[19] 中华人民共和国国家知识产权局

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

[12] 发明专利申请公开说明书

[21] 申请号 200610051555.4

[43] 公开日 2006 年 10 月 4 日

[11] 公开号 CN 1841377A

[22] 申请日 2006.2.28
[21] 申请号 200610051555.4
[30] 优先权
[32] 2005. 3.29 [33] US [31] 11/096,429
[71] 申请人 微软公司
地址 美国华盛顿州
[72] 发明人 A·C·卡帕迪亚 H·M·克劳
J·S·布克 R·D·帕克
[74] 专利代理机构 上海专利商标事务所有限公司
代理人 张政权

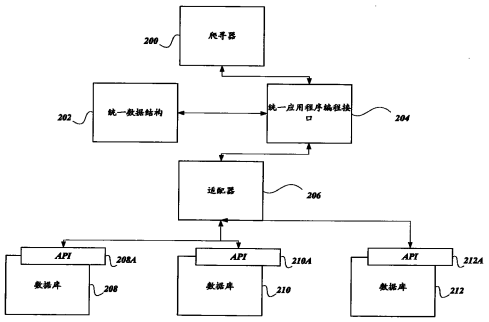
权利要求书 3 页 说明书 10 页 附图 10 页

[54] 发明名称

爬寻数据库以找出信息

[57] 摘要

一种系统便于搜索引擎搜索各种数据库和应用程序文档中的信息片断，而无需知道用于与各种数据库和诸如 Microsoft Excel 文档等应用程序文档通信的专用协议。 尽管有爬寻器连接到数据库必须使用的多种可能的协议，但是系统包括一种统一应用程序编程接口，以统一访问数据库和应用程序文档中的信息片断的方式。



1．一种用于探查不同数据库以找出信息的系统，包括：

储存在计算机可读介质中的统一数据结构，用于枚举要探查的一组数据库以及用于约束要返回的信息的过滤器；

储存在另一计算机可读介质中的统一应用程序编程接口，用于将用来与该组数据库通信的协议变换成一种协议；以及

爬寻器，所述爬寻器使用所述一种协议来调用所述统一数据结构中描述的方法，来探查该组数据库中的期望信息片断，而无需知道对该组数据库中的每一数据库专用的协议。

2．如权利要求 1 所述的系统，其特征在于，所述统一数据结构包括系统下的一个或多个实体，每一系统包括来自该组数据库的一个数据库，每一实体向该组数据库中的一个或多个数据库中的信息片断施加结构。

3．如权利要求 2 所述的系统，其特征在于，所述统一数据结构包括实体下的一个或多个方法，每一方法可由所述统一数据结构来描述，以允许所述爬寻器调用并探查数据库中的期望信息片断。

4．如权利要求 3 所述的系统，其特征在于，所述统一数据结构包括方法下的一个或多个参数，一个或多个参数可由所述过滤器注释以约束要返回的信息片断。

5．如权利要求 4 所述的系统，其特征在于，所述过滤器包括过滤自从特定日期以来修改的信息片断的过滤器。

6．一种其上储存有数据结构的计算机可读介质，所述数据结构用于增强爬寻不同数据库中的信息片断的过程，所述数据结构包括：

指示要由爬寻器探查的数据库的系统字段、指示要给予储存在所述数据库中的信息片断的意义的实体字段、指示用于找出所述数据库中的实体字段的示例的可调用逻辑片断的方法、指示要由值来填充的变量的参数字段、以及指示约束要通过所述方法调用返回的信息片断的注释的过滤器字段。

7．如权利要求 6 所述的计算机可读介质，其特征在于，还包括所述实体字段的实例化，用于创建表示选自由顾客、订单号和邮政编码组成的组的信息的类别的结构。

8．如权利要求 6 所述的计算机可读介质，其特征在于，还包括所述方法字段

的实例化，用于创建在由所述爬寻器探查的数据库中找出特定行的方法。

9．如权利要求 6 所述的计算机可读介质，其特征在于，还包括所述方法字段的实例化，用于创建找出由所述爬寻器探查的数据库中的多个行的方法。

10．如权利要求 6 所述的计算机可读介质，其特征在于，还包括所述过滤器字段的实例化，用于创建过滤自从特定日期以来被修改的信息片断的过滤器。

11．一种用于探查数据库和应用程序文档以找出信息片断的计算机实现的方法，所述方法包括：

通过访问统一数据结构，以通过用一个或多个过滤器调用查寻器方法来探查每一数据库下的每一实体，来爬寻期望的信息片断，每一过滤器约束要返回的信息片断；以及

接收用户选中超链接时的选择，所述超链接与一标识符相关联以获取一个或多个数据库中涉及与所述标识符相关联的实体的信息片断。

12．如权利要求 11 所述的方法，其特征在于，还包括在所述爬寻动作之前填充所述统一数据结构的动作，以枚举要爬寻的一个或多个数据库、每一数据库下的实体、每一实体下的方法、每一方法下的参数以及与参数相关联的过滤器。

13．如权利要求 12 所述的方法，其特征在于，还包括索引所述数据库中的关键字段。

14．如权利要求 13 所述的方法，其特征在于，还包括将统一应用程序编程接口的方法调用转换成对数据库专用的应用程序编程接口的调用。

15．如权利要求 14 所述的方法，其特征在于，还包括过滤在一个或多个数据库中找到的信息片断，以符合与所述方法调用相关联的过滤器。

16．一种其上储存有指令的计算机可读介质，所述指令用于实现探查数据库和应用程序文档以找出期望的信息片断的计算机实现的方法，所述方法包括：

通过访问统一数据结构，以调用通过带有一个或多个过滤器的查寻器方法来探查每一数据库下的每一实体，来爬寻期望的信息片断，每一过滤器约束要返回的信息片断；以及

接收用户选中超链接时的选择，所述超链接与一标识符相关联以获取一个或多个数据库中涉及与所述标识符相关联的实体的信息片断。

17．如权利要求 16 所述的计算机可读介质，其特征在于，还包括在所述爬寻动作之前填充所述统一数据结构的动作，以枚举要爬寻的一个或多个数据库、每一数据库下的实体、每一实体下的方法、每一方法下的参数、以及与参数相关联的过

滤器。

18．如权利要求 17 所述的计算机可读介质，其特征在于，还包括索引所述数据库中的关键字段。

19．如权利要求 18 所述的计算机可读介质，其特征在于，还包括将统一应用程序编程接口的方法调用转换成对数据库专用的应用程序编程接口的调用。

20．如权利要求 19 所述的计算机可读介质，其特征在于，还包括过滤在一个或多个数据库中找到的信息片断以符合与所述方法调用相关联的过滤器。

爬寻数据库以找出信息

技术领域

本发明一般涉及搜索，尤其涉及爬寻（crawl）诸如应用程序文档和数据库等不同的信息源，以增强搜索引擎的关键词匹配。

背景技术

因特网 104 是全球使用 TCP/IP 协议套件来彼此通信的网络和网关的集合。在因特网 104 的中心是由数以千计的商业、政府、教育和其它计算机系统构成的主节点（或主计算机）以及路由数据和消息的其它计算机系统之间的高速数据通信的主干。一个或多个因特网 104 节点可在不危及整个因特网 104 的情况下离线或使得因特网 104 上的通信停止，因为没有单个计算机或网络控制它。因特网 104 的起源是由（美国）国防部在 1969 年创建的称为 ARPANET 的分散型网络，以便于在核攻击的情况下进行通信。当前，因特网 104 向用户提供了各种服务，诸如电子邮件和万维网。

大量互连的超文本文档 106 驻留在遍及世界的 HTTP 服务器上。构成万维网的这些文档称为万维网页，它们是以 HTML（超文本标记语言）来编写的，由指定可用于访问文件的特定机器和路径名的 URL（统一资源定位符）来标识，并通过 HTTP（超文本传输协议）从服务器发送到最终用户。这些万维网页可由搜索引擎 102 来搜索，搜索引擎 102 收集可用的万维网页的列表，并将这些列表储存在用户可按关键词搜索的数据库中。较老的搜索引擎的示例包括 Lycos 和 Excite。更新近的搜索引擎的示例包括 Google 和 A9。

万维网页易于在因特网 104 上搜索是因为它们中的许多是使用可由一致的目标 URL 访问并可通过公共协议 HTTP 来通信的公共语言 HTML 来编写的。内联网 108 上的搜索造成了通常因特网 104 上看不到的问题。内联网 108 是基于诸如 TCP/IP 等因特网协议的专用网络，但是是为公司或组织内的信息管理而设计的。其用途包括诸如文档分发、软件分发、数据库访问和训练等服务。称为内联网 108 是因为它看上去类似于万维网站且基于类似的技术，然而它对于组织仍是严格的内

