



[12] 发明专利说明书

[21] ZL 专利号 99805477.1

[45] 授权公告日 2005 年 3 月 9 日

[11] 授权公告号 CN 1192320C

[22] 申请日 1999.12.15 [21] 申请号 99805477.1

[30] 优先权

[32] 1998.12.28 [33] US [31] 09/221951

[86] 国际申请 PCT/EP1999/010228 1999.12.15

[87] 国际公布 WO2000/039708 英 2000.7.6

[85] 进入国家阶段日期 2000.10.25

[71] 专利权人 皇家飞利浦电子有限公司

地址 荷兰艾恩德霍芬

[72] 发明人 程以宁

审查员 王艳坤

[74] 专利代理机构 中国专利代理(香港)有限公司

代理人 吴立明 陈景峻

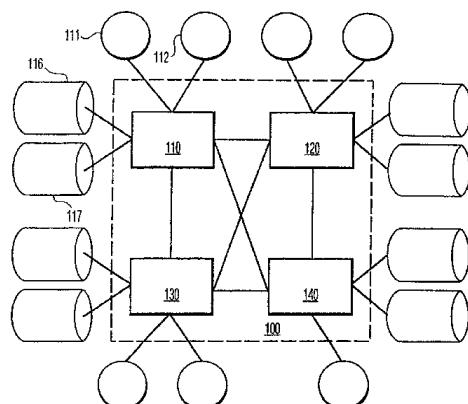
权利要求书 3 页 说明书 11 页 附图 3 页

[54] 发明名称 具有自动预过滤和路径选择的协作式主题服务器

基于主题的信息组织、路径选择和获取服务。文件可以与一个或多个主题相关，并通过由信息服务器维护的主题结构与每个主题联系在一起。

[57] 摘要

公开了一种基于主题内容的信息组织和获取系统，它有效地组织文件，目的在于快速而高效地搜索和获取。这种信息组织和获取系统经过完善，仅组织和获取那些关于给定的预定义的一组主题的文件。如果该文件不具有这套给定主题中的主题，它将被排除在所提供的服务之外。与此相似，如果该文件具有某个被所提供服务特别禁止的主题，它也将被排除在外。正是以这种模式，提供者有目的地限制了所提供的搜索和获取服务的范围，可是这样做提供了一种针对用户需求的更有效的服务。这种信息组织和获取系统也支持上下文敏感搜索和获取技术，包括使用预先定义或用户定义的意图，以及使用用户专门词汇。在一种优选实施方案中，所选的这套主题组织成有多个重叠的分层结构，并有一种分布的软件结构用来支持这些



1. 一种信息处理系统 (100)，包括：

服务器 (110)，拥有一组相关的服务器主题 (21, 211, 212)，

5 主题提取器 (310)，被配置成从源文件 (201, 301) 中提取文件主题 (211)，

文件选取器 (320)，与文件提取器 (310) 运行连接，被配置成依据文件主题 (211) 是否是服务器 (110) 的一组相关服务器主题 (21, 211, 212) 的成员主题 (211) 确定源文件 (201, 301) 作为被选中文件，

10 文件路径选择器 (330)，与文件选取器 (320) 运行连接，使被选中文件与成员主题 (211) 相关联。

2. 根据权利要求 1 所述的信息处理系统 (100)，其中，

服务器 (110) 是多个服务器 (110, 120) 中的一个，多个服务器

(110, 120) 的每个服务器拥有一组相关的服务器主题 (21, 211, 212;

15 22, 221, 2222)，

主题提取器 (310) 被进一步配置成从源文件 (201, 301) 中提取多个文件主题 (211, 2111, 2112, 2222)，

文件选取器 (320) 进一步设定为确定包括多个文件主题 (211, 2111, 2112, 2222) 的至少一个的多个文件服务器 (110, 120) 的服务器主题 (21, 211, 212; 22, 221, 2222) 的相关组的多个成员主题中每一个，

