(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 102725759 A
(43)申请公布日 2012.10.10

(21)申请号 201180008423.2
(22)申请日 2011.01.19
(30)优先权数据
12/701338 2010.02.05 US
(85)PCT申请进入国家阶段日
2012.08.03
(86)PCT申请的申请数据
PCT/US2011/021596 2011.01.19
(87)PCT申请的公布数据
WO2011/097066 EN 2011.08.11
(71)申请人 微软公司
地址 美国华盛顿州
(72)发明人 V. 瓦拉马尼 A. 斯里瓦斯塔瓦 T. 纳姆 A.C. 苏伦德兰
(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 董宁 汪扬
(51)Int.Cl.
G06F 17/30(2006.01)

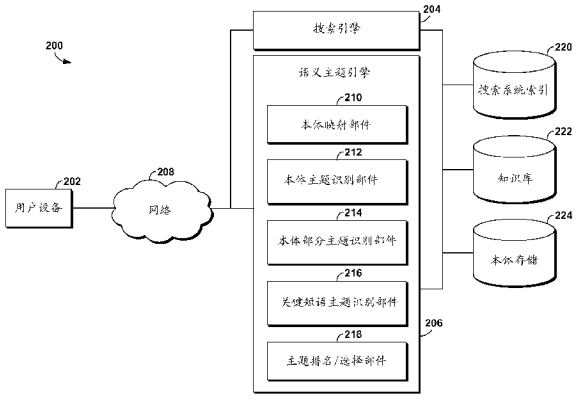
权利要求书 3 页 说明书 9 页 附图 8 页

(54)发明名称

用于搜索结果的语义目录

(57)摘要

通过将语义概念识别为主题以包括在目录中为响应于搜索查询的搜索结果生成目录。当接收到搜索查询时，执行搜索以识别搜索结果。将搜索结果与主题的本体进行比较以识别相关的主题。另外，将搜索结果与部分主题的本体进行比较以识别被命名的相关部分主题。进一步根据搜索结果生成独立关键短语，并将独立关键短语识别为关键短语主题。对识别出的主题、被命名的部分主题以及关键短语主题进行排名，并且选择主题以包括在目录中。响应于搜索查询，返回包含搜索结果和生成的目录的搜索结果页面。



CN 102725759 A

1. 一种或多种存储计算机可用指令的计算机可读的媒体，当一个或多个计算设备使用所述指令时，使所述一个或多个计算设备执行方法，该方法包括：

接收搜索查询；

使用所述搜索查询进行搜索；

从搜索中接收多个文档片段；

通过将一个或多个文档片段与主题的本体进行比较，识别一个或多个候选主题的第一集合；

通过将一个或多个文档片段与部分主题的本体进行比较，识别一个或多个候选主题的第二集合；

通过从一个或多个文档片段生成关键短语主题，识别一个或多个候选主题的第三集合；

对来自所述第一、第二、第三候选主题集合中的候选主题进行排名；

基于对候选主题的排名选择一个或多个主题；

提供搜索结果页面以响应于所述搜索查询，该搜索结果页面具有包含所述一个或多个主题的目录，以及用于呈现一个或多个搜索结果的搜索结果区域。

2. 如权利要求 1 所述的一种或多种计算机可读的媒体，其中通过将一个或多个文档片段与部分主题的本体进行比较来识别一个或多个候选主题的所述第二集合包括识别一个或多个部分主题并命名每个部分主题，其中所述部分主题通过以下操作命名：

在一个或多个文档片段中识别所述部分主题的部分主题标识符单词的出现；

提取所述一个或多个文档片段中围绕所识别的所述部分主题标识符单词的出现而出现的单词和 / 或短语；

对每个提取的单词和 / 或短语的频率进行计数；

选择频率最高的单词或短语；以及

使用所述部分主题标识符和所述使用频率最高的单词或短语对所述部分主题进行命名。

3. 如权利要求 1 所述的一种或多种计算机可读的媒体，其中通过从一个或多个文档片段生成关键短语主题来识别一个或多个候选主题的所述第三集合包括通过以下操作从文档集合中剩余的文档片段计算独立关键短语：

从一个或多个文档片段生成候选关键短语；

评估候选关键短语的独立性；

合并相互依赖的候选关键短语；以及

为每组合并的相互依赖的关键短语识别频率最高的候选关键短语。

4. 如权利要求 1 所述的一种或多种计算机可读的媒体，其中对候选主题排名基于选自如下中的一项或多项：分配给候选主题的文档总数、分配给候选主题的每个文档的排名、分配给主题的经过选择的文档的排名以及候选主题的单词长度。

5. 一种或多种存储计算机可用指令的计算机可读的媒体，当一个或多个计算设备使用这些指令时，使得所述一个或多个计算设备执行一种方法，该方法包括：

接收搜索查询；

确定所述搜索查询的本体映射是否存在；

如果所述搜索查询的本体映射存在，基于所述本体映射检索第一主题集合，并将所述第一主题集合添加到主题列表中；

使用所述搜索查询进行搜索以获得多个搜索结果，每个搜索结果与文档片段相对应；

接收至少部分文档片段作为文档集合以用于进一步分析；

将文档集合中的每个文档片段与主题的本体进行比较；

对于其中确定肯定的主题识别的每个文档片段，将所述文档片段分配给相应的主题，并从所述文档集合中移除所述文档片段；

将至少一个从主题本体识别出的主题添加到所述主题列表；

将所述文档集合中剩余的每个文档片段与部分主题的本体进行比较；

对于其中确定肯定的部分主题识别的每个文档片段，将所述文档片段分配给相应的部分主题，并从文档集合中移除所述文档片段；

对至少一个具有一个或多个被分配的文档片段的部分主题进行命名；

将至少一个命名的部分主题添加到所述主题列表；

根据所述文档集合中剩余的文档片段计算独立关键短语；

将文档分配给独立关键短语；

识别至少一个独立关键短语主题；

将所述至少一个关键短语主题添加到所述主题列表；

对所述主题列表中的主题进行排名；

基于排名选择主题；

使用所述选择的主题生成目录；以及

提供搜索结果页面以响应于所述搜索查询，所述搜索结果页面包括所述目录以及用于呈现搜索结果的搜索结果区域。

6. 如权利要求 5 所述的一种或多种计算机可读的媒体，其中将所述文档集合中的每个文档片段与主题本体进行比较包括：基于每个文档片段中包含的单词为每个文档片段计算特征向量，并将每个特征向量与所述主题本体中的主题进行比较，

并且其中通过确定文档片段的特征向量在给定主题的预定距离内来确定针对该文档片段的肯定主题识别，

并且其中将至少一个从所述主题本体识别出的主题添加到所述主题列表包括：添加具有大于预定数目的所分配的文档片段的每个主题。

7. 如权利要求 5 所述的一种或多种计算机可读的媒体，其中将所述文档集合中剩余的每个文档片段与所述部分主题的本体进行比较包括：基于每个文档片段中包含的单词为每个文档片段计算特征向量，并将每个特征向量与所述部分主题本体中的部分主题进行比较，

并且其中通过确定文档片段的特征向量在给定部分主题的预定距离内来确定针对该文档片段的肯定部分主题识别。

8. 如权利要求 5 所述的一种或多种计算机可读的媒体，其中对至少一个具有一个或多个所分配的文档片段的部分主题进行命名包括：对具有大于预定数目的所分配的文档片段的每个部分主题进行命名。

9. 如权利要求 5 所述的一种或多种计算机可读的媒体，其中对部分主题命名包括：

在分配给所述部分主题的一个或多个文档片段中识别所述部分主题的部分主题标识符单词的出现；

提取所述一个或多个文档片段中围绕所识别的所述部分主题标识符单词的出现而出现的单词和/或短语；

对每个提取的单词和/或短语的频率进行计数，其中对每个提取的单词和/或短语的频率进行计数包括：跟踪涉及所述部分主题标识符单词的每个提取出的单词和/或短语的位置；并且其中命名所述部分主题包括：基于使用频率最高的单词或短语的位置信息来确定所述部分主题标识符单词和使用频率最高的单词或短语的顺序；

选择使用频率最高的单词或短语；以及

使用所述部分主题标识符和使用频率最高的单词或短语对所述部分主题进行命名。

10. 如权利要求 5 所述的一种或多种计算机可读的媒体，其中从所述文档集合中剩余的文档片段计算独立关键短语包括：

从所述文档集合中剩余的文档片段生成候选关键短语；

评估候选关键短语的独立性；

合并相互依赖的候选关键短语；以及

为每组合并的相互依赖的关键短语识别频率最高的候选关键短语。

11. 如权利要求 5 所述的一种或多种计算机可读的媒体，其中识别至少一个关键短语主题包括：将具有大于预定数目的所分配到的文档片段的每个关键短语识别为关键短语主题。

12. 如权利要求 5 所述的一种或多种计算机可读的媒体，其中对主题排名基于选自如下的一项或多项：分配给主题的文档总数、分配给主题的每个文档的排名、从分配给主题的经过选择的文档的排名以及主题的单词长度。