部的，且不在严格意义上连接到因特网。在内联网 108 内可用的万维网页可由常规的搜索引擎 102 来搜索。典型地，诸如文档 112 等连接到内联网 108 的许多文档不是使用诸如超文本等公共语言来形成的，而是采用诸如 Microsoft Word、Microsoft Excel 等更特定的格式。诸如搜索引擎 102 等常规的搜索引擎无法搜索文档 112 内不是从诸如超文本等公共语言形成的信息片断。

类似的问题在搜索诸如数据库 110 等数据库时发生。在数据库中，数据不与诸如万维网页等文档相关联。但是在数据库 110 中储存了需要向内联网 108 的用户展示的信息片断。不幸的是，数据库 110 缺少用于搜索的任何组织良好的结构。搜索数据库 110 的另一问题是存在可用于搜索数据库 110 的许多协议。例如，搜索引擎 102 可使用查询语言来访问数据。然而，诸如数据库 110 等数据库通常不是直接访问的，而是搜索引擎 102 必须通过各种应用程序编程接口集。不像因特网 104，其中搜索引擎 102 必须仅知道一个协议（HTTP）来与万维网页通信以提取数据，搜索内联网 108 可要求搜索引擎 102 知道可能适合或不适合从诸如数据库 110 等数据库或应用程序文档 112 提取信息片断的多种协议。

尽管上文在内联网搜索的上下文中讨论了这些和其它问题，但是其它数据库搜索和文档搜索在通常与内联网相关联的不同种类的环境中也具有类似的（如果不是相同的）问题。如果没有对于有效地响应用户的内联网查询的问题的解决方案，用户可能最终不再信任搜索引擎 102 能够提供可再现储存的信息片断的期望的计算体验，且对于搜索引擎的需求将从市场上消失。由此，需要一种用于响应内联网查询，同时避免或减轻与现有系统相关联的上述和其它问题的系统、方法和计算机可读介质。

发明内容

依照本发明，提供了一种用于爬寻数据库和应用程序文档的系统、方法和计算机可读介质。本发明的系统形式包括用于探查不同数据库以找出信息的系统。该系统包括储存在计算机可读介质中用于枚举要探查的一组数据库的统一数据结构，以及用于约束返回的信息的过滤器。该系统还包括储存在另一计算机可读介质中用于将用来与该组数据库通信的协议变换成一种协议的统一应用程序编程接口。该系统还包括爬寻器，该爬寻器使用一种协议来调用以统一数据结构描述的方法以探查该组数据库中的期望信息片断，而无需知道对该组数据库中的每一数据库专用的协议。

依照本发明的另外的方面，本发明的计算机可读介质形式包括具有储存在其上用于增强在不同数据库中爬寻信息片断的过程的数据结构的计算机可读介质。该数据结构包括指示要由爬寻器探查的数据库的系统字段、指示要给予储存在数据库中的信息片断的意义的实体字段、指示用于在数据库中找出实体字段的实例的可调用逻辑片断的方法、指示要由值填充的变量的参数字段、以及指示约束要通过方法调用返回的信息片断的注释的过滤器字段。

依照本发明的另外的方面，本发明的方法形式包括用于探查数据库和应用程序文档以找出期望的信息片断的计算机实现的方法。该方法包括通过访问统一数据结构爬寻期望的信息片断，以通过调用具有一个或多个过滤器的查寻器来探查每一数据库下的每一实体。每一过滤器约束要返回的信息片断。该方法还包括当用户选中一超链接时接收该选择。该超链接与标识符相关联以获取一个或多个数据库中涉及与该标识符相关联的实体的信息片断。

附图说明

当结合附图参考以下详细描述时，可以更容易明白且更好地理解本发明的上述方面和许多附加优点，附图中：

图 1 是示出一个常规系统的框图，它示出了搜索应用程序文档和数据库的各种问题；

图 2 是示出用于爬寻数据库和应用程序文档以找出信息片断的示例性系统的框图；

图 3 是示出描述共同定义了储存在数据库和应用程序文档中的信息片断上的结构的项目组的广义类别的云图；以及

图 4A-4G 是示出依照本发明的一个实施例，用于在数据库和应用程序文档中爬寻信息的过程的流程图。

具体实施方式

图 2 中示出了其中可搜索连接到内联网的业务系统的系统。该系统是具有软件应用程序片断的联网计算环境，这些软件应用程序协作以允许爬寻器 200 寻找储存在业务系统中的信息片断，该业务系统是从数据库 208-212 以及应用程序文档（未示出）形成的。爬寻器 200 是搜索内联网以找出期望的信息片断的自动化应用程序。爬寻器 200 以允许由搜索引擎（未示出）检查匹配的方式索引数据库 208-212

和内容相关信息中的关键字段。

图 2 所示的系统提供了使得搜索引擎能够搜索数据库 208-212 中的信息片断，而无需知道用于与数据库 208-212 以及诸如 Microsoft Excel 文档等应用程序文档通信的专用协议的工具。尽管有爬寻器 200 为连接到数据库 208-212 必须使用的多种可能的协议，然而本发明的各实施例包括用于统一访问数据库 208-212 和应用程序文档中的信息片断的方式的统一应用程序编程接口 204。统一应用程序编程接口 204 提供了允许爬寻器 200 与数据库 208-212 通信，而无需知道用于与数据库 208-212 和应用程序文档通信的专用协议的转换。诸如数据库 208-212 等许多数据库理解结构化查询语言，该查询语言是用于查询、更新和管理数据库的常规数据库子语言。然而，数据库 208-212 通常具有可能模糊结构化查询语言的业务逻辑来从数据库 208-212 获取期望的信息片断。因此，每一数据库 208-212 具有对于特定数据库专用的可定制协议，用于查询特定数据库以找出信息片断。许多软件应用程序在对爬寻器 200 不可访问或不可见的嵌入式数据库中储存应用程序文档。如果符合适当的协议，则软件应用程序中的这些嵌入式数据库是可访问的。本发明的各实施例允许用户将关键词输入到搜索引擎中，且通过爬寻器 200，可找到数据库 208-212 中的期望的信息片断，并以使得这些信息片断可作为搜索结果返回给用户的方式来索引它们。

统一应用程序编程接口 204 连同统一数据结构 202 一起允许爬寻器 200 与数据库 208-212 通信，而无需知道用于与数据库 208-212 和应用程序文档通信的专用协议。即使数据库 208-212 的协议是已知的，然而数据库 208-212 的专用应用程序编程接口 208A-212A 也必须已知以获得数据库 208-212 内的期望信息片断。统一应用程序编程接口 204 连同统一数据结构 202 一起使得爬寻器 200 无需知道专用协议以及专用应用程序编程接口 208A-212A 来获得期望的信息片断。

如本领域的技术人员可以理解的，储存在数据库 208-212 中的信息片断可以相当大，其范围以万亿字节来计算。即使数据库 208-212 可以用极大的速度来处理各种信息片断，然而各种信息片断非常迅速地变得极其大。较佳的是，爬寻器 200 在关于它能够探查来索引数据库 208-212 中的多少信息片断方面受到约束。不感兴趣的传统数据或历史数据应当较佳地避免被爬寻器 200 探查。本发明的各实施例包括统一数据结构 202 中的过滤器或注释，且这些过滤器或注释提供了对于爬寻器 200 的探查的各种约束。当爬寻器 200 调用统一应用程序编程接口 204 时，统一应用程序编程接口 204 与统一数据结构 202 协商以理解用于探查数据库 208-212 以及

应用程序文档中的信息片断的约束。提供过滤器或注释的一种合适的技术是对数据描述语言的使用，该数据描述语言是专门为描述数据结构而设计的语言。

假定数据库 208 具有 API 108A，它可被调用以获得关于特定业务的顾客的信息片断。如果调用该 API，则数据库返回可被搜索的顾客的列表。本发明的各实施例用诸如最后修改日期等约束来过滤或注释对该 API 的调用的输入参数，以限制对数据库 208 的探查。爬寻器 200 可使用储存在统一数据结构 202 中的约束以找出比最后修改日期更新的信息片断。API 208A 返回对应于由统一数据结构 202 提供的约束（最后修改日期）的信息片断。由统一数据结构 202 储存的约束以及其它内容在本示例中向 API 208A 提供了意义。换言之，返回的信息片断已知是对于比最后修改日期更新的顾客信息。对过滤器或注释的使用允许爬寻器 200 更有效地探查。