文件路径选择器 (330) 被进一步设定来将被选中文件与所说的多个成员主题的每一个主题相关联。

3. 根据权利要求 1 所述的信息处理系统 (100)，其中，

主题提取器 (310) 包括一个词语映射装置 (340)，将源文件 (201, 301) 中的词语转化便于文件主题 (211) 的提取。

4. 根据权利要求 1 所述的信息处理系统 (100)，进一步包括：

查询服务设备 (390)，当搜索主题包括成员主题 (211) 时，它确定被选中文件作为找到文件，和

文件获取器 (350)，与查询服务设备 (390) 运行连接，它被配置

成，当搜索主题包括成员主题（211）时，实现搜索文件的获取。

5. 根据权利要求 4 所述的信息处理系统（100），其中，

查询服务设备（390）包括一个词语映射装置（340），依据用户询问和用户上下文确定搜索主题。

6. 一种创建文件库（335）的方法，包括以下步骤：

定义多个主题（21, 211, 22, 22），

创建数据结构（210, 220），它具有多个节点，其中每一个节点分别与多个主题（21, 211, 22, 221）的各一个主题相对应，

10 扫描文件（201）以查找包含在多个主题（21, 211, 22, 221）里的成员主题（211），

将文件（201）与对应成员主题（211）的节点相关联。

7. 根据权利要求 6 所述的方法，进一步包括以下步骤：

将带有多个主题（21, 211, 22, 221）中相应的一组主题（21, 211; 22, 221）分配给多个服务器（110, 120）中相应的一个。

15 8. 根据权利要求 6 所述的方法，进一步包括以下步骤

创建词语的转化映射（340），

和其中扫描文件查找成员主题的步骤中包括步骤：

基于词语转化映射（340）转化文件（201）

9. 根据权利要求 6 所述的方法，其中，数据结构（210, 220）是分层的数据结构。

20 10. 根据权利要求 6 所述的方法，其中，扫描文件（201）查找成员主题（211）的步骤是依据至少一个另一文件的成员主题（211）的先前的确定。

11. 一种实现文件（201）确定的方法，包括以下步骤：

25 实现用户查询的接收，

实现基于用户查询确定搜索主题，搜索主题是多个预先确定主题中的一个，

实现对应于搜索主题的数据结构（210, 220）上的主题节点的确定，实现基于文件（201）与主题节点之间的相关性确定文件（201）。

12. 根据权利要求 11 所述的方法，其中，实现确定搜索主题的步骤包括

实现确定用户上下文的步骤，

其中搜索主题的确定进一步依据用户上下文。

5 13. 根据权利要求 12 所述的方法，其中，实现确定搜索主题的步骤包括以下步骤，

依据词语映射实现用户查询的转化，从而实现搜索主题的确定。

14. 根据权利要求 12 所述的方法，其中，

数据结构（210，220）是一种分层的结构，和

10 搜索主题的确定进一步依赖于这种分层结构。

15. 根据权利要求 12 所述的方法，其中，

多个预先确定的主题分配给多个服务器（110，120），搜索主题与
多个服务器（110，120）中的一个相联系，

主题节点的确定包括确定所说的多个服务器（110，120）中的一个。

15 16. 根据权利要求 12 所述的方法，其中，搜索主题的确定依赖于至少
一个先前另一个用户查询的搜索主题的确定。

具有自动预过滤
和路径选择的协作式主题服务器

5

技术领域

本发明涉及数据处理和交换领域，特别是文件的存储、组织和获取领域。

10

背景技术

可供访问的信息量在不断增加，并且信息量的增加速率也在加大。这种不断增加的信息增长，导致了用来存储、组织和获取信息的不断膨胀的资源。

15

传统的搜索引擎，例如因特网上用来查找文件的搜索引擎，使用了各种技术响应用户查询以快速找到用户要求的文件。其中的一种技术便是建立一个对应于万维网上文件的索引数据库。通过找出用户请求和索引数据库中信息的某种联系，完成用户请求的处理，而不是真的到万维网上去找来响应用户请求。传统的搜索引擎使用一种“爬行器(crawler)”来定位文件或更新文件。一旦一个新的或是更新的文件被定位，搜索引擎便生成一个对应于该文件的目录，其中包括比如文件中最常见单词和词组的列表。

20

还存在一些可以代替以上步骤的技术，即让文件的创建者在文件中直接增添一些关键词或词组，而这些词或词组用来给文件编制索引。为了方便起见，下文中的关键字一词就指文件索引中包含的某个词，而不管将其放置在索引中的方法。当用户输入一个查询，搜索结果依据用户查询中的词和文件索引中的关键词的匹配。本领域的技术人员可以理解一个文件的目录可能较大，万维网上基本上所有文件的索引数据库实际上极其庞大，而且将以不断增长的速率继续增加。1998年，因特网上每天约增加150万页，并且每天增速预计将继续扩大。除了增加了的存储资源的代价外，随着数据库的扩张数据库查找技术的性能在下降。

随着包含某个关键词的文件越来越多，依据关键词查找的文件获取效率越来越低，也越来越不可行。在因特网上一次关键词查找返回成千上万与此关键词相关的文件不足为奇，而其中的很多和用户的查询并无关联。为了减少对于关键词被识别文件的数目，用户必须增加提供额外的关键词或词组来增加搜索参数。可是这样做，如果用户没有选择文件中的相同词语，与用户请求有关的文件又可能被排除在外。搜索引擎可以通过在用户查询自动增加同义词从而增强性能，不过如此增加又将加剧所识别文件与用户查询无关的问题，尽管这些文件也包含了这些关键词。

主题式分类为查找与用户请求相关的文件提供了一种选择更精细的方法，因为那些与用户请求具有相同主题的文件要比那些仅仅包含匹配单词集合的文件更具备相同的信息。不过，确定文件的一个或多个主题比确定文件包含的词更复杂。传统上，主题的确定是一项人工密集的任务，需要很多人阅读和分类每个文件。信息科学领域中，基于统计学的算法和神经网，以及基于遗传学的算法，和自动分类相似文件的研究方面不断取得进展。主题分类也为一般的浏览提供了一种高效的方法，用户既可以选择感兴趣的主題又可以选取不感兴趣的主題控制浏览过程。

对于可以通过主题分类的文件，那种基于关键词的文件组织、存储和获取技术并不理想或令人满意。仅仅用主题词组代替关键词的搜索引擎，并不能为不断增加的信息量提供所需的搜索和存储上的改进。传统的方式是建立越来越大的引擎和对应与网上每个文件的索引数据库，这种索引是基于单词或词组在文件中出现的频率，这种方法对于组织和获取基于主题的文件可能完全不可行。主题决定技术的一个不加选择的应用，举例来说，也许仅仅是建立一个甚至更大的词汇集，用户必须使用这些词汇来筛选相关文件，其固有的危险是用户可能选择不同与文件索引中的词或词组。因为大多数文件包含多个主题，额外主题信息加进已有索引实质上也将增加存放这些信息的数据库的大小。