13. 一种用于从搜索结果集合识别主题以生成搜索结果的目录的方法，该方法包括：

接收搜索查询；

使用所述搜索查询进行搜索；

从搜索中接收多个文档片段；

从所述文档片段的至少部分生成候选关键短语；

评估候选关键短语的独立性；

合并相互依赖的候选关键短语；

为每组相互依赖的关键短语识别频率最高的候选关键短语，以生成多个独立关键短语；

将一个或多个文档片段分配给每个独立关键短语；以及

基于对独立关键短语的文档片段分配选择关键短语主题。

14. 如权利要求 13 所述的方法，其中使用基于马尔可夫链的方法来生成候选关键短语。

15. 如权利要求 13 所述的方法，其中使用选自以下的一项或多项来评估候选关键短语的独立性：候选关键短语共享的单词数目、对候选关键短语中单词的首字母缩写词的分析以及候选关键短语共享的文档数目。

用于搜索结果的语义目录

背景技术

[0001] 计算机系统可以存储大量的信息，但用户往往难以找到具体的信息或者有效地探索感兴趣的特定主题区域。现有的许多搜索引擎允许用户通过输入搜索查询的方式来搜索信息，该搜索查询包含用户可能感兴趣的一个或多个关键字。接收到来自用户的搜索请求后，搜索引擎会基于关键字识别相关的文档和 / 或网页。通常，搜索引擎返回非常多的文档或网页地址，并且随后用户需要从这些文档、链接以及相关信息的列表中进行筛选，找出想要的信息。对于用户来说，这一过程可能会很繁琐、令人泄气并且很耗时。

[0002] 为了帮助用户在搜索结果中导航并找到相关的文档，搜索引擎采用了许多技术。一种方法是提供目录(TOC)，其包含与搜索查询相关的主题列表。用户可以从 TOC 中选择主题并查看与该选择的主题相关的搜索结果。在一些实现中，用户从 TOC 中选择不同的主题时目录保持静态的，这就允许用户在原始搜索查询的上下文中导航至不同的搜索结果集合。

[0003] 通常，TOC 由搜索引擎专员手工生成。特别地，搜索引擎专员识别顶端查询(即针对搜索引擎具有最大搜索量的搜索查询)，并手工识别与每个搜索查询相关的主题。然而，这种方法劳动密集型的程度很高，也不切实际于为中间(torso)和尾端的查询生成 TOC (即针对搜索引擎具有较低搜索量的搜索查询)。在一些实例中，TOC 可以由算法确定用于搜索查询，例如，通过识别搜索查询所属的领域(例如，汽车、金融等)并基于该领域提供 TOC。然而，这种方法对有些搜索查询可能不起作用，从而导致为有些搜索查询(例如中间和尾端的查询) 提供的 TOC 不存在或质量很差。这样会使搜索用户的体验不一致。

[0004] 发明概述

提供此发明内容来以简要形式介绍一些概念选集，其将在以下具体实施例中进一步描述。此发明内容不旨在识别要求保护的主题的关键特征或必要特征，也不旨在用来帮助确定要求保护的主题的范围。

[0005] 本发明的实施方案涉及到响应于搜索查询，将语义概念识别为主题以包含在目录中。在接收到搜索查询时，识别搜索结果，并生成包含主题列表的目录以用于浏览搜索结果的目录。在一些实施例中，通过对主题的本体进行分析来识别概念表的主题，以识别与搜索结果相关的主题。在进一步的实施例中，对部分主题的本体进行分析，以识别被命名的相关部分主题。在更进一步的实施例中，由搜索结果生成关键短语，并对关键短语进行分析，以识别关键短语主题。识别过的主题经过排名和选择以包含在目录中。

附图说明

[0006] 以下参考附图来详细描述本发明，其中：

图 1 是适合用于实现本发明实施例的示范性计算环境的框图；

图 2 是可以部署本发明实施例的示范性系统的框图；

图 3A、3B 和 3C 是示出根据本发明实施例的用于为搜索查询识别主题并生成目录的方法的流程图；

图 4 是示出根据本发明实施例的用于对部分主题进行命名的方法的流程图；

图 5 是示出根据本发明实施例的用于根据文档片段计算独立关键短语的方法的流程图；

图 6 是示出具有根据本发明实施例生成的目录的搜索结果页的示例性屏幕显示。

具体实施例

[0007] 此处特别地描述本发明的主题以满足法定要求。然而，描述本身不旨在限定这个专利的范围。相反，发明人已经预期到要求保护的主题还可以结合现有的或未来的技术以其它方式体现其它，以包含与本文描述的步骤类似但不同的步骤或步骤的组合。此外，尽管此处可能使用术语“步骤”和 / 或“框”来表示所采用的方法中的不同元件，但不应将这些术语解释为暗示了此处所公开各步骤之间的任何特定顺序，除非和除了在显式描述了各个步骤的顺序时。

[0008] 本发明的实施例通常针对为响应于搜索查询的搜索结果生成目录（TOC）。当接收到搜索查询时，对搜索结果进行检索。另外识别与搜索查询和搜索结果相关的主题，并根据识别出的主题生成 TOC。响应于搜索查询返回包含搜索结果和生成的 TOC 的搜索结果页面。用户可以从 TOC 中选择主题来浏览与每个主题相关的不同搜索结果。在一些实施例中，用户从 TOC 中选择不同的主题以查看不同的搜索结果集合时 TOC 是静态的，从而允许用户在初始搜索查询的上下文中浏览搜索结果。

[0009] 在本发明的各实施例中，可以以多种不同的方式为搜索查询识别主题以包含在 TOC 中。在一些实施例中，当接收到搜索查询时，确定该搜索查询的本体映射是否已经存在。例如，对于与接收到的搜索查询相匹配的搜索查询，可能已经手工生成了许多主题。再如，对于与接收到的搜索查询相匹配的搜索查询，可能之前已经接收过了，并且已经识别并缓存了主题。在这些实施例中，为 TOC 检索来自现有本体映射的主题。在进一步的实施例中，针对搜索查询检索搜索结果，并且将搜索结果与主题的本体和 / 或部分主题的本体进行比较，以识别相关的主题。在更进一步的实施例中，分析搜索结果以识别出独立关键短语，并选择关键短语主题。当识别出大量的主题时，对主题进行排名，选择排名最高的主题以用于生成针对搜索查询的 TOC。

[0010] 相应地，一方面，本发明的实施例针对存储计算机可用指令的一种或多种计算机可读的媒体，当一个或多个计算设备使用这些指令时，使得所述一个或多个计算设备执行方法。所述方法包括接收搜索查询，使用搜索查询进行搜索，以及从搜索中接收多个文档片段。所述方法还包括通过将一个或多个文档片段与主题的本体进行比较，识别一个或多个候选主题的第一集合。所述方法进一步包括通过将一个或多个文档片段与部分主题的本体进行比较，识别一个或多个候选主题的第二集合。所述方法还包括通过根据一个或多个文档片段生成关键短语主题，识别一个或多个候选主题的第三集合。所述方法进一步包括对来自所述第一、第二、第三候选主题集合中的候选主题进行排名，并基于对候选主题的排名选择一个或多个主题。所述方法更进一步包括提供搜索结果页面以响应于所述搜索查询，该搜索结果页面具有包含一个或多个主题的目录，以及用于呈现一个或多个搜索结果的搜索结果区域。

[0011] 在另一实施例中，发明的方面针对一种或多种存储计算机可用指令的计算机可读

的媒体，当一个或多个计算设备使用这些指令时，使得所述一个或多个计算设备执行方法。所述方法包括接收搜索查询，并确定搜索查询的本体映射是否存在。如果搜索查询的本体映射存在，所述方法包括基于本体映射检索第一主题集合，并将第一主题集合添加到主题列表中。所述方法还包括使用搜索查询进行搜索以获得多个搜索结果，其中每个搜索结果与文档片段相对应，并接收至少一部分文档片段作为文档集合以用于进一步分析。所述方法进一步包括将文档集合中的每个文档片段与主题的本体进行比较。对于其中确定肯定的主题识别的每个文档片段，所述方法包括将该文档片段分配给相应的主题，并从文档集合中移除该文档片段。所述方法还包括将至少一个根据主题本体识别出的主题添加到主题列表中。所述方法进一步包括将文档集合中剩余的每个文档片段与部分主题的本体进行比较。对于其中确定肯定的部分主题识别的每个文档片段，所述方法包括将该文档片段分配给相应的部分主题，并从文档集合中移除该文档片段。所述方法还包括对具有一个或多个被分配的文档片段的至少一个部分主题进行命名，并将至少一个被命名的部分主题添加到主题列表中。所述方法进一步包括根据文档集合中剩余的文档片段计算独立关键短语，将文档分配给独立关键短语，识别至少一个关键短语主题，并将至少一个关键短语主题添加到主题列表。所述方法进一步包括对主题列表中的主题进行排名，基于排名选择主题，并使用所选的主题生成目录。所述方法更进一步包括提供搜索结果页面以响应于所述搜索查询，该搜索结果页面包括目录以及用于呈现搜索结果的搜索结果区域。

[0012] 本发明的进一步实施例针对一种用于从搜索结果集合中识别主题以生成针对搜索结果的目录的方法。所述方法包括接收搜索查询，使用搜索查询进行搜索，以及从搜索中接收多个文档片段。所述方法还包括从至少部分文档片段中生成候选关键短语。所述方法进一步包括评估候选关键短语的独立性，合并相互依赖的候选关键短语，并针对每组相互依赖的候选关键短语识别频率最高的候选关键短语以生成多个独立关键短语。所述方法进一步包括将一个或多个文档片段分配给每个独立关键短语。所述方法更进一步包括基于文档片段到独立关键短语的分配来选择关键短语主题。