在由爬寻器 200 探查了数据库 208 之后，信息片断按照字段名和字段值返回。每一字段具有某一意义。如果数据库 208 被实现为关系型数据库，则字段可以是标识特定行的唯一标识字段。本发明的各实施例可从这些唯一标识字段中创建超链接，以向用户呈现类别结果，从而允许作出到与信息片断相关联的值的连接。当用户选中一超链接时，与该特定超链接相连接的所有信息片断通过诸如浏览器等适当的应用程序呈现给用户。较佳地，该技术用于避免通过要求数据库 208-212 返回爬寻器 200 所探查的所有内容而使数据库 208-212 过载。统一数据结构 202 中的过滤器或注释以及将关键字段用作标识符防止一次返回所有的信息片断，以避免破坏由数据库 208-212 形成的计算系统。本发明的各实施例使用统一数据结构 202 中的过滤器和注释，并使用关键字段作为标识符，来将搜索请求分解成多个片断。当用户选中与一特定标识符连接的超链接时，则返回被指示为用户感兴趣的信息片断的一部分，其中该特定标识符进而与唯一地标识数据库 208 中的行的字段连接。

数据库 208-212 由记录组成，每一记录包含字段以及用于搜索、分类、重组和其它功能的一组操作。记录中的字段是唯一地标识被指定为关键字的一部分的信息片断的结构。这些关键字有助于在数据库 208-212 中检索信息片断的执行。本发明的各实施例提供了与这些关键字段相关联的描述或名称。当用户选中这些关键字段中的一个时，包含期望信息片断的适当记录被返回给用户。

统一数据结构 202 描述了表示爬寻器 200 可以探查的数据库 208-212 的系统的目录。换言之，统一数据结构 202 枚举了业务线系统的种类，诸如组织内的数据库 208-212，以允许爬寻器 200 探查。统一数据结构 202 中的过滤器或注释向数据库

208-212 中由爬寻器 200 探查到的信息片断提供了意义。通过统一应用程序编程接口 204，爬寻器 200 可与数据库 208-212 的任一个通信。爬寻器 200 只需与统一应用程序编程接口 204 通信。统一应用程序编程接口 204 知道适配器 206 中用于与数据库 208-212 通信的特定适配器。适配器 206 中所选择的适配器进而知道用于访问数据库 208-212 中的信息片断的特定 API。

适配器 206 通常是由数据库 208-212 的软件制造商提供的。然而，第三方软件制造商也可提供适配器 206。尽管通常适配器 206 不提供语义变换（或给予意义），然而统一应用程序编程接口 204 连同统一数据结构 202 一起提供了语义变换，因为如本领域的技术人员可以理解的，爬寻器 200 通常不期望理解储存在数据库 208-212 中的信息片断的语义。在探查过程中，爬寻器 200 看见信息片断的字段或表，而不理解这些信息片断意味着什么。诸如“最后修改日期”等由统一应用程序编程接口 204 和统一数据结构 202 提供的语义变换使得爬寻器 200 能够更智能地搜索。。

图 3B 示出了更好地定义统一数据结构 202 的云图 300。系统 300 示出了云之间的关系，每一云是描述一组更具体项目的广义类别。见图 3。云是程序中用于定义表征云的任何成员的一组属性和/或一组服务（对程序的其它部分可用的动作）的描述性工具。本质上，每一云定义了它所包括的实体的类型以及其实体运作的方式。每一云可使用合适的编程技术来实现，诸如面向对象的程序或关系型数据库表。

系统云 302 定义了表示包含与业务有关的信息片断的业务线的属性和服务。例如，数据库 208-212 的每一个是可由系统类 302 表示的系统。系统 302 的实例包括 SAP 系统 302A，它定义了与 SAP 系统有关的属性和服务。从 SAP 系统 302A 发出且终止于系统云 302 处的箭头形图的边指示在 SAP 系统 302A 的实例化和系统云 302 之间存在对应关系。

实体云 304 定义了与施加在诸如数据库 208-212 或 SAP 系统 302A 等系统中的信息片断上的结构有关的属性和服务。例如，假定 SAP 系统 302A 是使用关系型数据库来实现的。使用实体云 304 的适当实例化，爬寻器 200 无法看见关系型数据表，但是可以看见 SAP 系统 302A 中的顾客行。从实体云 304 发出且终止于系统云 302 处的箭头形图的边指示与系统云 302 的一对多关系（换言之，一个系统和一个系统中的多个实体的关系）。

顾客实例 304A 定义了与购买商品或服务的顾客的表示有关的属性和服务。顾客实例 304A 可用于对 SAP 系统 302A 中的信息片断施加意义和结构，以指示包含

在它们中的某些信息片断是涉及顾客的信息。由实体云 304 表示的每一实体是对于存在于由系统云 302 表示的系统，诸如 SAP 系统 302A 中的信息片断上的结构的描述。从顾客实体 304A 发出且终止于实体云 302 处的箭头形图的边指示在顾客实体 304A 的实例化和实体云 302 之间存在对应关系。

每一实体云 304 具有由方法云 306 表示的特征。方法云 306 定义了表示实体可具有的行为的属性和服务。从方法运 306 发出且终止于实体云 304 处的箭头形图的边指示与实体云 304 的一对多关系（换言之，一个实体与一个实体内的多个方法的关系）。

查寻器方法 306A 提供了在系统内查找实体的能力。例如，查寻器方法 306A 可用于找出作为诸如 SAP 系统 302A 等系统内的实体的顾客。较佳地，爬寻器 200 使用系统内的实体的特定查寻器方法以找出该实体的所有实例。从查寻器方法 306A 发出且终止于方法云 306 处的箭头形图的边指示查寻器方法 306A 和方法云 306 之间存在对应关系。

每一方法具有一个或多个参数。这些参数是由参数云 308 表示的，参数云 308 定义了表示方法的参数的属性和服务。从参数云 308 发出且终止于方法云 304 处的箭头形图的边指示与方法云 306 的一对多关系（换言之，一个方法与一个方法中的多个参数的关系）。

一种类型的参数是由过滤器云 308A 表示的过滤器。过滤器云 308A 定义了注释或限制由爬寻器 200 通过调用诸如查寻器方法 306A 等方法来探查和返回的信息片断的属性和服务。从过滤器云 308A 发出且终止于参数云 308 处的箭头形图的边指示与参数云 308 的一对多关系（换言之，一个参数与一个参数中的多个过滤器的关系）。

一个较佳的过滤器由“自从……以来修改”过滤器 308B 来表示，过滤器 308B 定义了与限制具有超过特定点的修改日期的信息片断有关的属性和服务。从“自从……以来修改”过滤器 308B 发出且终止于过滤器云 308A 处的箭头形图的边指示在“自从……以来修改”过滤器 308B 的实例化和过滤器云 308A 之间存在对应性。过滤器云 308A 可被实例化以从诸如 SAP 系统 302A 等系统中注释或过滤期望的信息片断。

较佳地，诸如查寻器方法 306A 等方法的名称及其签名，诸如参数云 308 的实例和过滤器 308A 的实例（即，“自从……以来修改”过滤器 308B）在文本上描述并被储存在统一数据结构 202 中。方法云 306 是由参数云 308 表示的参数的集合。

由方法云 306 表示的方法集合是由实体云 304 表示的实体的特征或行为。系统云 302 的实例、实体云 304 的实例以及方法云 306 的实例共同指向系统中储存的逻辑片断，诸如数据库 208-212 的 API 208A-212A，它们可被调用以获得期望的信息。

假定用户向浏览器的搜索输入区中输入搜索表达式“98829”。搜索串可意味着多种内容，诸如订单号、顾客号、邮政编码等等。搜索引擎和爬寻器 202 都不知道该搜索表达式意味着什么。实体类 304 提供了统一数据结构 202 中的语义，以允许返回的数据有意义。例如，爬寻器 200 可返回 SAP 系统 302A 中其中可找到搜索表达式“98829”的实体的类别。这些类别可包括订单号、顾客、邮政编码等等。搜索表达式可出现在储存在 SAP 系统 302A 或诸如数据库 208-212 等其它系统中的任何字段中。为获得系统外的信息片断，用户需要通过选中表示特定实体的超链接来指示。从这一选择中，爬寻器 200 知道要调用的具体方法或查寻器方法，以从 SAP 系统 302A 中提取期望的信息片断。

本发明的各实施例替代了编写直接与对诸如数据库 208-212 等特定系统专用的协议交互的新协议处理程序的需求。相反，使用统一应用程序编程接口 204 和统一数据结构 202，爬寻器 200 需要直接与统一应用程序编程接口 204 交互，而无需知道数据库 208-212 所需的协议的具体细节。与数据库 208-212 有关的具体细节在某种意义上由本发明的各实施例变换成单个统一协议。更具体地，变换是由爬寻器 200 对特定的实体调用查寻器方法。为找出信息，爬寻器 200 只需找出用户感兴趣的特定实体的期望的查寻器方法。爬寻器 200 无需知道其它复杂的协议来从数据库 208-212 中提取期望的信息片断。