发明内容

本发明的一个目的是提供一种信息组织和获取系统，以有效组织文件

从而基于主题内容快速和高效的查找、获取。本发明进一步的目的是提供一种可以增强的信息组织和获取系统。本发明另一个目的是提供一种支持上下文敏感搜索和获取技术的信息组织和获取系统。本发明还有一个目的，即提供这样一种信息组织和获取系统，它允许用户使用不同于用来组织这些信息的单词。

这些目标即其他目标的达到，是通过提供一种信息组织和获取系统，该系统优化为仅获取那些与给定的一组主题相关的文件。本发明提供了一种方法和设备，通过协作式主题信息服务器网络，完成文件的自动预筛选和路径选择。信息服务器用来根据所选主题组组织和获取文件。所选的该组主题组织成具有多个重叠的分层结构，和一个分布式软件结构用来支持基于主题的信息组织、路径选择和获取服务。文件被自动预筛选以确定它们是否与所选主题组相关，只有相关的文件才被确认以供以后的获取。文件可能和一个或多个主题相关，它通过由信息服务器支持的主题分层结构与每个主题联系在一起。

在一个优选实施方案中，通过提供一种支持使用基于用户正在其中查找的上下文而增加查找准则的预定义或用户定义视图从而增强获取处理的方法和设备。

本发明中的组织和获取处理也通过使用内部一致的主题词汇而得到增强。文件作者或搜索文件的用户使用的用语和词组，都被翻译成通用的内部词汇，因此在允许单词和词组的多种选择的同时，提供了增强了的组织和搜索能力。

附图说明

以下以举例的方式参考附图详细说明本发明，其中：

图 1 示出根据本发明的一种信息处理系统的示例性方块图。

图 2 示出在根据本发明的一种信息处理系统中，文件和多个主题节点的联系的示例。

图 3 是通过根据本发明的一种信息处理系统来组织、搜索和获取文件的流程图示例。

具体实施方式

根据本发明，文件是通过主题分类和组织的。单独的服务器的网络用
来识别和获取文件。通过设计，每个服务器负责一个主题或多个主题的独
立的选定组。主题由服务网络的提供者选择，例如，基于预计的用户对特
5 定主题范围的请求。当每个新的主题被确认需要加入，把它加在一个已有的
的服务器上，或者加入另外一个新增加的服务器上。这样，主题的范围就
控制住了，并且通过增加网络上的服务器仍然保持其可扩展性。如果某个
文件不具有网络主题所包含的主题，它将被排除在提供的服务之外。相似
10 的，如果某个文件包含所提供的服务特地禁止的主题，它也被排除在外。在
这个模型中，提供者有目的性的限制了所提供的搜索和获取的范围，但是这
样做也提供了一种针对预期用户请求的更有效的服务。随着请求的增加，
提供额外的主题和服务器，因此允许了所提供的服务扩展。

图 1 描述了根据本发明的信息处理系统一个例子。信息处理系统 100
包括主题服务器 110, 120, 130, 140 组成的网络。方便起见，把主题服务
器网称作一个联合 100。每个主题服务器负责所述的一组主题，这个联合中
15 服务器主题组的集合称作联合主题。确认和某个主题相关的文件与包含该
主题的主题服务器中的主题相关。通过提供一个服务器网络，每个服务器
负责所选的一组主题，与组织和搜索文件有关的工作量分配给服务器。

在一个优选实施方案中，进一步分配工作量，某个服务器还负责指定
20 的客户机和指定的文件源。如图 1 所示，主题服务器 110 具有指定的客户
机 111、112，和指定的文件源 116、117。指定的客户机 111、112 比如说是
万维网浏览器，用户用它与系统 100 相交互。文件源 116、117 比如说是
因特网上的存储设施。为了理解方便，文件一词这里指一段信息，比如一
25 页或多页文本，也可能是其它形式的信息，例如视频和音频片断，图形，
图画，计算机程序和其它。

和传统的搜索引擎一致的是，主题服务器 110，周期性地发送网络爬
行器给文件源 116、117，收集新的或更新的文件。服务器 110 扫描爬行器
发现的文件，确定每个文件的主题。和传统搜索引擎不同的是，只有文件
的一个或多个主题包含在联合主题中，服务系统 100 才选择该文件以供识

别。如下文所述，可以用自动装置来确定主题，比如使用语义处理，试探学，基于知识的系统，机器学习，和其它类似的装置。还可以通过附加在文件后的信息确定主题。例如视频“文件”可能具有相关的摘要，音频文件可以根据风格或作者存储在文件源 116、117 中，如此等等。用相似的方法，可以把手工确定的主题结果和文件存在一起，然后服务器 110 具此确定文件的主题为系统 100 所用。正如如下所述，这样一个问题对本领域的技术人员是显然的，因为可能的主题事先定义好，相比与盲目的寻找定位文件的所有可能主题，决定某个文件与某一个主题相关的能力提高了。服务器 110 和联合中包含有一个或多个联合主题服务器 120、130、140 交换与文件相关的标识符和文件主题。同样对于本领域的技术人员显然的是，存储与联合主题的预确定主题组相关的文件标识符可望比在传统搜索引擎这存储如前所述通常的关键词索引或类似内容消耗少得多的资源。

