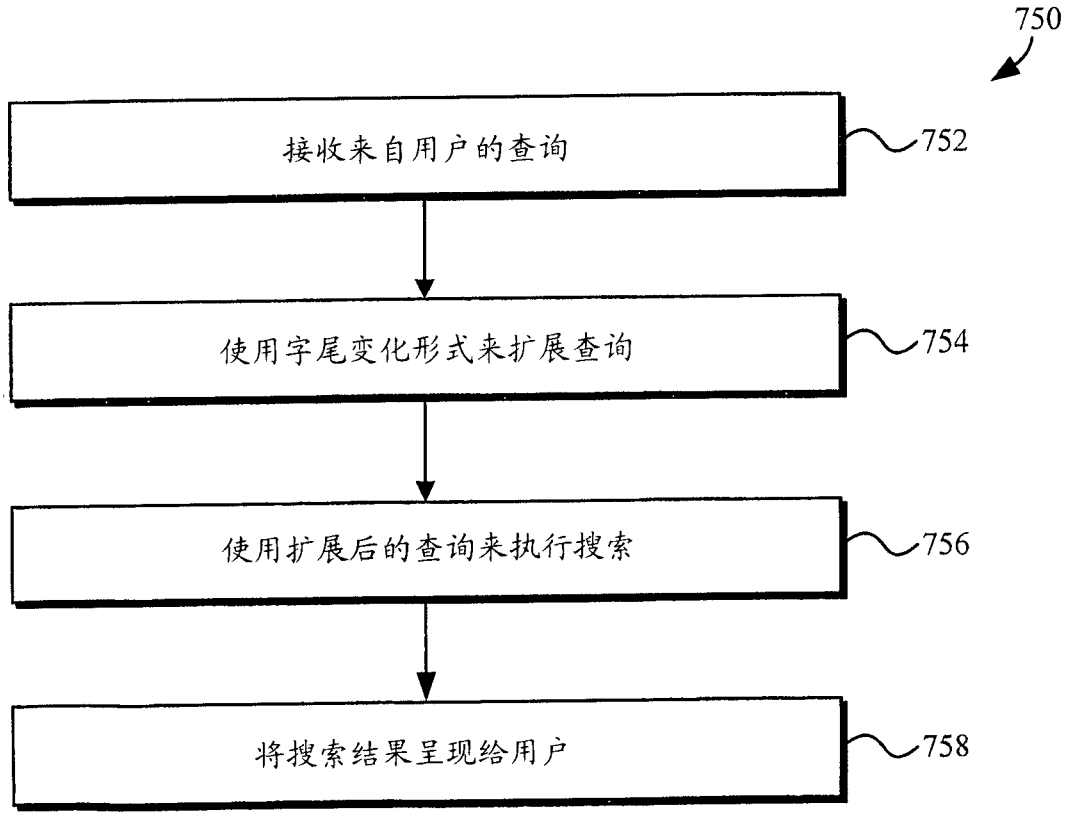

750
接收来自用户的查询
752
使用字尾变化形式来扩展查询
754
使用扩展后的查询来执行搜索
756
将搜索结果呈现给用户
758

图 7B

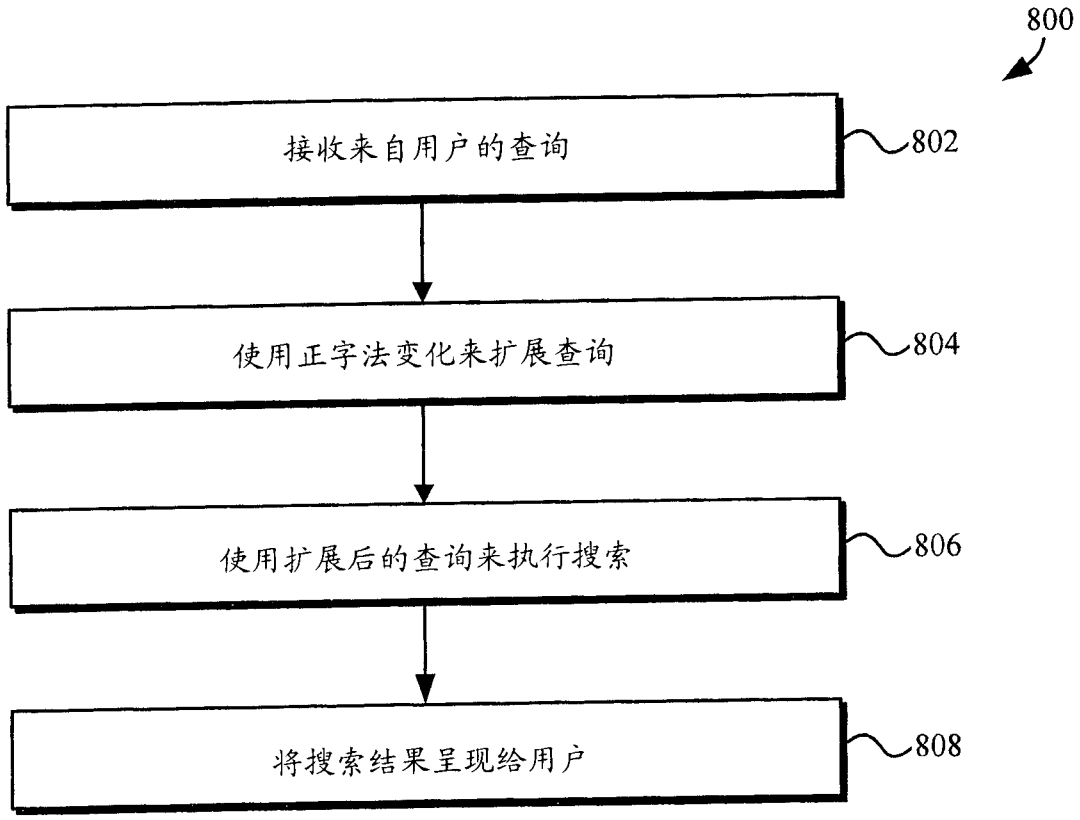

800
接收来自用户的查询
802
使用正字法变化来扩展查询
804
使用扩展后的查询来执行搜索
806
将搜索结果呈现给用户