统一数据结构 202 文本上包含由系统云 302 定义的各种系统、由实体云 304 定义的各种实体、由方法云 306 定义的各种方法（诸如查寻器方法 306A）、由参数云 308 定义的各种参数、以及由过滤器云 308A 定义的各种过滤器（诸如“自从……以来修改”过滤器 308B）。较佳的是组织内具有诸如数据库 208-212 等系统的知识的信息技术（IT）工作者填充统一数据结构 202。在实现统一数据结构 202 时无需编写使用本发明的各实施例来的代码。较佳的是向组织中的信息技术工作者提供用户界面来填充统一数据结构 202。

可以有一个以上查寻器方法 306A。较佳地，一个查寻器方法允许找出诸如数据库 208-212 等系统内包含信息片断的特定行。另一合适的查寻器是找出包含期望的信息片断的多个行的查寻器。由方法云 306 表示的某些方法可具有由参数类 308 表示的复杂参数。为简化方法的调用，本发明的各实施例允许描述各方法的默认参

数并将其储存在统一数据结构 202 中。默认参数将自动被插入到方法调用中，而无需爬寻器 200 指定参数的细节。

图 4A-4G 示出了用于在数据库和应用程序文档中爬寻信息片断的过程 400，且为清楚起见，以下对过程 400 的描述参考了结合云图 300（图 3）和爬寻器 200（图 2）所示的各元素。从开始框中，过程 300 前进到一组在延续终端（“终端 A”）和退出终端（“终端 B”）之间定义的处理步骤 402。该组处理步骤 402 描述了用于填充统一数据结构 202 的过程。

从终端 A（图 4B），过程 400 前进到框 408，其中信息技术工作者文本上指定一个或多个系统。过程 400 前进到框 410，其中 IT 工作者文本上指定系统中的一个或多个实体。IT 工作者然后文本上指定与实体有关的一个或多个方法（即，查寻器方法）。见框 412。过程 400 然后流至框 414，其中 IT 工作者文本上指定与方法有关的一个或多个参数。接着，在框 416，IT 工作者文本上指定与方法有关的一个或多个过滤器。IT 工作者也文本上指定一个或多个默认参数。见框 418。该方法然后继续到退出终端 B。

从退出终端 B，过程 400 前进到在延续终端（“终端 C”）和退出终端（“终端 D”）之间定义的一组处理步骤 404。该组处理步骤 404 描述了对数据库和应用程序文档执行的爬寻过程。

从终端 C（图 4C），过程 400 前进到框 420，其中使用统一 API 的爬寻器仔细检查统一数据结构，并选择要爬寻的系统。接着，在框 422，使用统一 API 的爬寻器找出要爬寻的系统的实体（在统一数据结构内）。系统的一个示例是数据库 208。实体的另一示例是顾客实体 304A。使用统一 API 的爬寻器找出实体的查寻方法。见框 424。过程 400 然后前进到框 426，其中爬寻器使用统一 API 找出与查寻方法相关联的一个或多个参数。爬寻器使用统一 API 找出一个或多个默认参数。见框 428。过程 400 然后继续到框 430，其中爬寻器使用统一 API 找出一个或多个过滤器，以约束爬寻器的截止器。接着，过程 400 继续到另一延续终端（“终端 C1”）。

从终端 C1（图 4D），爬寻器使得统一 API 用由一个或多个过滤器约束的一个或多个默认参数调用方法。接着，在框 434，统一 API 将方法和参数转换成带有可由诸如数据库 208 等特定系统的适配器理解的参数的另一方法。适配器将方法调用转换成可由诸如 API 208A 等其它 API 系统理解的协议。见框 436。过程 400 然后继续到框 438，其中适配器将方法调用转换成可由能找出期望信息片断的系统的

特定 API 理解的协议。过程 400 然后继续到另一延续终端（“终端 C2”）。

从终端 C2（图 4E），过程 400 前进到判别框 442，其中执行测试以确定爬寻器是否使用所选择的方法完成了其爬寻。如果在判别框 442 处测试的回答是“否”，则过程 400 循环回来以再次执行判别框 442 处的测试。否则，如果回答是“是”，则过程 400 继续到框 444，其中方法索引唯一地标识包含期望信息片断的行的关键字段（以及关键字段的描述）。在判别框 446 执行另一测试以确定在系统内是否还有更多实体要爬寻。如果对判别框 446 处的测试的回答是“是”，则过程前进到另一延续终端（“终端 C3”）。从终端 C3（图 4C），过程 400 循环回到框 422，其中重复上述处理步骤。否则，对判别框 446 的测试的回答是“否”，且过程 400 前进到另一延续终端（“终端 C4”）。

从终端 C4（图 4F），过程 400 前进到判别框 448，其中执行测试以确定是否还有更多系统要爬寻。如果对测试的回答是“是”，则方法继续到另一延续终端（“终端 C5”）。从终端 C5（图 4C），该过程循环回到框 420，其中重复上述处理步骤。否则，对判别框 448 处的测试的回答是“否”，且过程 400 前进到退出终端 D。

从退出终端 D，过程 400 前进到在延续终端（“终端 E”）和退出终端（“终端 F”）之间定义的一组处理步骤 406。该组处理步骤 406 选择性地显示了由爬寻器找到的信息片断的各个身份。从终端 E（图 4F），用户通过浏览器输入搜索表达式。见框 450。接着，在框 452，过程 400 收集系统中包含搜索表达式内容的实体。过程 400 然后继续到另一延续终端（“终端 E1”）。

从终端 E1（图 4G），过程 400 前进到框 455，其中该过程向用户呈现包含搜索表达式的内容的实体的描述。接着，在框 456，用户选择特定实体。使用索引的统一 API 提取涉及搜索表达式的内容以及特定系统的特定实体的实际信息片断。过程 400 然后前进到框 460，其中在浏览器上向用户呈现信息片断。过程 400 然后继续到判别框 462，其中执行另一测试以确定是否有更多搜索要执行。如果对判别框 462 处的测试的回答是“是”，则过程 400 继续到终端 E，并循环回到图 4F，其中重复上述处理步骤。否则，对判别框 462 处的测试的回答是“否”，则过程 400 继续到退出终端 F 并终止执行。