通过把文件和其所包含的联合主题中的每一个主题连系在一起形成文件库。每个服务器根据服务器覆盖的主题，通过文件标识符组织文件。在一个优选实施方案中，主题组织成树节点，往树根方向的节点具有越来越一般性的概念，往树叶方向的节点具有越来越具体的概念。一个所选文件和一个或多个主题节点，每个节点指向零个或多个文件。图 2 所示为服务器 110 中的一棵树 210 和服务器 120 中的一棵树 220 的结构示例。树 210 是对应于艺术的一棵分枝树，树 220 是对应于工程的一棵分枝树。如图所示一般性主题艺术 21，具有更具体的枝节点文艺复兴时期 211 和现代 212。和传统树术语一致的是，分支文艺复兴 211 和现代 212 的全名是艺术. 文艺复兴时期 211 和艺术. 现代 212。艺术. 文艺复兴时期节点 211 包含分支艺术. 文艺复兴时期. 油画 2111，艺术. 文艺复兴时期. 绘画 2112，艺术. 文艺复兴时期. 雕刻 2113，和艺术. 文艺复兴时期. 表演 2114。同样，工程节点 22 包括分支工程. 电子 221 和工程. 航空 222。工程. 航空节点 222 包括分支工程. 航空. 固定机翼 2221 和工程. 航空. 旋转机翼 2222。

在图 2 的例子里，服务器 110 的爬行器已经在文件源 116 中找到了文件 201。比如文件 201 包含的信息与达芬奇的素描画直升飞机和油画蒙娜丽莎有关。依据文件 201 的内容，服务器 110 从文件 201 中提取出主题，其

其中包括艺术. 文艺复兴时期(和达芬奇相关), 艺术. 文艺复兴时期. 油画(和蒙娜丽莎相关), 艺术. 文艺复兴时期. 素描(和直升飞机相关)以及工程. 航空. 旋转机翼(也是和直升飞机相关)。请注意, 由于主题是预先定义的, 服务器 110 可以经过组织从而优化主题提取过程。例如, 每个主题有一组相关的关键词和词组, 一种传统的加权和阈值处理是根据关键词和词组在文件中出现的频率, 可以据此确定一个文件是否与某个主题相关。在一个优先实施方案中, 传统的技术通过基于词组在文件特定位置的启发式方法得到提高, 比如标题, 或词组的字体(粗体、斜体等), 单词和词组存在于元标记里, 等等。使用预先定义的主题也为改进了的组织技术的使用提供了便利。比如, 在一个优先实施方案里, 采用了机器学习技术来增强服务器确定文件主题的能力。典型地, 决定一个给定主题是否包含在文件里是依据许多独立的和非独立的决策。在一个训练模式下, 根据确定每个主题的正确性, 主题提取器得到一些反馈。反馈用来调整主题提取器以后的确认, 比如使用基于每个确定的正确性调整与每个确定元素相关的相似因子的贝利斯网络。正确的确认增大与每个决策元素相关的相似性因子, 而错误的确认则减小因子。同样, 机器学习技术可以用来依据可见的文件主题聚类和其它一些因素建立或修改主题的分层组织结构。这些及其它一些文件组织分类技术, 比如基于认知的系统, 机器学习, 模糊逻辑, 及与此类似的技术在已有技术中是很常见的。

在一个优先实施方案里, 服务器 110 找出爬行器找到的每个文件中的每一个联合主题。或者, 因为可以优化使每个服务器提取其所负责的每个主题, 如此联合 100 可以组织成每个爬行器找到的每个文件由某个服务器独立处理。在图 2 的最佳实施例的例子里, 服务器 110 把文件 201 的一个标识符传给服务器 120, 告诉它文件 201 含有工程. 航空. 旋转机翼这一主题。文件 201 的标识符可以是, 比如文件 201 的网络地址, 或者其它可以唯一定位文件 201 的标识。根据本发明, 文件 201 和树 210 及 220 的 211, 2111, 2112 和 2222 分支连系在一起, 比如通过把文件 201 的标识符加进每个节点的相关文件表中。

图 2 显示的是一个传统的树结构。这方面, 常见的其它数据组织结构

也是可行的。在优选实施方案里，分层结构，比如树，是首选的，因为它允许了现有与人类组织信息方法适应的搜索技术。重叠的，和“缠结的”树结构被应用在优选实施方案里，使用户通过多种搜索途径到达某个给定节点。比如，一棵包含物理主题的树可能具有节点物理. 飞行. 直升飞机，
5 它与前述工程. 航空. 旋转机翼是同一个主题。与此类似，优选实施方案中还具有姊妹节点间的联系，比如图 2 中，让文艺复兴时期的素描 2111 和油画联系在一起。

图 3 描述了对应本发明的一个信息处理系统流程图示例。该流程图显示了，例如，由服务器提供者提供的以实现基于主题内容的组织、查找和获取的信息的资源，和这些元件之间数据的传输。在优选实施方案里，为了方便，每个服务器都具有图 3 中任一个功能块，尽管这些功能块可以分布在联合中。主题提取模块 310 提取出文件 301 的主题词语和词组。主题提取模块 310 和文件选择分类模块 320 一起完成这项工作，如前所述，在使用预先定义主题的基础上，分类模块 320 增强了提取过程。而词语映射服务模块 340，通过执行比如把提取出的词语和词组翻译成联合中所用的词和词组一类的操作，协作这一过程。举例来说，“直升飞机”一词转化成“旋转翼飞行器”，方便确定包含主题“直升飞机”的文件是否和表达为“旋转翼飞行器”的主题相关，而不是“直升飞机”。因为使用了选择的主题，同义词和词组的正确辨认相比于独立主题的翻译实际上可以得到改进。
10
15
20