[0013] 前面已经简要描述了本发明实施例的概览，下面描述其中可以实现本发明实施例的示范性操作环境，以便为本发明的各方面提供一个一般性的上下文。特别地，首先特别参考图 1，示出用于实现本发明实施例的示范性操作环境，并且通常被指定为计算设备 100。计算设备 100 只是一个合适的计算环境示例，并且不旨在表明对本发明的使用范围或功能性的任何限定，也不应将计算设备 100 解释为与图中的任何部件或部件的组合有任何依赖关系或需求关系。

[0014] 本发明可以在计算机代码或机器可用指令的一般上下文中描述。所述计算机代码或机器可用指令包括诸如程序模块的计算机可执行的指令，其由计算机或其它机器(如个人数据助理或其它手持设备等)执行。一般说来，程序模块包括例程、程序、对象、部件、数据结构等，指的是执行特定任务或实现特定抽象数据类型的代码。本发明可以在许多系统配置中实现，包括手持设备、消费电子、通用计算机、更专用的计算设备等。本发明还可以在分布式计算环境中实现，在该环境下，任务由通过通信网络连接的远程处理设备执行。

[0015] 参考图 1，计算设备 100 包括直接或间接连接如下设备的总线 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 / 输出端口 118、输入 / 输出部件 120 以及示例性的电源 122。总线 110 表示一条或多条总线，例如地址总线、数据总线或它们的

组合。为了清晰起见，图 1 中的各个框都用线条示出，但是实际上，这些框表示逻辑部件而不一定是实际的部件。例如，可以将诸如显示设备的呈现部件视为 I/O 部件。此外，处理器也有存储器。我们认识到这是本领域的性质，并重申图 1 的框图仅说明能够结合本发明的一个或多个实施例使用的示范性计算设备。诸如“工作站”、“服务器”、“膝上计算机”、“手持设备”等类型不作区分，因为它们都被设想在图 1 的范围内并称为“计算设备”。

[0016] 计算设备 100 典型地包括多种计算机可读的媒体。计算机可读的媒体可以是任何能够由计算设备 100 访问的可用媒体，并包括用任何方法或技术实现的用于存储诸如计算机可读指令、数据结构、程序模块或其它数据等信息的易失性的媒体和非易失性的媒体、可移动的媒体和不可移动的媒体。计算机可读的媒体包括但不限于 RAM、ROM、EEPRROM、闪存或其它存储器技术，CD-ROM、数字化多功能盘（DVD）或其它光盘存储，磁盒、磁带、磁盘存储或其它磁存储设备，或任何其它可以用于存储所需信息并且能够被计算设备 100 访问的媒体。上面所述媒体的任何组合也应包括在计算机可读的媒体范围内。

[0017] 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媒体。所述存储器可以是可移动的、不可移动的或它们的组合。示范性的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 包括一个或多个处理器，其从诸如存储器 112 或 I/O 部件 120 的各种实体中读取数据。（多个）呈现部件 116 向用户或其它设备呈现数据指示。示范性的呈现部件包括显示设备、扬声器、打印部件、振动部件等。

[0018] I/O 端口 118 允许计算设备 100 逻辑上连接到包括 I/O 部件 120 在内的其它设备，其中有些设备可能是内置的。示例性的部件包括麦克风、操纵杆、游戏手柄、卫星接收器、扫描仪、打印机、无线设备等。

[0019] 现在参考图 2，提供了示出其中可以部署本发明实施例的示范性系统 200 的框图。应当理解，本文所描述的这个和其它布局都仅作为示例阐述。其它布局和元件（例如机器、接口、功能、顺序以及功能的分组等）可以用于补充所示出的布局或元件，或可以用于替代所示出的布局或元件，并且有些元件也可以完全省略。进一步地，此处描述的许多元件是功能性的实体，它们可以实现为离散的或分布式的部件，或与其它部件结合，并且可以以任何合适的组合和位置实现。此处所描述的由一个或多个实体执行的各种功能可以由硬件、固件和 / 或软件实现。例如，各种功能可以通过处理器执行存储在存储器中的指令来实现。

[0020] 除了其它没有示出的部件外，系统 200 包括用户设备 202、搜索引擎 204 和语义主题引擎 206。图 2 中示出的每个部件可以是任何类型的计算设备，例如参考图 1 所描述的计算设备 100。这些部件可以经由网络 208 相互通信，其中网络 208 可以包括但不限于一个或多个局域网（LAN）和 / 或广域网（WAN）。这样的连网环境在办公室、企业范围的计算机网络、内部网络和因特网中都很常见。应当理解，在本发明范围内，系统 200 中可以部署任何数量的用户设备、搜索引擎和语义主题引擎。每个都可以包含单个设备或在分布式环境下协同工作的多个设备。例如，搜索引擎 204 和语义主题引擎 206 可以是搜索系统的一部分，该搜索系统包含多个布置在分布式环境下的设备，其共同提供此处所描述的搜索引擎和语义主题引擎的功能。另外，其它未示出的部件也可以包括在系统 200 中。

[0021] 在本发明的实施例中，系统 200 包括搜索系统，其包括除其它未示出的部件外的搜索引擎 204 和语义主题引擎 206 其它。用户可以采用用户设备 202 输入搜索查询并向搜索系统提交搜索查询。例如，用户可以采用用户设备 202 上的网页浏览器访问搜索系统的

搜索输入网页，并输入搜索查询。再如，用户可以经由例如位于网页浏览器内、用户设备 202 的桌面或其它位置的搜索引擎工具条其它提供的搜索输入框输入搜索查询。本领域技术人员将认识到，在本发明实施例的范围内，其它多种方法也可以用来提供搜索查询。

[0022] 当搜索系统接收到来自诸如用户设备 202 的用户设备的搜索查询时，搜索引擎 204 对搜索系统索引 220、知识库 222 和 / 或其它包含由搜索系统维护的其它可搜索内容的数据存储进行搜索。搜索系统索引 220 一般可以包含非结构化的和 / 或半结构化的数据，而知识库 222 一般可以包含结构化的数据。相应地，搜索引擎 204 响应于接收到的搜索查询，识别许多搜索结果。另外，语义主题引擎 206 对接收到的搜索查询进行操作，以识别用于 TOC 的生成的相关主题。响应于搜索查询，可以向用户设备 202 提供包括具有 TOC 的搜索结果的搜索结果页面，该 TOC 包括由语义主题引擎 206 识别的主题。

[0023] 如图 2 所示，语义主题引擎 206 一般包括本体映射部件 210、本体主题识别部件 212、本体部分主题识别部件 214、关键短语主题识别部件 216 和主题排名 / 选择部件 218。语义主题引擎 206 采用部件 210、212、214 和 216 中的任意一个来识别语义主题。在本发明的一些实施例中，部件 210、212、214 和 216 的每一个都可以被采用以识别针对给定的搜索查询的主题，并且经识别的主题可以由排名 / 选择部件 218 进行排名并选择出某些主题以包括在 TOC 中。在其它实施例中，可以只通过部件 210、212、214 和 216 中的一部分来识别主题。例如，在一个实施例中，一旦通过部件 210、212、214 和 216 中的一个或多个识别阈值数目的主题，就不进行通过剩余部件的进一步分析。在进一步的实施例中，语义主题引擎 206 可以只包括图 2 所示的部件 210、212、214 和 216 中的一部分。任意以及所有这些变体都被设想在本发明实施例的范围内。

[0024] 当接收到来自诸如用户设备 202 的用户设备的搜索查询时，本体映射部件 210 操作以识别该搜索查询的本体映射是否已存在。例如，搜索查询可能是顶端搜索查询，搜索系统专员已手工为其识别用于该搜索查询的 TOC 的相关主题。再如，接收到的搜索查询可能与这样的搜索查询一致，即针对该搜索查询的相关主题已经被识别并缓存。如果本体映射部件 210 确定接收到的搜索查询的本体映射已经存在，那么就基于本体映射检索主题。在一些实施例中，只基于通过本体映射部件 210 检索出的主题生成 TOC。在其它实施例中，通过其它部件 212、214 和 216 中的一个或多个来识别额外的主题，详细描述如下。

[0025] 本体主题识别部件 212 结合本体存储部件 224 中存储的主题本体对所接收的搜索查询进行操作以识别针对搜索查询的相关主题。本体存储部件 224 可以存储一个或多个本体；本体主题识别部件 212 使用这些本体来将语义概念识别为所接收的搜索查询的主题。每个本体包括单词和短语的选集，它们定义了概念以及概念之间的关系。在一些实施例中，对搜索系统索引 220 和 / 或知识库 222 进行搜索，以为搜索查询检索搜索结果；并且本体主题识别部件 212 结合主题本体分析搜索结果来识别相关主题以用于可能包括在搜索查询的 TOC 中。

[0026] 本体部分主题识别部件 214 以类似于本体主题识别部件 212 的方式起作用，但它使用部分主题的本体而不是主题的本体。此处所使用的部分主题指的是部分命名的主题。每个部分主题包括部分主题标识符单词，其可以与另外的单词或短语组合以创建用在 TOC 中的主题。例如，“评论”可以是部分主题。当在上下文中分析时，部分主题标识符单词“评论”可以与诸如“专家”或“用户”等另外的单词组合来生成主题“专家评论”或“用户评论”。

相应地，一旦识别了搜索查询的部分主题，本体部分主题识别部件 214 或相关的部件对部分主题进行命名以用于可能包括在搜索查询的 TOC 中。

[0027] 关键短语主题识别部件 216 针对接收到的搜索查询分析搜索结果，以生成候选的关键短语。一般说来，关键短语主题识别部件 216 根据搜索结果生成关键短语并识别独立关键短语。对独立关键短语进行评估来识别候选主题以用于可能包括在搜索查询的 TOC 中。