尽管示出和描述了本发明的较佳实施例，然而可以理解，可在其中作出各种改变而不脱离本发明的精神和范围。

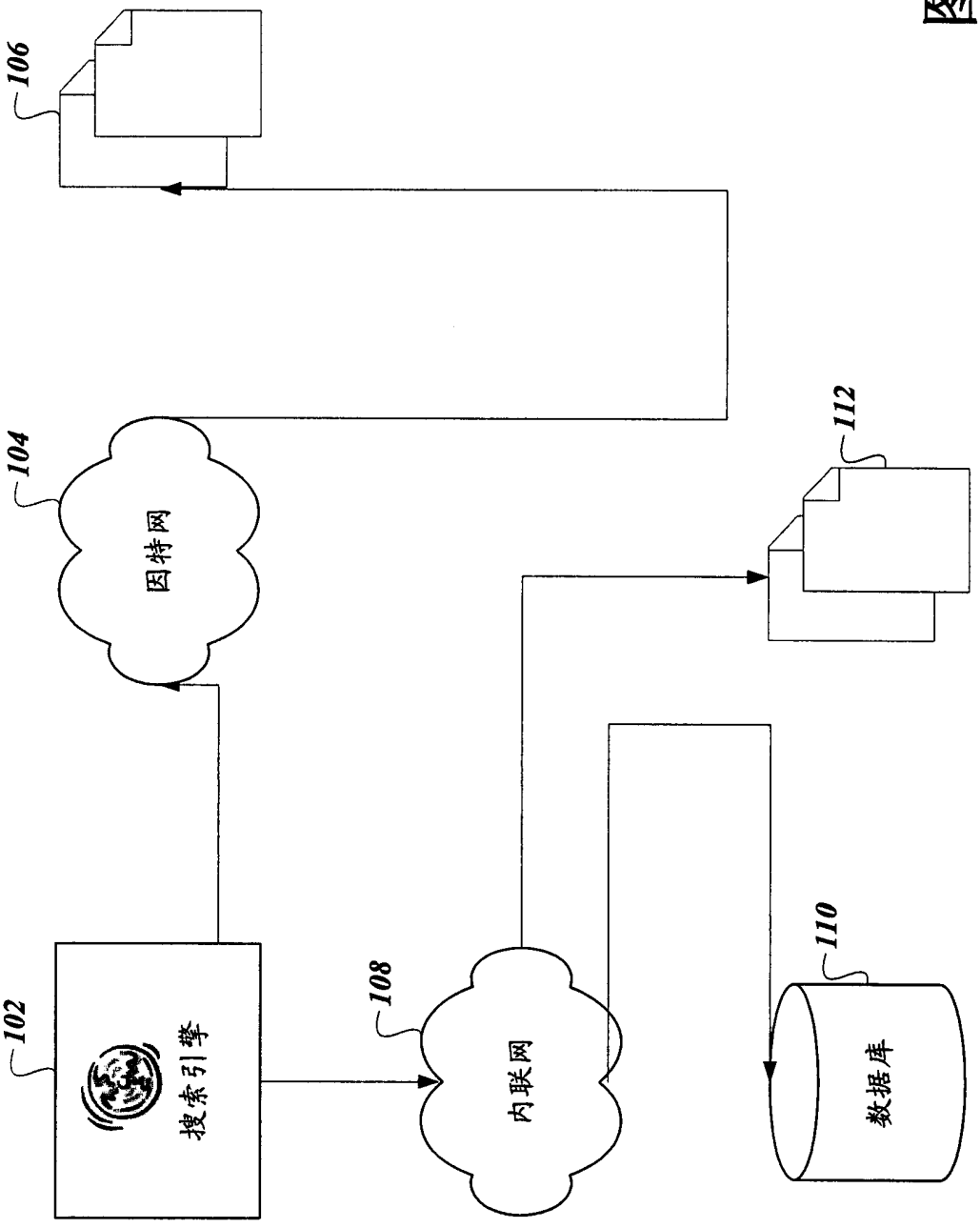


图 1

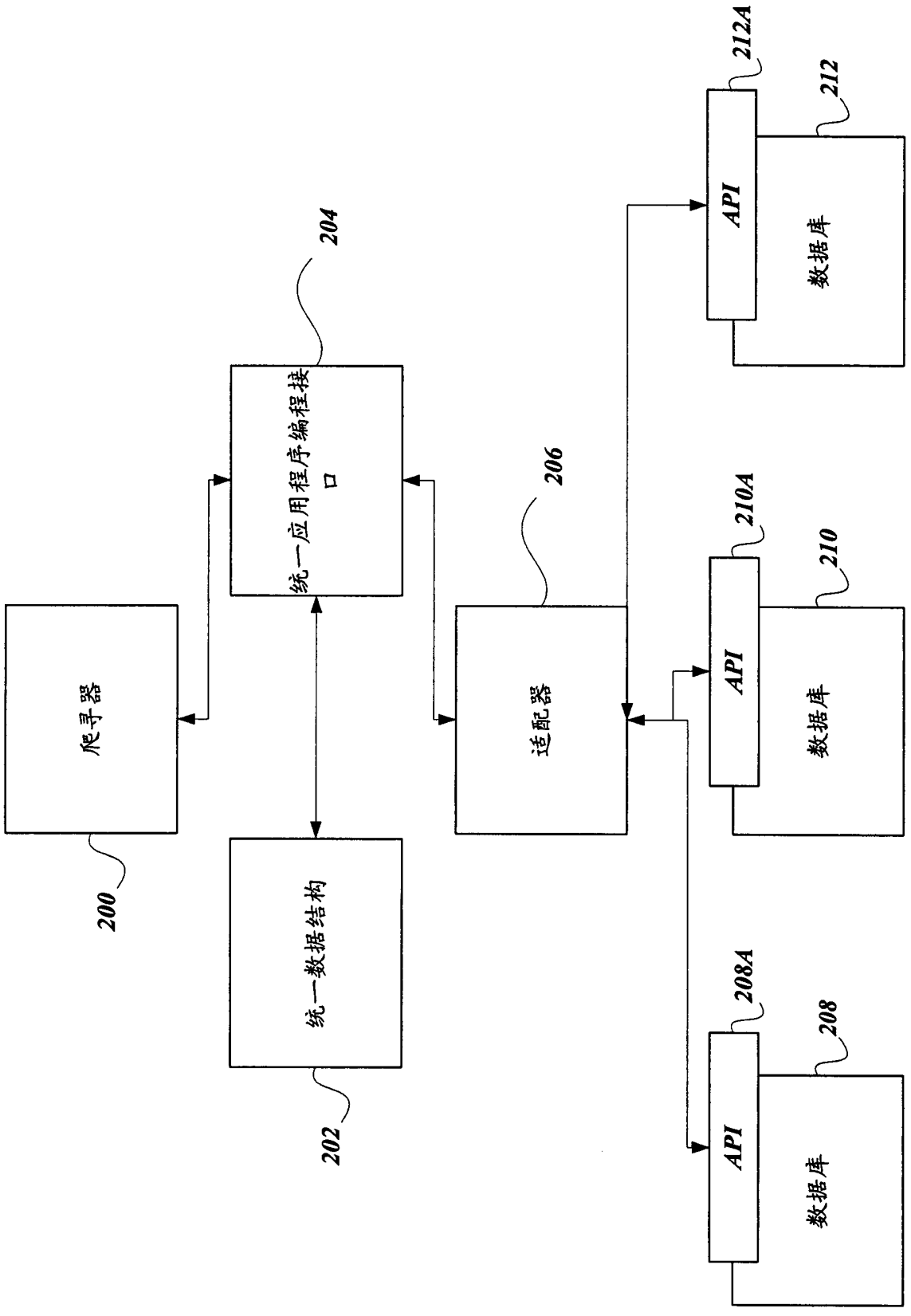

爬寻器
200
统一应用程序编程接口
204
统一数据结构
202
适配器
206
API
212A
数据库
212
API
210A
数据库
210
API
208A
数据库
208