如果主题提取模块 310 可以找到文件 301 中的一个联合主题，文件 301 将被文件选择分类模块 320 选中并分类，文件的标识和主题被送往文件路径选择模块 330。文件路径选择模块 330 把这些信息送往数据库 335，因为所找到的文件主题属于该联合。数据库 335 包含图 2 讨论过的基于主题的数据。在优选实施方案里，数据库 335 根据和每个服务器相连的主题分布在联合中。同样如上所述，在优选实施方案里，每个服务器包含图 3 所示模块。为了便于参考，“客户 - 服务器”一词，用来指包含一个给定模块的服务器。文件路径选择模块 330 直接更新数据库 335 与它的客户 - 服务器相关的每个主题，并把文件的标识和主题送给其它含有该文件主题的
25

服务器里的路径选择模块。那些其它的文件路径选择模块更新数据库 335 的相应客户 - 服务器主题。相应地，文件路径选择模块 330 应设定成可以从其它服务器接收文件标识和主题，直接更新数据库 335 每个与它的服务器相关的主题。也就是说，举例来说，如果文件路径选择模块 330 位于图 2 5 的服务器 120 中，该文件路径选择模块 330 为更新与服务器 120 相关的数据库，与工程 22 相关的所有文件标识，当文件主题包含艺术 21 时，它会把文件标识和主题送往服务器 110 上的文件路径选择模块。同样，当服务器 110 上的文件路径选择模块发现其找到的文件包含主题工程 22，它会把文件的标识和主题送往服务器 120 上的文件路径选择模块。

10 图 3 示出了一个任选的外部联合模块 360 和一个任选的代理服务模块 370，它们实现信息处理系统内部的多个联合的结合。在优选实施方案里，多个联合用来提供每个联合内的一定专门化程度。相关的主题放在一个联合里，而不相关主题放在不同的联合里。这样，在联合的特殊领域内，每个联合可以根据用户的反馈进行控制和扩展，以提供高效的获取。多个联合代理服务也来访问其它提供者的资源，从而使得服务提供者为用户提供更广泛的主题，而无需服务提供者为如此广泛的主题分类所有的文件。代理服务模块 370 调节其本机联合资源的访问程度。比如，在同一个提供者的各个联合间，当文件主题包含一个或多个别的联合的主题时，文件标识和主题将从一个联合转发至另一个联合。在不同的提供者的联合之间，20 代理服务模块 370 会允许搜索别的联合和获取文件，但可能禁止本机联合的文件选择分类模块 320 确认得文件标识和主题送往别的联合。

25 在优选实施方案里，多个联合结构中的每个联合的一个服务器被用作代理服务器，用于与其它联合的对应代理服务器接口。代理服务器总结与本联合相关的信息，使用代理服务模块 370 把这些信息适当地送往其它联合，并且从其它联合各自的代理服务器接收相关信息。代理服务模块 370 还影响词语映射服务模块 340 和文件选择分类模块 320 的更新处理，从而实现确认和选择文件 301 中外部联合主题。如果发现文件 301 含有外部联合主题，文件路径选择模块 330 把文件标识和主题送往外部联合主题/视图服务模块 360。如果得到代理服务模块 370 的同意，如上所述，外部联合主

题/视图服务模块 360 把文件标识和主题提供给每个包含一个或多个该文件主题的外部联合。

管理服务模块 380，提供管理信息处理系统的服务，包括主题的建立和修改，服务器的增加和去除，代理服务区的建立，和其它类似的服务。

5 图 3 还介绍了文件搜索和获取得流程示例。一个用户通过客户设备 305 与该系统交互。查询/结果服务模块 390 处理用户请求以确定搜索主题。对于文件选择分割模块 320 的处理，词语映射服务模块 340 把用户的查询转换和增加为信息处理系统所使用的术语，从而帮助查询处理。由于使用预先定义的主题，在优选实施方案里的查询/结果服务模块 390，可以通过把查询词汇处理成与联合主题及主题层次结构相一致，从而优化搜索主题的 10 确定。

15 使用预先定义的主题和主题分层结构提供传统基于关键词的搜索引擎无法实现的优点。例如，在优选实施方案里，通过让系统提出顺着主题分层结构调整查询词语的意见，引导用户完成查询词语的规范表述。以图 2 为例，当用户选择“艺术”作为查询时，提供树 210 的图形表述；之后，用户沿着树 210 不断前进，使用键盘，鼠标，或其它输入设备，比如语音识别系统。当用户到达树 210 的每个节点时，与该节点相关的文件介绍就显示出来，然后用户选择，获取一个和多个发现的文件或浏览其它与这一 20 主题有联系但相关性较小的其它文件，或者继续搜索。在优选方案里，和子节点及姊妹节点相关的文件，也包含在认为与主题相关的文件集合里。由于在优选方案里，主题组织成分层结构，随着用户沿着分层结构走下去，相关文件的范围逐步缩小，从而提高了搜索的性能和效率。