808

图 8

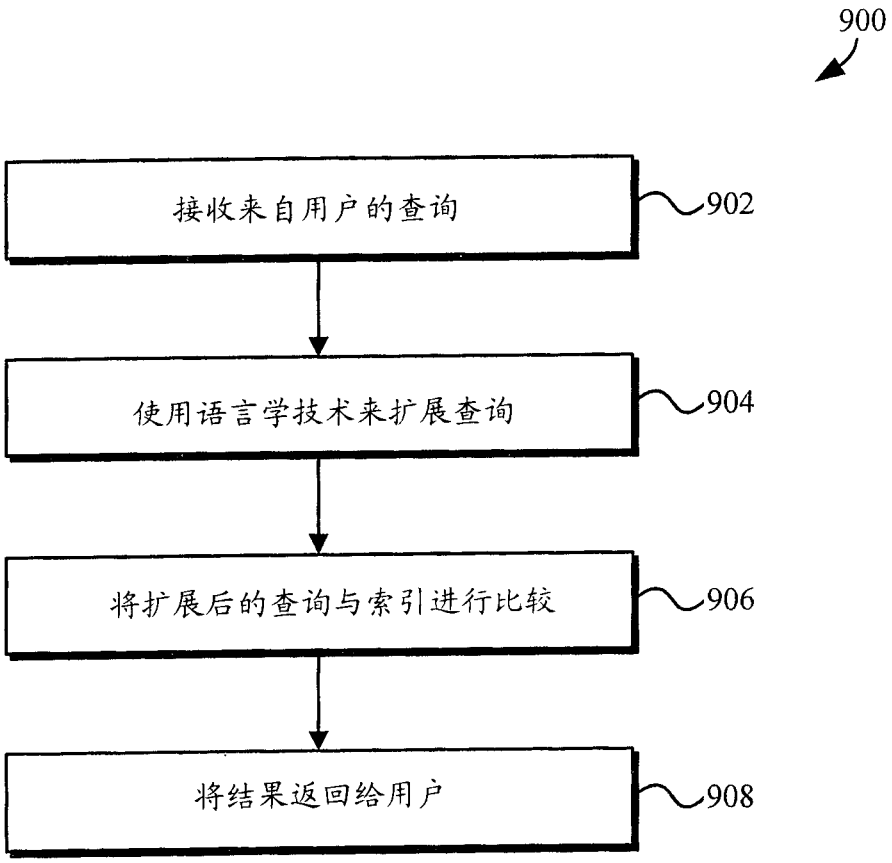

900
接收来自用户的查询
902
使用语言学技术来扩展查询
904
将扩展后的查询与索引进行比较
906
将结果返回给用户
908

图 9

abendzeitung / abend-zeitung	文档 302, 304, 和 306
abirrung / abirrungen	文档 302, 304, 和 308