[0028] 本体映射部件 210、本体主题识别部件 212、本体部分主题识别部件 214 和 / 或关键短语主题识别部件 216 可以为所接收的搜索查询识别许多主题。在一些实例中，所有识别出的主题都可以包括在响应于搜索查询在搜索页面上所提供的 TOC 中。在其它实例中，可以识别大量的主题，但只有识别出的主题的子集会包括在 TOC 中。主题排名 / 选择部件 218 操作以对主题进行排名，并选择主题以用于包括在 TOC 中。根据本发明的各种实施例，可以使用许多不同的要素对主题进行排名。仅以示例而非限制性的方式，可以基于分配给每个主题的文档总数来对每个主题进行排名。分配至给定主题的更大数目的文档为该主题提供了更高的排名。还可以基于分配给主题的每个搜索结果的排名（或经过选择的搜索结果，例如排名最高的 N 个搜索结果）对主题进行排名。对每个搜索结果的排名与搜索结果和搜索查询的相关性相对应。相应地，更高度相关的搜索结果被分配至给定主题可以为该主体提供更高的排名。进一步可以使用每个主题的长度（例如单词数目）来对主题进行排名。任意以及所有这些变体都被设想在本发明实施例的范围内。对候选主题进行排名后，排名 / 选择部件 216 选择用于 TOC 的主题。

[0029] 转至图 3A、3B 和 3C。提供了示出根据本发明实施例的用于针对在搜索系统接收的搜索查询生成 TOC 的方法 300 的流程图。如框 302 所示，接收搜索查询。本领域技术人员将认识到，搜索查询可以包括由用户输入的一个或多个搜索术语（尽管在有些实施例中搜索术语可以自动提供）。另外，搜索查询可以用许多不同的方式提供。仅以示例而非限制性的方式，用户可以采用网页浏览器来浏览至搜索引擎网页，并在输入框中输入搜索查询。再如，用户可以通过例如位于网页浏览器内、用户计算设备桌面或其它位置的搜索引擎工具条其它提供的输入框输入搜索查询。本领域技术人员将认识到，在本发明实施例的范围内，其它多种方法也可以用来提供搜索查询。

[0030] 根据图 3A 所示的实施例中，在框 304 处确定搜索查询的本体映射是否已经存在。例如，搜索查询可能是顶端查询，搜索引擎专员已经为其手工识别了针对该搜索查询的相关主题。可替换地，所接收的搜索查询可能与先前已被搜索系统处理的搜索查询相对应，来识别了相关主题，并且搜索系统可能已缓存针对该搜索查询识别出的主题。如果在框 306 处确定本体映射已经存在，那么在框 308 处检索搜索查询的主题。在一些实施例中，只有在框 308 处检索出的主题才被用来生成 TOC 以响应于搜索查询，并且该过程结束。在这样的实施例中，生成搜索页面，其包括根据在框 308 处检索出的主题生成的 TOC。在其它实施例中，过程在框 310 处继续，并且由算法识别另外的主题。

[0031] 如果在框 306 处确定搜索查询的本体映射不存在（或者如果在框 308 检索主题之后过程继续进行），那么使用搜索查询进行搜索，如在框 310 处所示。返回搜索查询的搜索结果，并且在框 312 处接收来自搜索中的排名最高的 N 个文档片段，以作为待分析的文档集合。

[0032] 如框 314 处所示，将文档集合中的每个文档片段与主题的本体（或本体选集）进行比较，以识别每个文档片段是否映射到本体中的主题。在本发明实施例的范围内，将文档片段识别为与主题本体中的主题相关联可以以许多不同的方式进行。仅以示例而非限制性的方式，在一个实施例中，基于文档片段中包含的单词将文档片段转换为特征向量，将特征向量与本体中的主题进行比较，以确定特征向量与主题的距离。通过确定文档片段的特征向量在给定主题的预定距离内来确定针对给定文档片段的肯定主题识别。如果在框 316 处，基于文档片段和本体的分析，针对给定文档片段的主题识别是肯定的，那么就将文档片段分配给所识别的主题，如框 318 处所示。另外，在框 320 处将文档片段从文档集合中移除。

[0033] 在为给定的文档片段识别相关主题（例如经由框 316~320）或确定没有来自本体的主题与该文档片段足够相关（例如经由框 316）之后，在框 322 处确定所分析的文档片段是否为待分析的文档集合中的最后文档片段。如果还有另外其它待分析的文档片段，那么重复框 316~322 的过程，直到文档集合中所有的文档片段都已与主题的本体进行了比较。在文档集合中所有的文档片段都与主题的本体比较过之后，把从主题本体中识别出的主题添加到候选主题列表中以供考虑，如图 3B 中的框 324 处所示。在一些实施例中，所有识别出的主题都被添加到列表。在其它实施例中，只有部分主题被添加。例如，在一些实施例中，仅具有预定数目的被分配的文档片段的主题被添加到主题列表。

[0034] 如框 326 处所示，将文档集合中的每个剩余文档片段与部分主题的本体（或本体选集）进行比较。如前所述，部分主题是仅部分命名的主题。每个部分主题包括部分主题标识符单词，该部分主题标识符单词可以与另外的单词或短语组合以创建用在 TOC 中的主题。

[0035] 在框 328 处确定给定的文档片段是否与部分主题本体中的部分主题相关联。在本发明实施例的范围内，将文档片段识别为与部分主题相关联可以以许多不同的方式进行。仅以示例而非限制性的方式，在一个实施例中，基于文档片段中包含的单词将文档片段转换为特征向量，并将特征向量与部分主题本体中的部分主题进行比较，以确定特征向量与部分主题的距离。通过确定文档片段的特征向量在给定部分主题的预定距离内来确定针对给定文档片段的肯定部分主体识别。如果在框 328 处，基于文档片段和部分主题本体的分析，对给定文档片段的部分主题识别是肯定的，那么就把文档片段分配给识别出的部分主题，如框 330 所示。另外，在框 332 处将文档片段从文档集合中移除。

[0036] 在为给定的文档片段识别相关部分主题（例如经由框 328~332）或确定没有来自本体的部分主题与给定的文档片段足够相关（例如经由框 328）之后，在框 334 处判断所分析的文档片段是否为待分析的文档集合中的最后文档片段。如果还有其它待分析的文档片段，那么重复框 328~334 的过程，直到文档集合中所有的文档片段都已与部分主题的本体进行了比较。

[0037] 在文档集合中剩余的每个文档片段都跟部分主题的本体比较过之后，在框 336 处对部分主题进行命名。在一些实施例中，对所有识别出的部分主题进行命名。在其它实施例中，只对一部分主题进行命名，而其它部分不被考虑用于进一步分析。例如，在一些实施例中，仅具有预定数目的被分配的文档片段的部分主题才被命名并被考虑用于进一步分析。图 4 提供了这样的流程图，其示出根据本发明实施例的用于命名部分主题的方法 400。如框 402 所示，识别分配给部分主题的文档片段内的部分主题标识符单词的出现。例如，部分

主题标识符单词可以是“评论”，并且该术语在文档片段中的每次出现都被识别。在框 404 处，提取围绕部分主题标识符单词的一个或多个单词和 / 或短语。如框 406 所示，对提取出的每个单词和 / 或短语的频率进行计数。在一些实施例中，跟踪并计数针对部分主题标识符单词提取出的每个单词和 / 或短语的位置。特别地，单词或短语可以出现在部分主题标识符单词的前面或后面。搜索系统可以分别跟踪每个单词和 / 或短语在部分主题标识符单词的前面出现多少次，以及每个单词和 / 或短语在部分主题标识符单词的后面出现了多少次。

[0038] 在分析完每个文档片段后，选择使用频率最高的单词或短语，如框 408 所示。另外，使用部分主题标识符单词和使用频率最高的单词或短语对部分主题进行命名，如框 410 所示。部分主题标识符单词和使用频率最高的单词或短语之间的先后顺序可以基于文档片段中分析过的文本里的多数排序来确定。例如，如果所选择的单词或短语出现在部分主题标识符单词前面多于出现在部分主题标识符单词后面，那么部分主题的名字的顺序将首先包括所选择的单词或短语，然后是部分主题标识符单词。回到图 3B，把命名的部分主题添加到主题列表，如框 338 所示。

[0039] 在将文档片段与主题的本体、部分主题的本体进行比较之后，在框 340 处从文档集合中剩余的文档片段生成独立关键短语。参考图 5，提供了这样的流程图，其示出根据本发明实施例的用于根据剩余的文档片段计算独立关键短语的方法 500。如框 502 所示，候选的关键短语根据文档集合中剩余的文档片段生成。根据本发明的一些实施例，使用基于马尔可夫链的方法来生成候选关键短语。

[0040] 评估候选关键短语的独立性，如框 504 所示。根据本发明的实施例，候选关键短语的独立性可以使用许多度量来评估。例如，可以基于如下度量的任意组合来确定独立性 :候选关键短语共享的单词数目、对关键短语中单词的首字母缩写词的分析以及候选关键短语共享的文档数目。

[0041] 对于每一组相互依赖的关键短语，在框 506 处对相互依赖的关键短语进行合并。若此，从相互依赖的关键短语群组中选择频率最高的关键短语用于后续分析，如框 508 所示。合并相互依赖的关键短语来识别关键短语以用于进一步分析的过程不断重复，直到不再有相互依赖的关键短语剩余。方法 500 的结果是一个或多个独立关键短语的选集，它们可以进一步被评估为可能的主题。