25 请注意，以上的过程提供了传统关键词搜索引擎没有的附加优点。比如，分层结构的显示让用户深入了解系统内部文件是如何组织的，并让用户据此调整他和她的搜索方法。显示的内容还给了用户即使得反馈，告知用户的术语用是否适合于系统的辨别。在优选实施方案里，词语映射服务模块 340，允许用户增加与系统所用词语相关的词语或词组，从而允许个人化的搜索词汇。

根据本发明的一个方面，查询/结果服务模块 390，通过将查询公式化

为上下文敏感查询或视图而提高了用户查询的质量。比如，用户查询的上下文可能依据用户是在家里或在办公室而不同。比如，如果用户在正常工作时间提交有关餐馆的查询，搜索过程可以侧重商业方面，如果在别的5时间提交这种查询，搜索可以侧重家庭方面。在优选实施方案里，查询/结果模块 390 还依据特定用户的爱好，并利用用户的偏好使搜索结果个人化。与主题提取中一样，在优选实施方案里，使用机器学习和其它的技术，根据可见的用户行为提供更有效的搜索方式。这里引用这样的一个应用实例作为参考，由 Chandra Dharap 于 1998 年 6 月 25 日提交的“基于上下文和10 用户个性驱动的信息获取”，律师备案目录表 PHA 23,422，序列号 09/104,491，该申请使用一种方法和设备在用户访问数据库时，根据用户之前的查询增添查询条件。仍然使用餐馆的例子，如果用户提交有关餐馆的查询后，总是打开法国餐馆文件并且总是忽略快餐馆文件，查询/结果服务模块 390 会给包含法国餐馆的文件以更大的选择加权，而给包含烧烤食品主题的文件以更少的加权。这一申请还允许其它形式的搜索输入，比如15 一个形状和图案的画，代表一段音乐的曲调或节奏，等等。因为使用了本发明的预先定义的主题，所以这种任选形式可以根据每个主题而设。比如，负责电路主题的服务器可以做成接收电路图作为用户查询输入，然后处理该图以查找对应相似电路的成员主题。或者，用户可以指向电路中某个元件，而服务器提供列有此种设备零售商的文件。在这一方面，对于依据本20 发明的这些或其它一些具体应用对于本领域的技术人员是容易理解的。

在优选实施方案里，其它一些学习技术被用来为那些具有不同含义的查询确定适合的搜索路径。比如，词语 “card” ，可以指贺卡、扑克牌、信用卡，印刷电路板卡，怪人等。在优选实施方案里，查询/结果服务模块 390 根据所提供的用户个人信息或词语的通常用法为查询词选择某一个25 主题。如果，对于所选的主题，用户修改了查询词以查找单词 “card” 可能对应的其它主题中的一个，则查询/结果服务模块 390 将根据用户对 “card” 一词的新的用法侧重选择其它的主题。考虑本发明阐述的内容，这些及其它一些基于经验和预先定义主题的使用来改进用户查询词语的方法对于这方面的某个普通技术来说是容易理解的。比如，对应于一个查询的多个

可能主题可以显示给用户选择，用户可以选择让查询/结果服务模块 390 对于类似查询总是选择被选主题，或每次都显示多个主题来选择。

在优选实施方案里，和关键词搜索系统中多个关键词的使用相似，查询/结果服务模块 390 也允许用户同时使用多个查询词来改进搜索请求，还可以使用布尔符号和模糊逻辑术语来组合主题。比如，用户可以选择搜索与主题政府. 美国和医学研究. 实验性的. 动物相关的文件，但是排除与主题大学. 医学相关的文件。每个包含一个或多个所选主题的服务器，把与每个主题相关的文件引用，通过本地 - 联合 - 主题/视图 - 服务模块 350 送往查询/结果服务模块 390，或者，外部联合主题/视图服务模块 360。查询/结构服务模块 390 根据上述的用户偏好和上下文，过去的经验，组合逻辑用语等，来整理这些文件引用显示给用户。

请注意，根据预先定义的主题和分层结构来组织文件可以大大地节省搜索定位文件的时间和资源。通过根据主题组织文件，响应查询而显示给用户的与查询无关的文件数也大大减少。通过提供上下文敏感的用户查询，把查询转换成预先定义的主题和分层结构中使用的词汇，将为用户给出一个合适节点的速度提高了很多。通过基于机器学习技术，动态地调整主题提取过程和用户查询处理本发明的信息处理方案的效果和效率不断提高。

上述内容只是阐述了发明的原理。因此，本领域的技术人员可以理解可以设计出各种各样的装置，这些装置虽然没有在这里直接描述或显示，但包含了本发明的原理，所以也属于本分明的本质和范围。比如，词语映射服务模块 340 可以借助现有的和未来的语言处理技术得到改进，包括在多种语言间的翻译能力。图中显示的结构只是示例性的，其它类似的结构也属于本发明的本质和范围。比如，联合中的前述服务器，可以单独用来组织和获取文件，而别的设备用来和客户机打交道。对于本领域的技术人员，这种可选的功能性划分是容易做到的。