abisolieren / abisolierte / abisolierten	文档 304, 306, 和 308
autotelefon / autotelephon	文档 302, 304, 和 308
bahnwaggon / bahnwagon	文档 306 和 308
betttuch / bettuch	文档 302 和 306
. . .	. . .

图 10

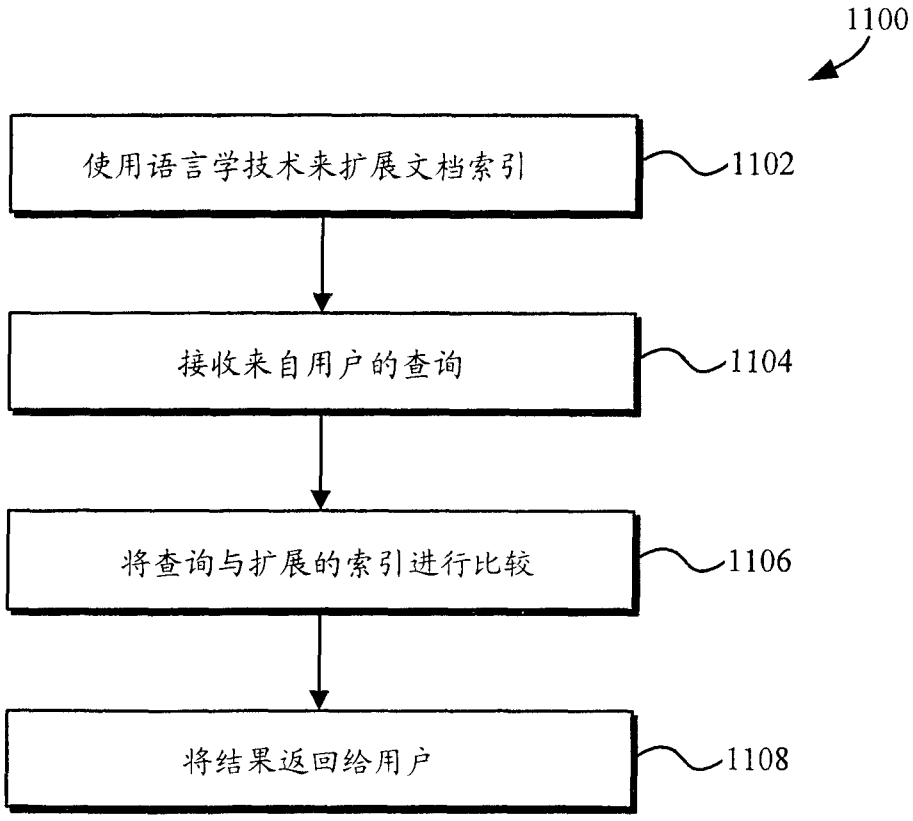


图 11