[0042] 参考图 3C，在从独立关键短语识别出候选主题后，将文档集合中剩余的文档片段分配给关键短语主题，如框 342 所示。在本发明实施例的范围内，将文档片段识别为与关键短语相关联可以以许多不同的方式进行。仅以示例而非限制性的方式，在一个实施例中，基于文档片段中包含的单词将文档片段转换为特征向量，并将特征向量与关键短语进行比较，以确定特征向量与关键短语的距离。通过确定文档片段的特征向量在给定关键短语的预定距离内来确定针对给定文档片段的肯定的关键短语识别。识别关键短语主题如框 344 所示，并且在框 346 处将关键短语主题添加到主题列表。在一些实施例中，所有独立关键短语都被识别为关键短语主题，并添加到主题列表。在其它实施例中，只有一部分关键短语会被承认是主题并添加到主题列表。例如，在一些实施例中，仅具有预定数目的被分配的文档片段的关键短语才被识别为关键短语主题并添加到主题列表。

[0043] 上述过程的结果是提供了候选主题列表，该候选主题列表可以包括从现有本体映

射识别出的主题、对主题本体的分析、对部分主题本体的分析和 / 或关键短语生成。在一些实例中，可能会识别出超出 TOC 需求的大量主题。若此，过程通过对主题进行排名和选择以包括在 TOC 中来继续进行。如框 348 所示，对候选主题进行排名。根据本发明的各种实施例，可以使用许多不同的要素对候选主题进行排名。仅以示例而非限制性的方式，可以基于分配给每个候选主题的文档总数来对每个候选主题进行排名。分配至给定候选主题的更多数目的文档可以为该候选主题提供更高的排名。还可以基于分配给候选主题的每个文档的排名（或经过选择的文档，即排名最高的 N 个文档）对候选主题进行排名。对每个文档的排名和每个文档与搜索查询的相关性相对应。相应地，更高度相关的文档被分配至给定候选主题可以为该候选主题提供更高的排名。进一步可以使用每个候选主题的长度（例如单词数目）来对候选主题进行排名。任意以及所有这些变体被设想在本发明实施例的范围内。

[0044] 如框 350 所示，基于排名从候选主题列表中选择主题以包括在 TOC 中，该 TOC 结合搜索结果被提供以响应于搜索查询。在一些实施例中，选择预定数目的主题。例如，搜索系统可以选择排名最高的五个主题。在其它实施例中，可以选择所有具有满足预定或动态阈值的排名的主题。在进一步的实施例中，选择排名显著高于其它主题的那些主题。基于排名选择主题的以上和 / 或其它方法的任意组合都可以在本发明实施例中采用。

[0045] 如框 352 所示，基于选择的主题生成 TOC。另外，搜索结果页面在框 354 处生成，并返回给提交搜索查询的用户。根据本发明的实施例，搜索结果页面包括搜索查询的搜索结果列表。另外，搜索结果页面包括 TOC，该 TOC 包括在框 350 处选择的主题。TOC 可以呈现在邻近搜索结果的边栏中，也可以在搜索结果页面的其它另一部分。

[0046] 以图示的方式，图 6 包括示出搜索结果页面 600 的示范性屏幕显示，该搜索结果页面 600 包括根据本发明实施例生成的 TOC。那些本领域的普通技术人员将会理解并明了，图 6 中的屏幕显示仅以示例的方式提供，并不旨在以任何方式限制本发明的范围。

[0047] 如图 6 所示，提供了搜索结果页面 600 响应于搜索查询 602（“瑟马米什娱乐”）。响应于搜索查询 602，搜索结果页面 600 在左侧窗格包括 TOC 604。TOC 604 包括如下主题 ：酒店目录、远足、划船、瑟马米什烟花、在线交友和长曲棍球。包括在 TOC 604 中的主题是基于此处讨论的本体实体和概念的分析以及关键短语提取，针对搜索查询识别的语义概念。搜索结果页面 600 还包括搜索结果区域 606，其用于显示与搜索查询 602 相关的搜索结果。在图 6 的屏幕显示中，搜索结果区域 606 当前显示的是“所有结果”。如果用户从 TOC 604 中选择主题，与该选择的主题相关的搜索结果会显示在搜索结果区域 606 中。如图 6 所示，搜索结果页面可以包括进一步的特征，例如相关搜索查询 608、搜索历史 610、赞助商网站 612 等。为了清晰起见，在搜索结果页面 600 中省略了这些部分的细节。

[0048] 不难理解，本发明的实施例将语义概念识别为主题以用于针对搜索结果的 TOC 的生成。本发明已关于特定实施例而被描述，其在各个方面都是说明性的而非限制性的。在不脱离本发明范围的情况下，其它替代的实施例对本发明所属领域的普通技术人员是显而易见的。

[0049] 如前所述，将会看出本发明很好地适于达到以上阐述的目标和目的，连同所述系统和方法显然具有以及内在的其它优点。将会理解，特定的特征和子组合是实用的，并可以被采用而无需参考其它特征和子组合其它。这通过权利要求的范围来设想并在其范围内。