图 2

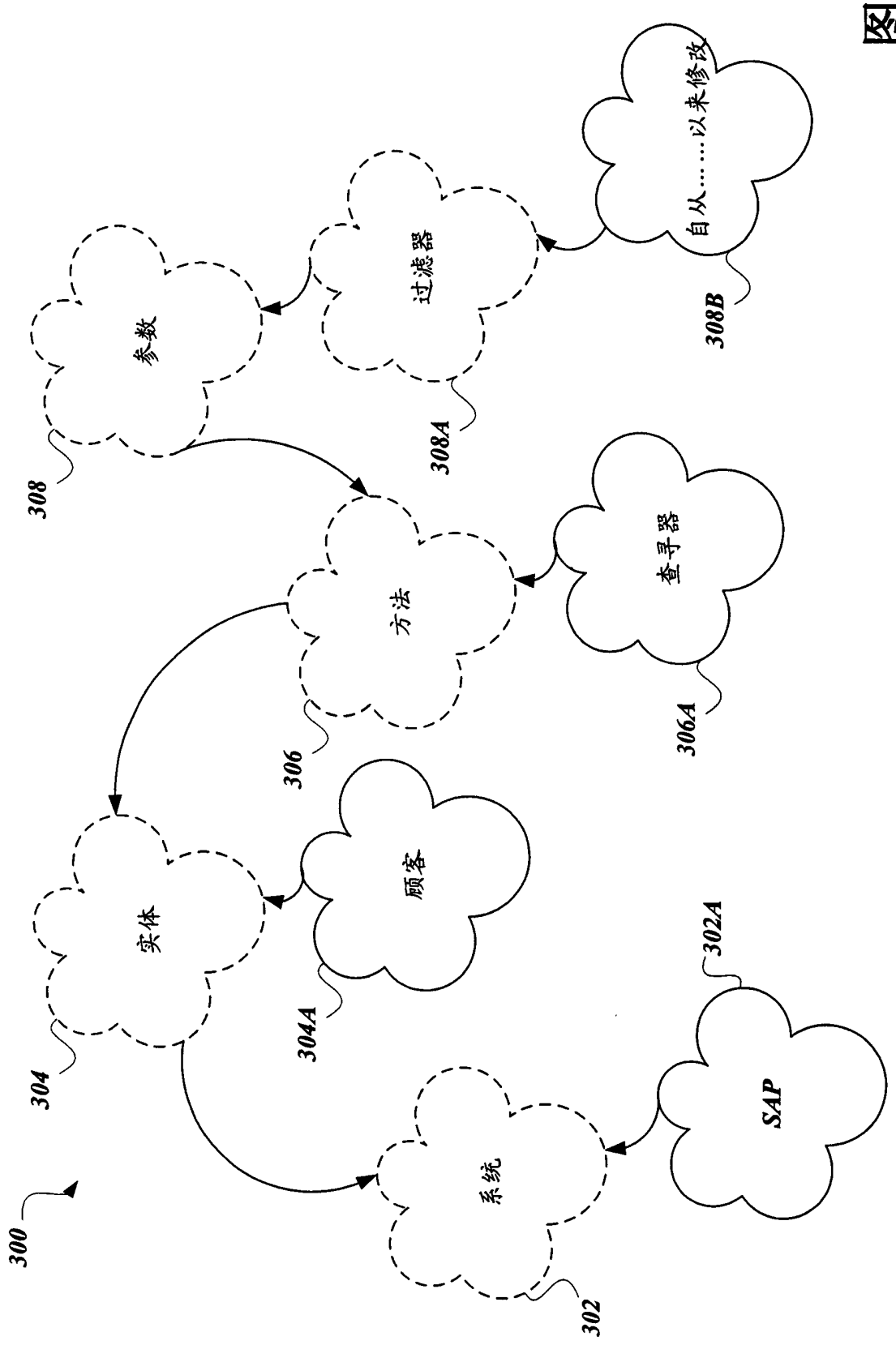

300
系统
302
SAP
302A
实体
304
顾客
304A
方法
306
查寻器
306A
参数
308
过滤器
308A
自从……以来修改
308B

图 3

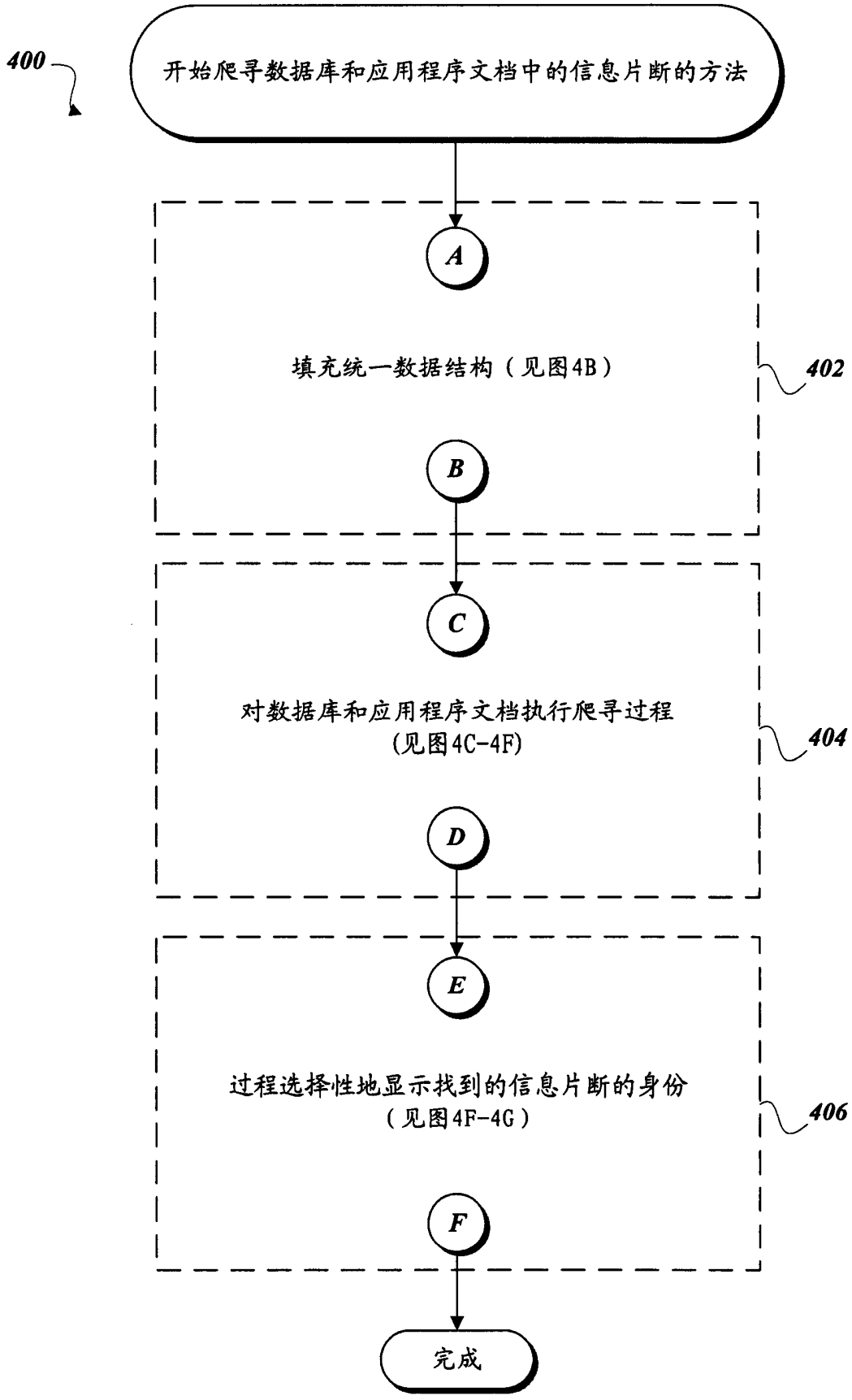

400
开始爬寻数据库和应用程序文档中的信息片断的方法
A
填充统一数据结构（见图4B）
402
B
C
对数据库和应用程序文档执行爬寻过程
(见图4C-4F)
404
D
E
过程选择性地显示找到的信息片断的身份
（见图4F-4G）
406
F
完成

图 4A

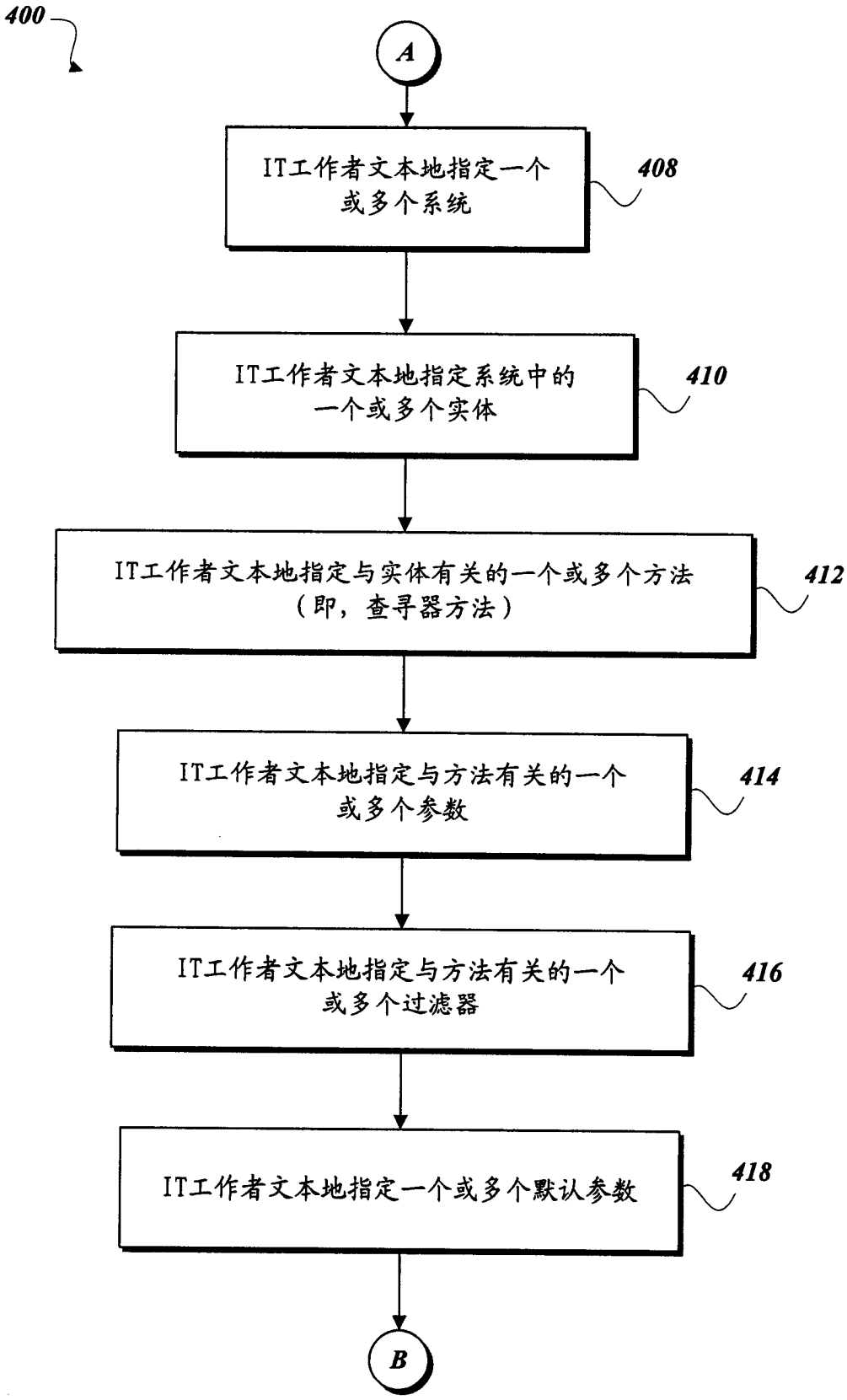

400
A
IT工作者文本地指定一个或多个系统
408
IT工作者文本地指定系统中的一个或多个实体
410
IT工作者文本地指定与实体有关的一个或多个方法（即，查寻器方法）
412
IT工作者文本地指定与方法有关的一个或多个参数
414
IT工作者文本地指定与方法有关的一个或多个过滤器
416
IT工作者文本地指定一个或多个默认参数
418
B