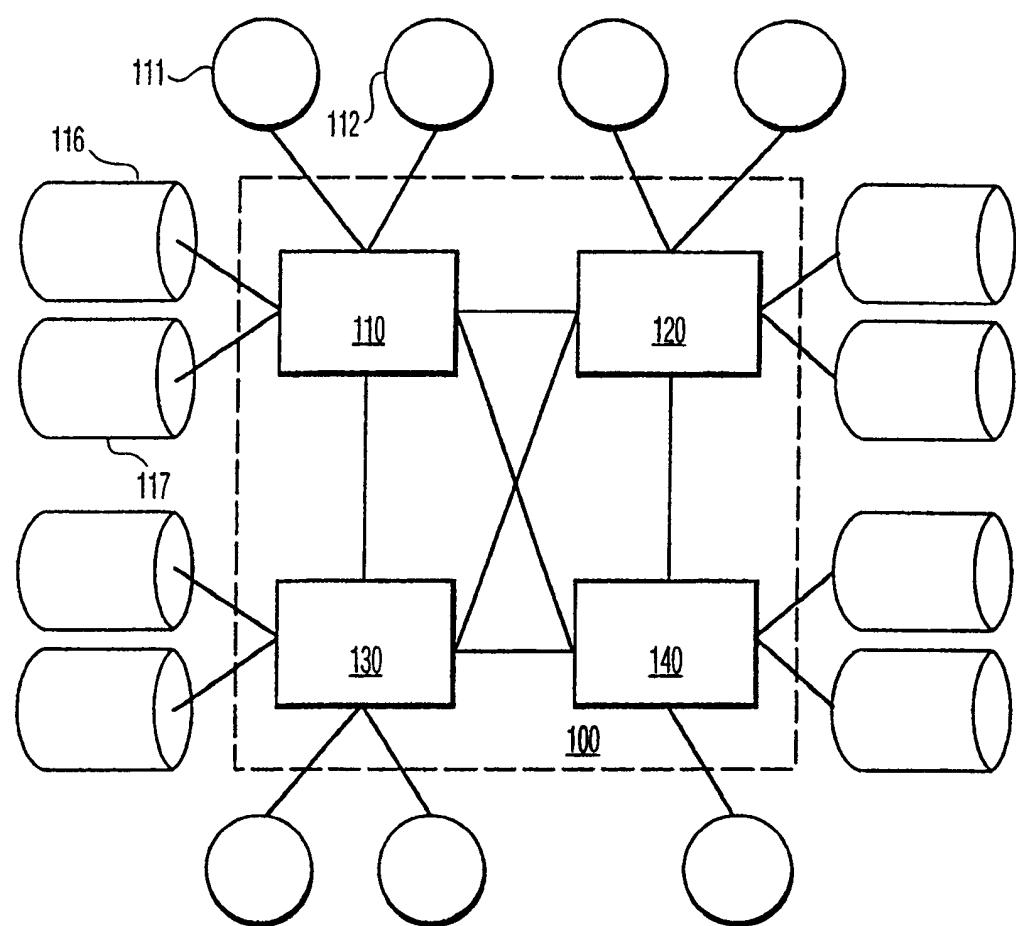


图 1

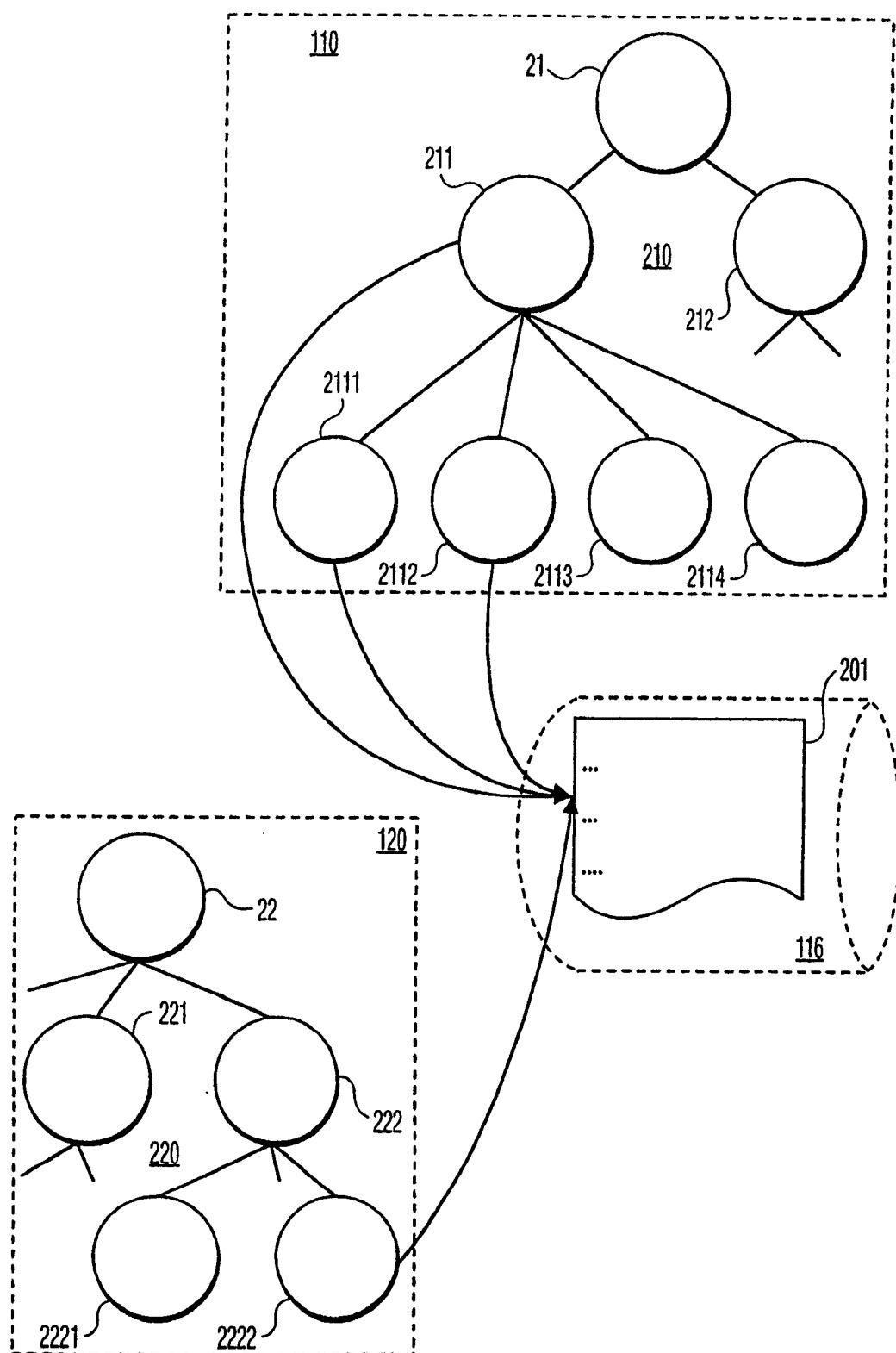