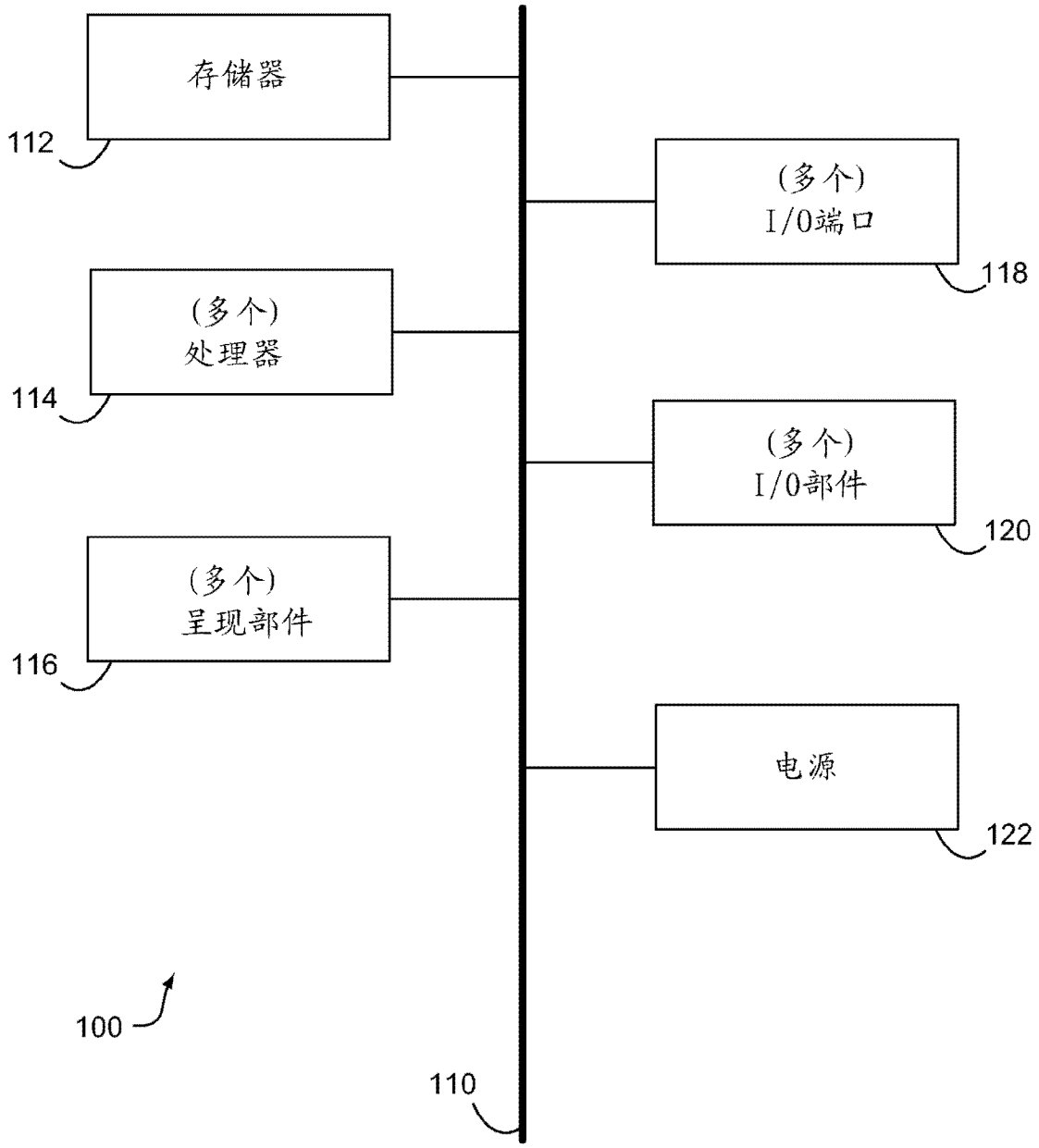

存储器
112
(多个)
处理器
114
(多个)
呈现部件
116
(多个)
I/O端口
118
(多个)
I/O部件
120
电源
122
100
110

图 1

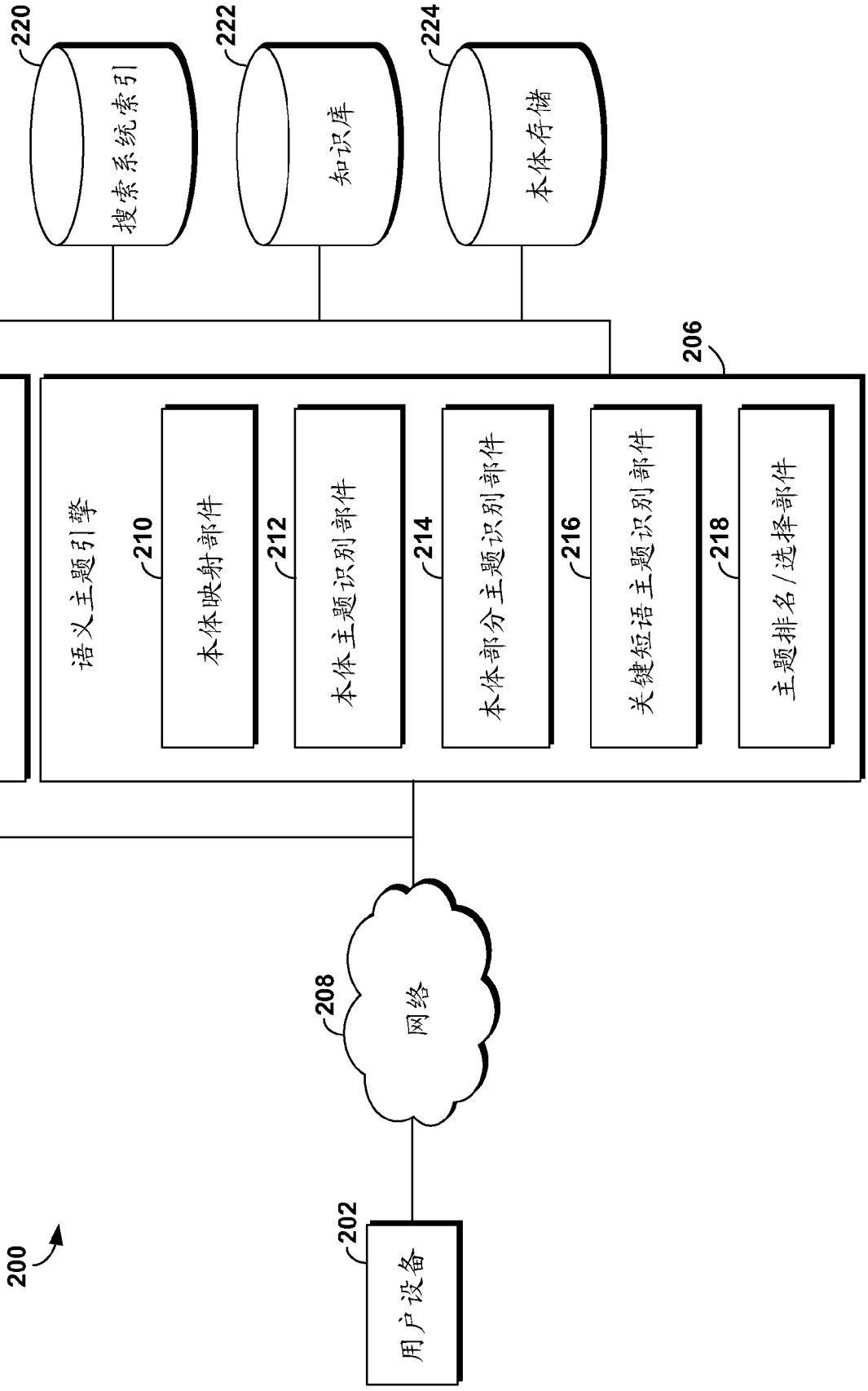

220
搜索系统索引
222
知识库
224
本体存储
204
搜索引擎
206
语义主题引擎
210
本体映射部件
212
本体主题识别部件
214
本体部分主题识别部件
216
关键短语主题识别部件
218
主题排名/选择部件
208
网络
202
用户设备
200

图 2

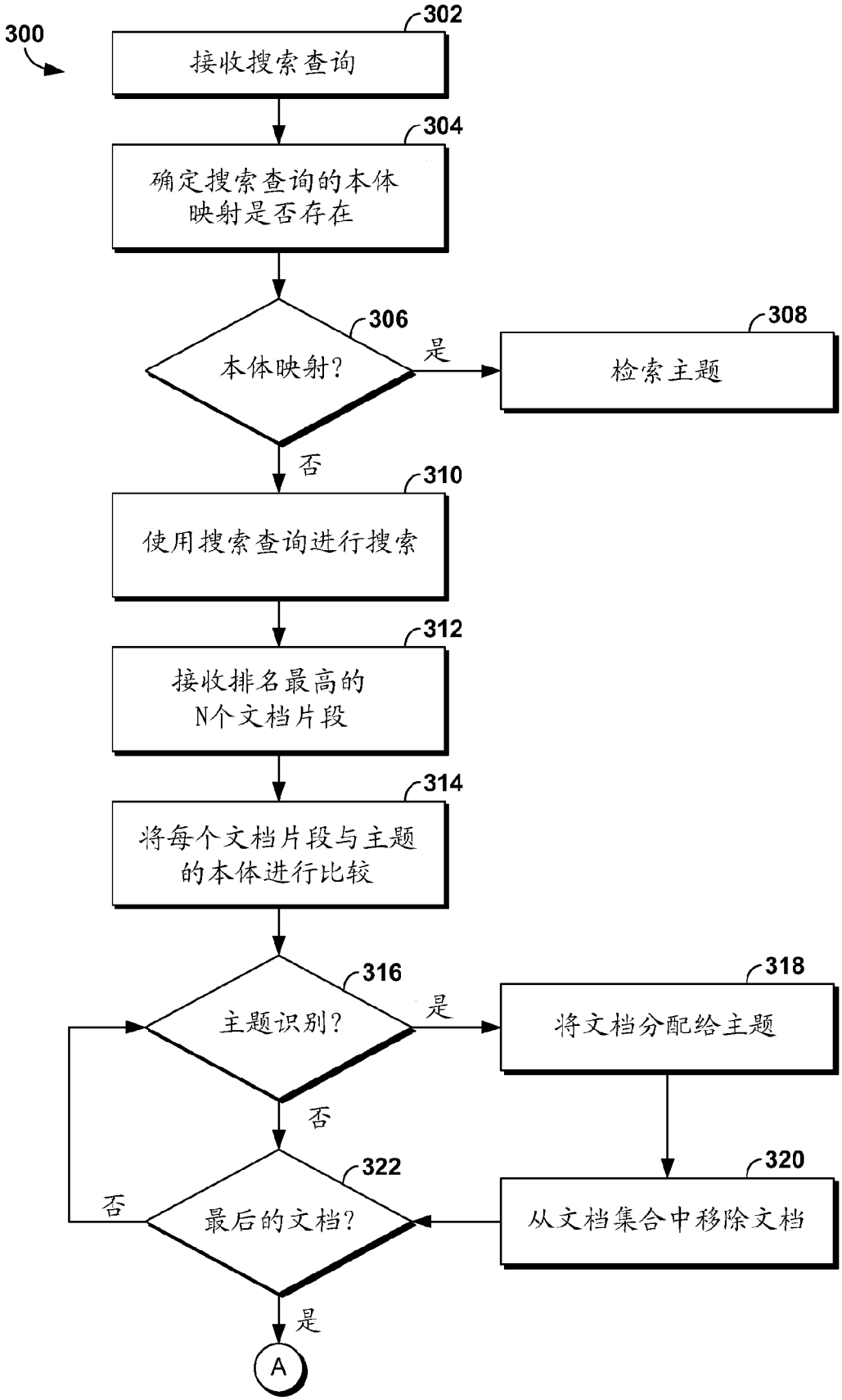

300
302
接收搜索查询
304
确定搜索查询的本体映射是否存在
306
本体映射?
是
308
检索主题
否
310
使用搜索查询进行搜索
312
接收排名最高的N个文档片段
314
将每个文档片段与主题的本体进行比较
316
主题识别?
是
318
将文档分配给主题
否
322
最后的文档?
否
320
从文档集合中移除文档
是
A