图 4B

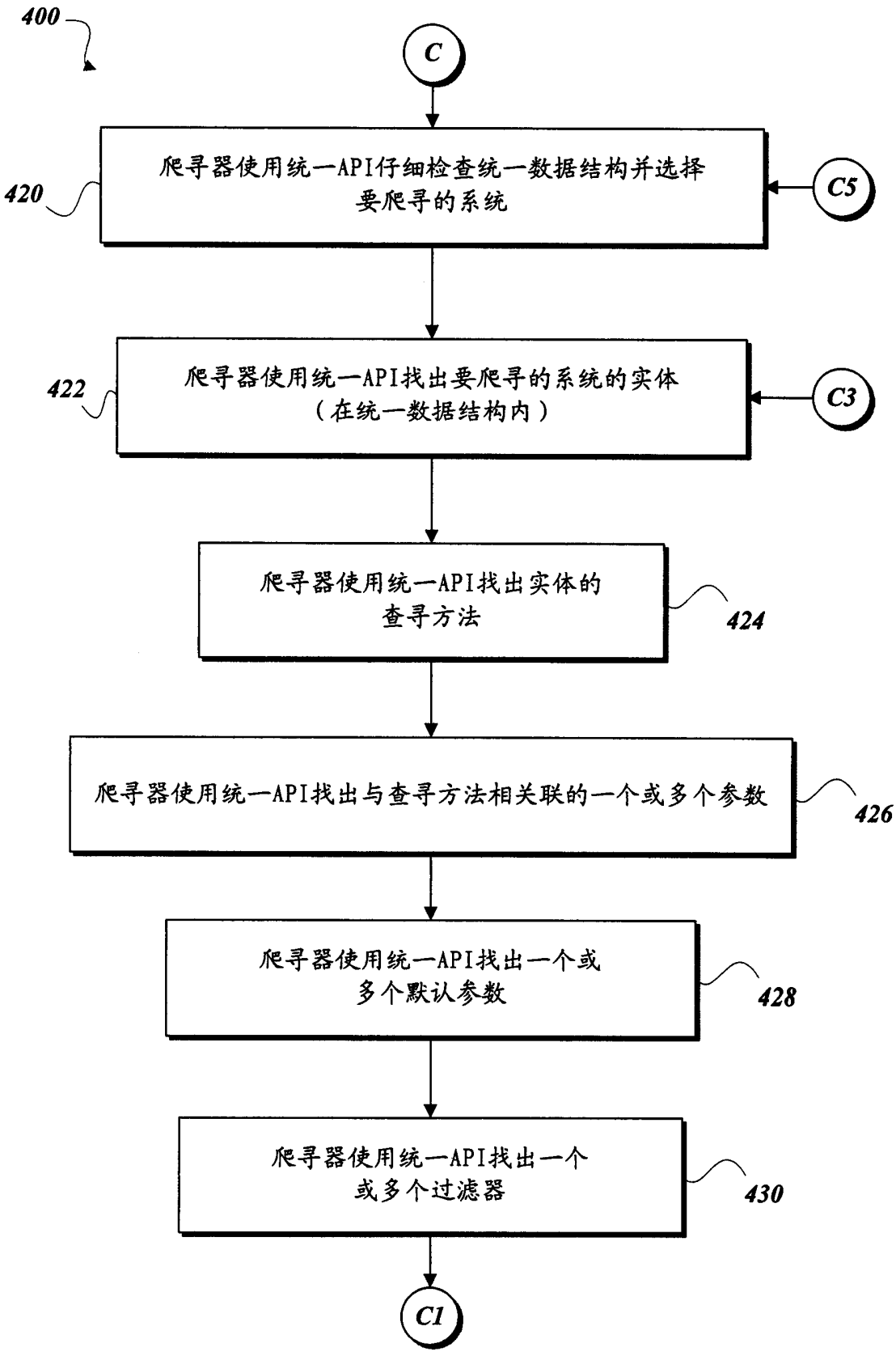

400
C
爬寻器使用统一API仔细检查统一数据结构并选择要爬寻的系统
420
C5
爬寻器使用统一API找出要爬寻的系统的实体（在统一数据结构内）
422
C3
爬寻器使用统一API找出实体的查寻方法
424
爬寻器使用统一API找出与查寻方法相关联的一个或多个参数
426
爬寻器使用统一API找出一个或多个默认参数
428
爬寻器使用统一API找出一个或多个过滤器
430
C1

图 4C

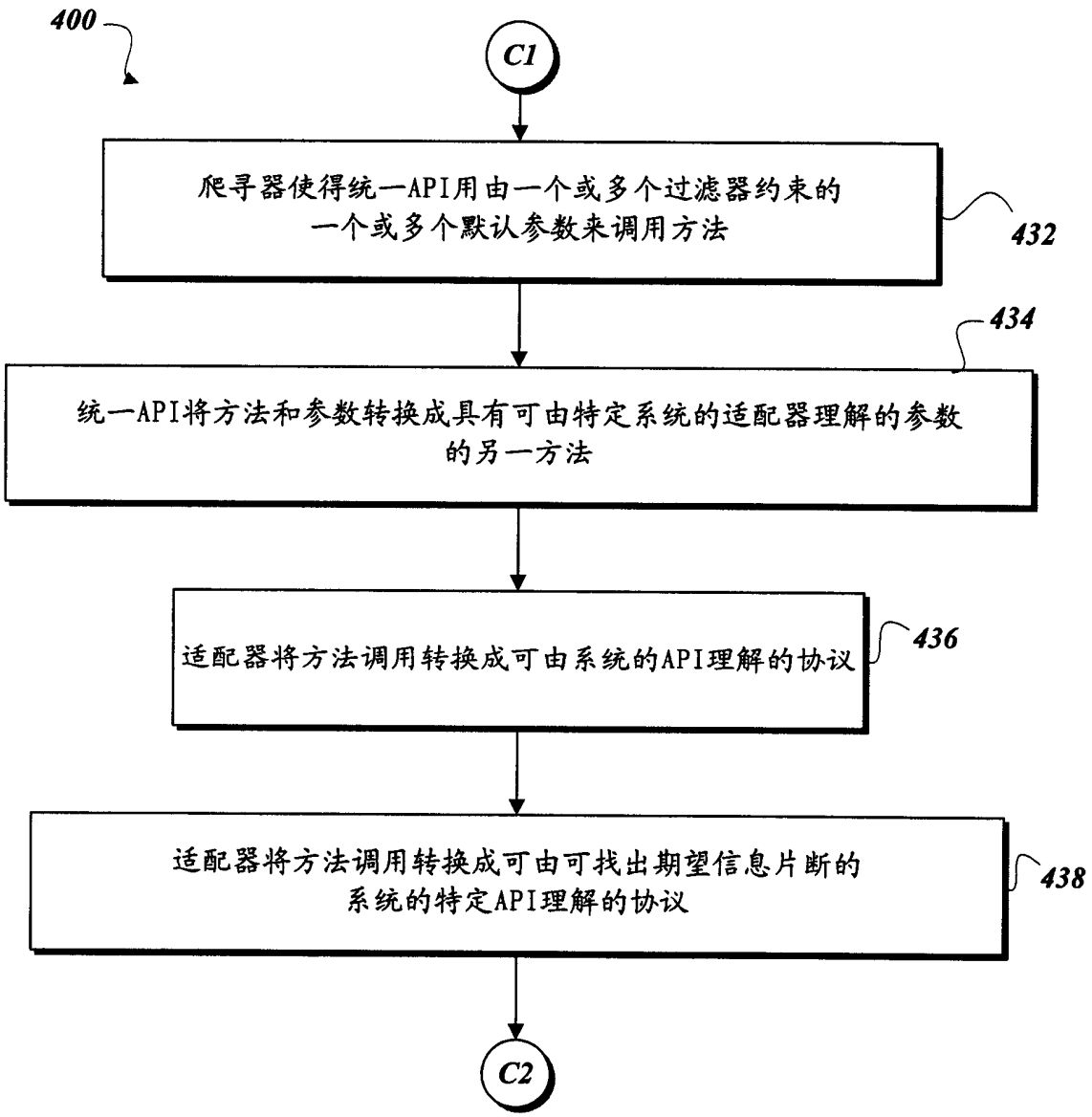

400
C1
爬寻器使得统一API用由一个或多个过滤器约束的
一个或多个默认参数来调用方法
432
434
统一API将方法和参数转换成具有可由特定系统的适配器理解的参数
的另一方法
适配器将方法调用转换成可由系统的API理解的协议
436
适配器将方法调用转换成可由可找出期望信息片断的
系统的特定API理解的协议
438
C2