图 2

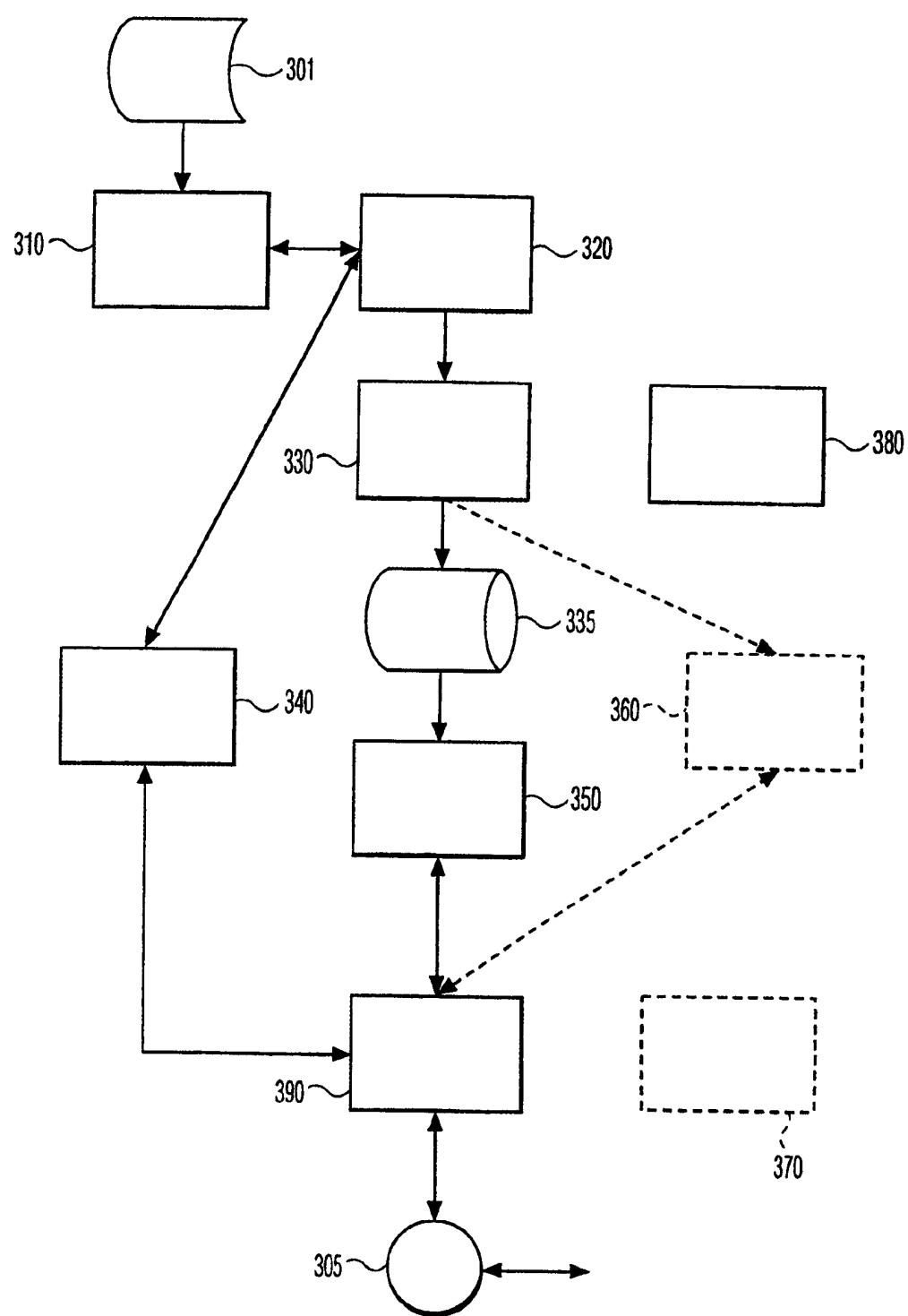


图 3