图 3A

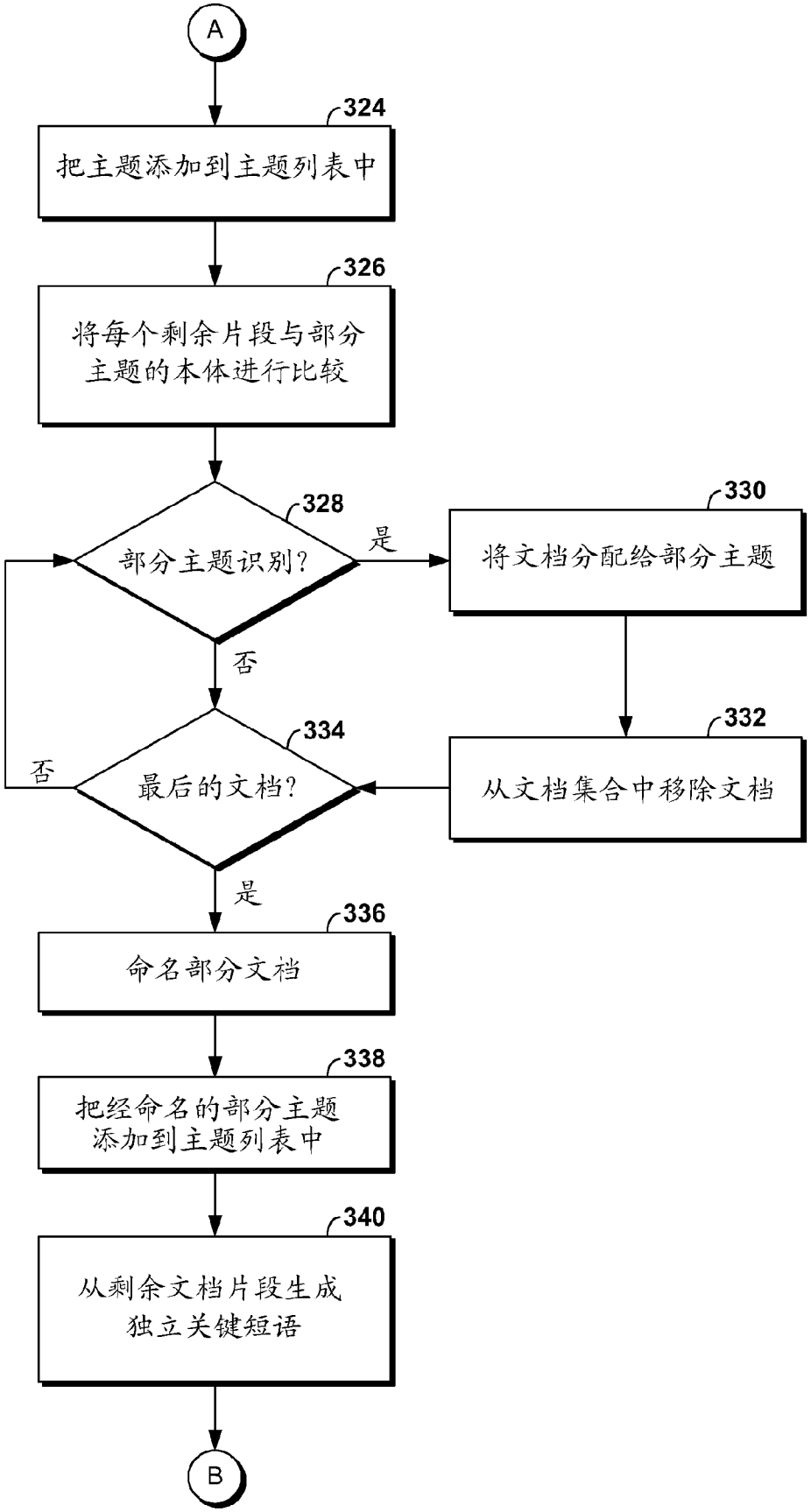

A
324
把主题添加到主题列表中
326
将每个剩余片段与部分主题的本体进行比较
328
部分主题识别?
是
330
将文档分配给部分主题
否
332
从文档集合中移除文档
334
最后的文档?
否
是
336
命名部分文档
338
把经命名的部分主题添加到主题列表中
340
从剩余文档片段生成独立关键短语
B

图 3B

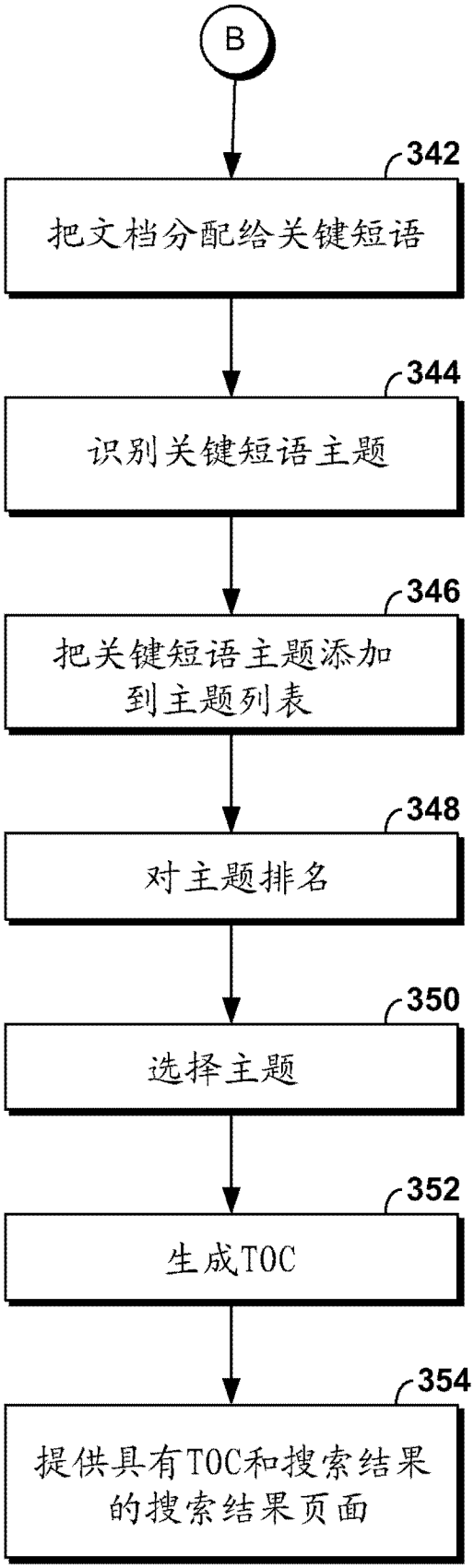

B
342
把文档分配给关键短语
344
识别关键短语主题
346
把关键短语主题添加到主题列表
348
对主题排名
350
选择主题
352
生成TOC
354
提供具有TOC和搜索结果的搜索结果页面

图 3C

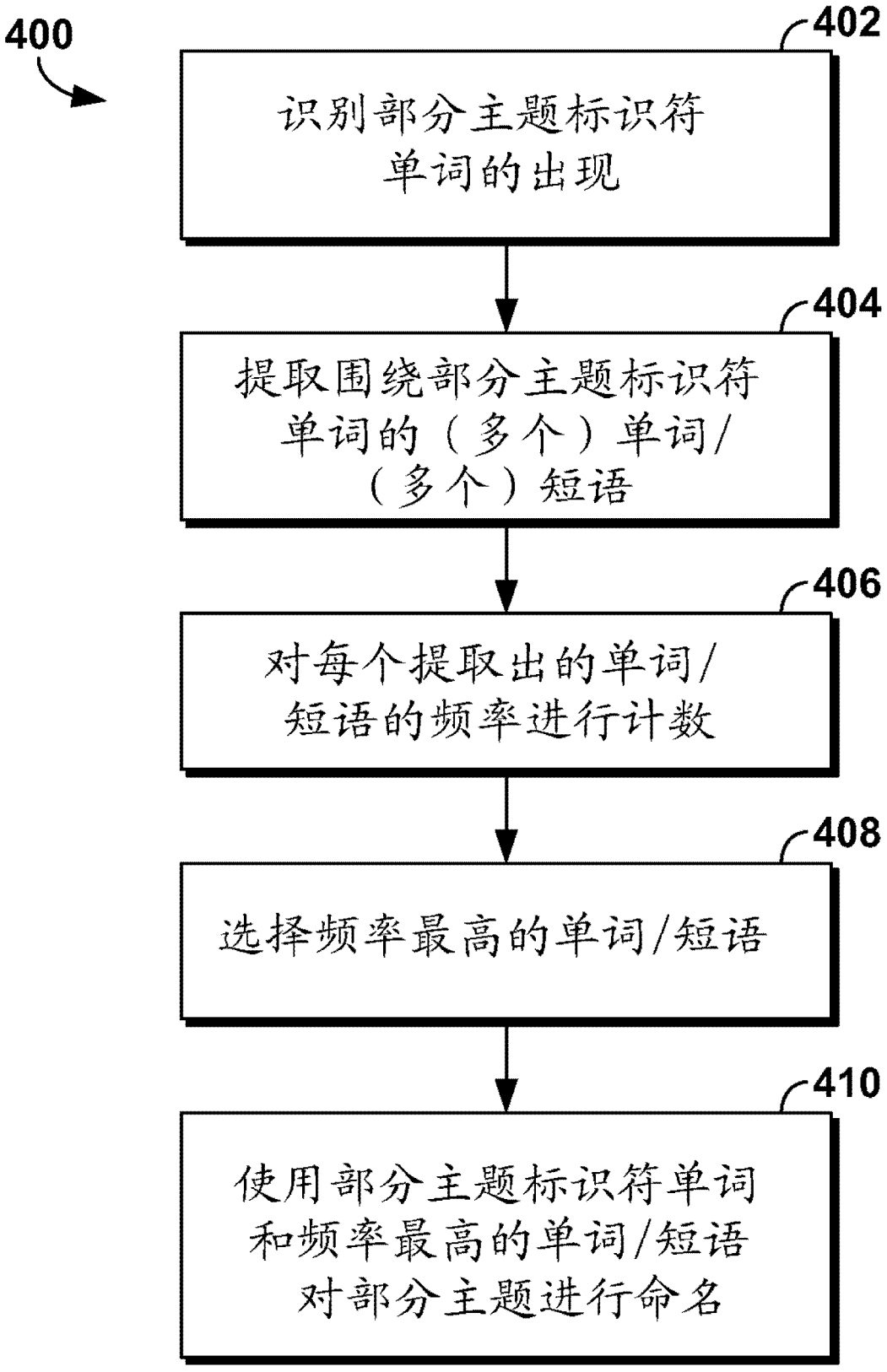

400
402
识别部分主题标识符单词的出现
404
提取围绕部分主题标识符单词的（多个）单词/（多个）短语
406
对每个提取出的单词/短语的频率进行计数
408
选择频率最高的单词/短语
410
使用部分主题标识符单词和频率最高的单词/短语对部分主题进行命名

图 4

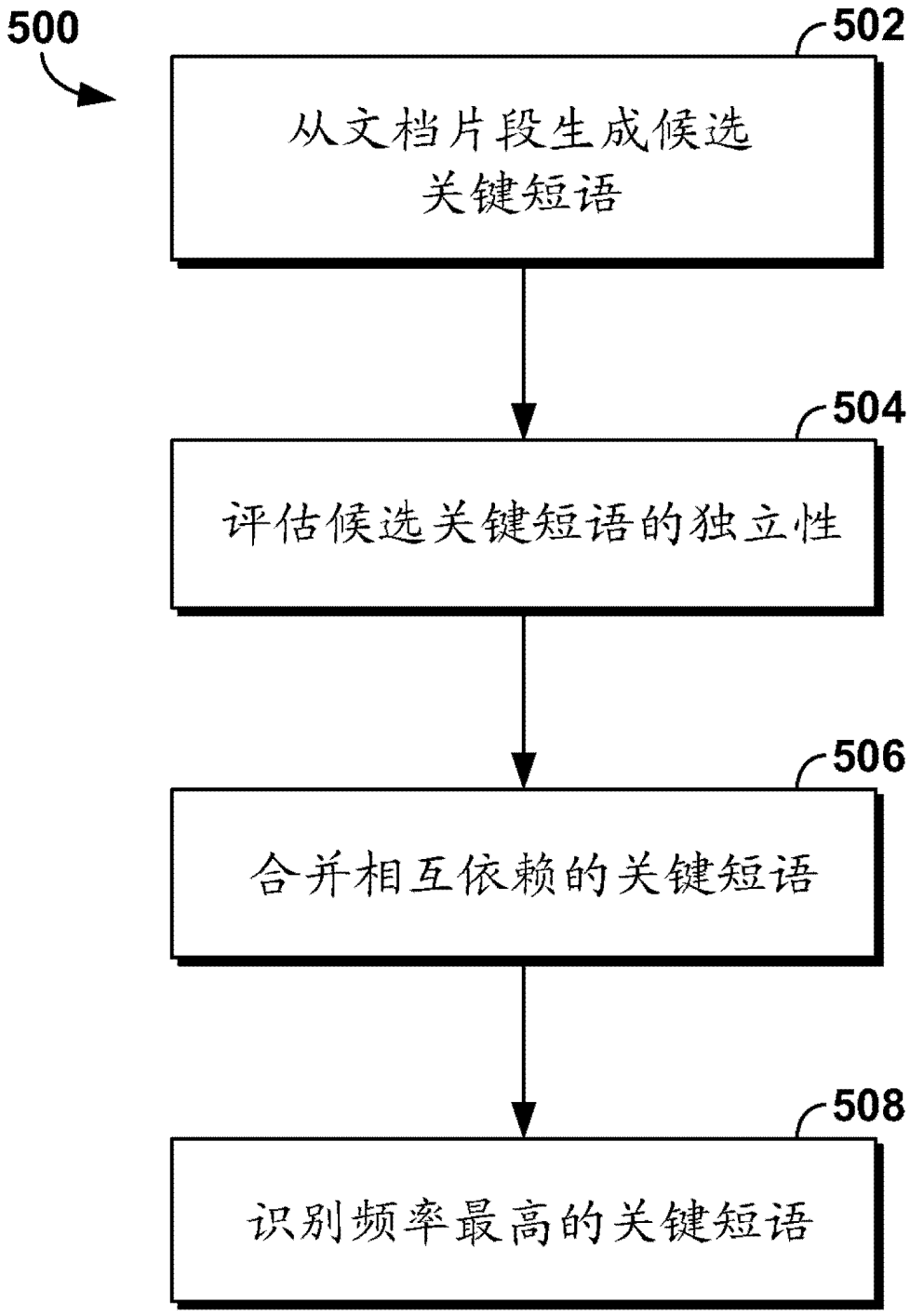

500
502
从文档片段生成候选关键短语
504
评估候选关键短语的独立性
506
合并相互依赖的关键短语
508
识别频率最高的关键短语

图 5

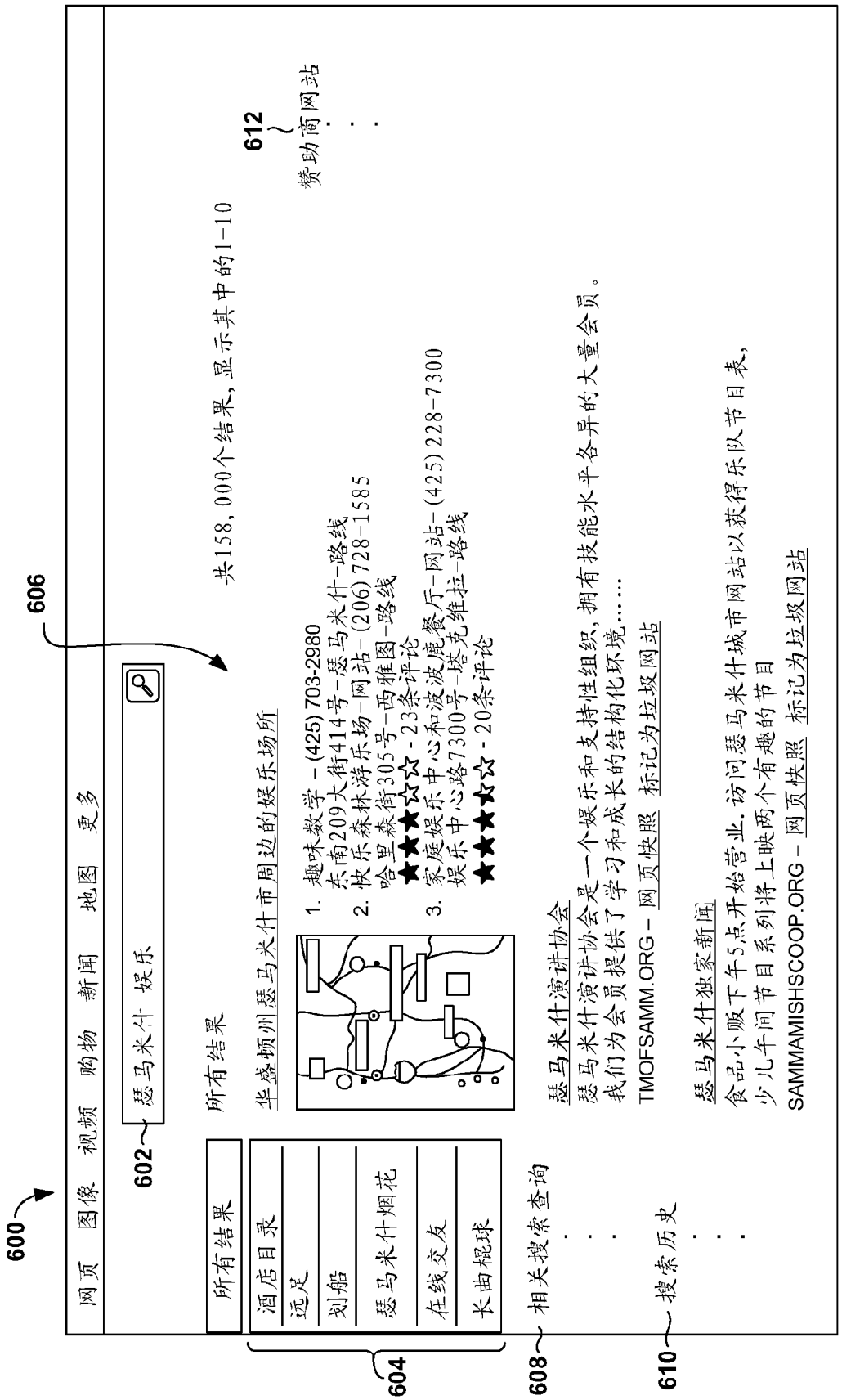

600
606
网页 图像 视频 购物 新闻 地图 更多
602
瑟马米什 娱乐
所有结果
共158,000个结果,显示其中的1-10
612
赞助商网站
华盛顿州瑟马米什市周边的娱乐场所
604
所有结果
酒店目录
远足
划船
瑟马米什烟花
在线交友
长曲棍球
1. 趣味数学 - (425) 703-2980
东南209大街414号-瑟马米什-路线
2. 快乐森林游乐场-网站-(206) 728-1585
哈里森街305号-西雅图-路线
★★★☆☆ - 23条评论
3. 家庭娱乐中心和波波鹿餐厅-网站-(425) 228-7300
娱乐中心路7300号-塔克维拉-路线
★★★☆☆ - 20条评论
608
相关搜索查询
瑟马米什演讲协会
瑟马米什演讲协会是一个娱乐和支持性组织,拥有技能水平各异的大量会员。
我们为会员提供了学习和成长的结构化环境……
TMOFSAMM.ORG - 网页快照 标记为垃圾网站
610
搜索历史
瑟马米什独家新闻
食品小贩下午5点开始营业。访问瑟马米什城市网站以获得乐队节目表,
少儿午间节目系列将上映两个有趣的节目
SAMMAMISHSCOOP.ORG - 网页快照 标记为垃圾网站

图 6