图 4D

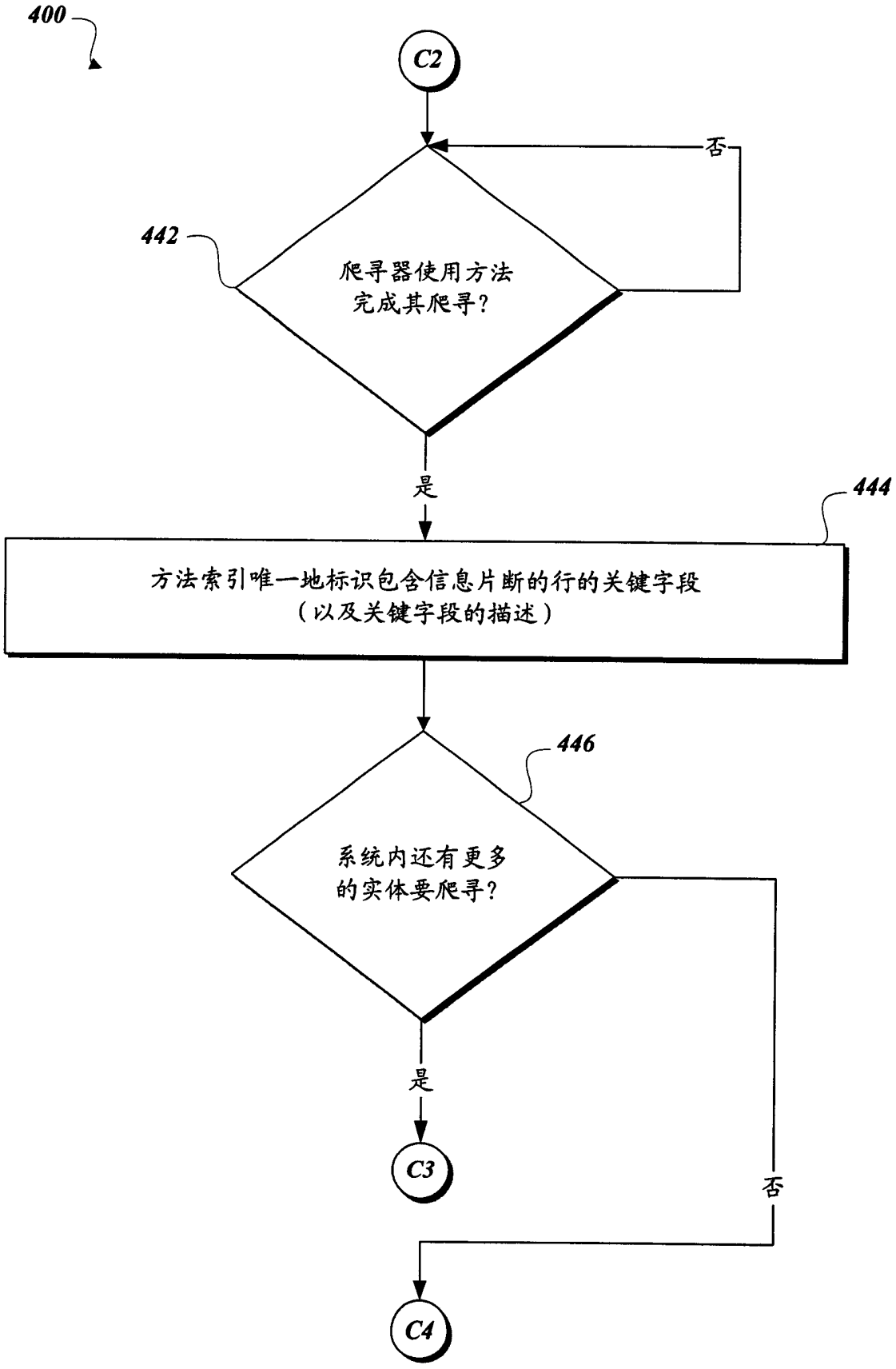

400
C2
否
442
爬寻器使用方法
完成其爬寻?
是
444
方法索引唯一地标识包含信息片断的行的关键字段
(以及关键字段的描述)
446
系统内还有更多
的实体要爬寻?
是
C3
否
C4

图 4E

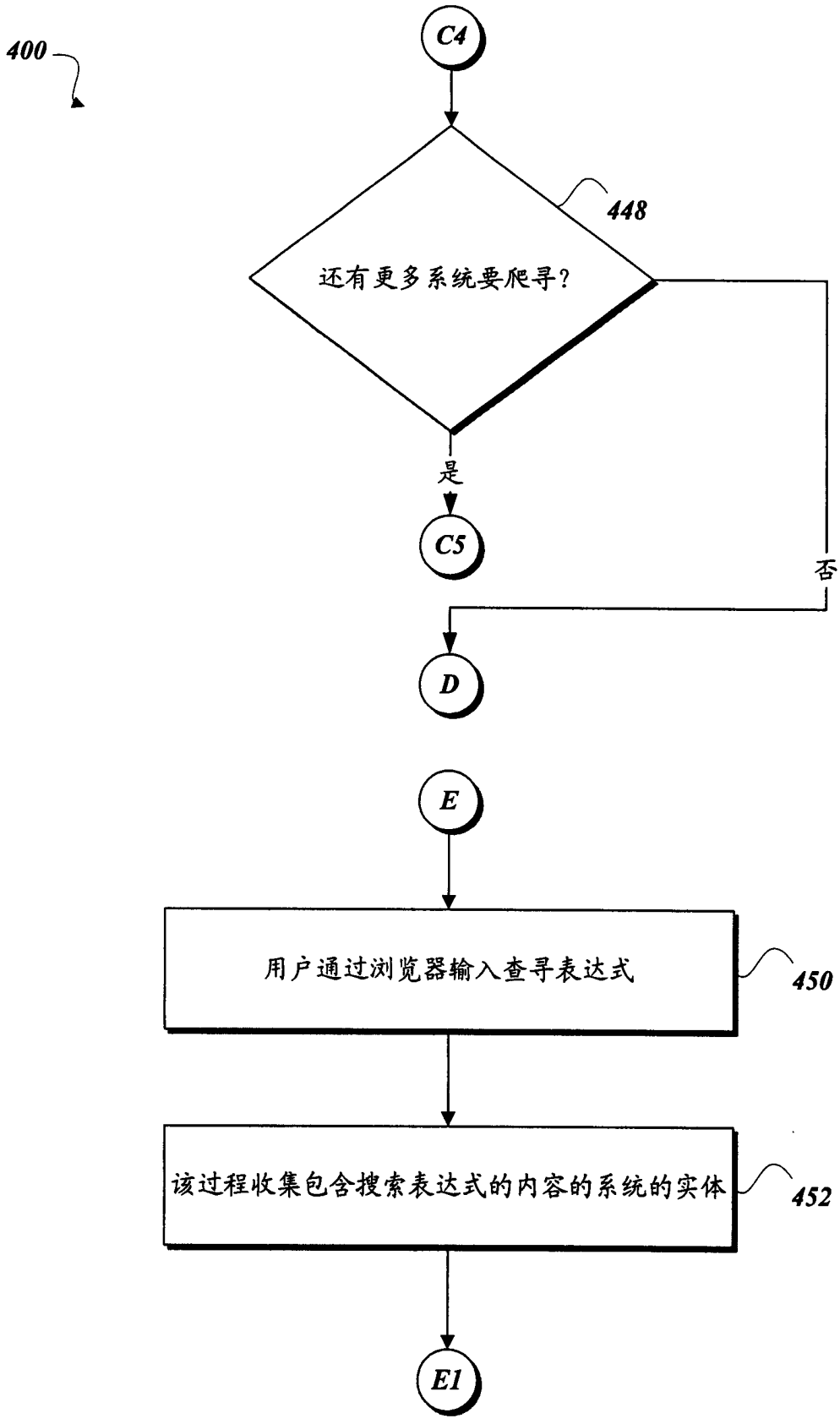

400
C4
448
还有更多系统要爬寻?
是
C5
否
D
E
用户通过浏览器输入查寻表达式
450
该过程收集包含搜索表达式的内容的系统的实体
452
E1

图 4F

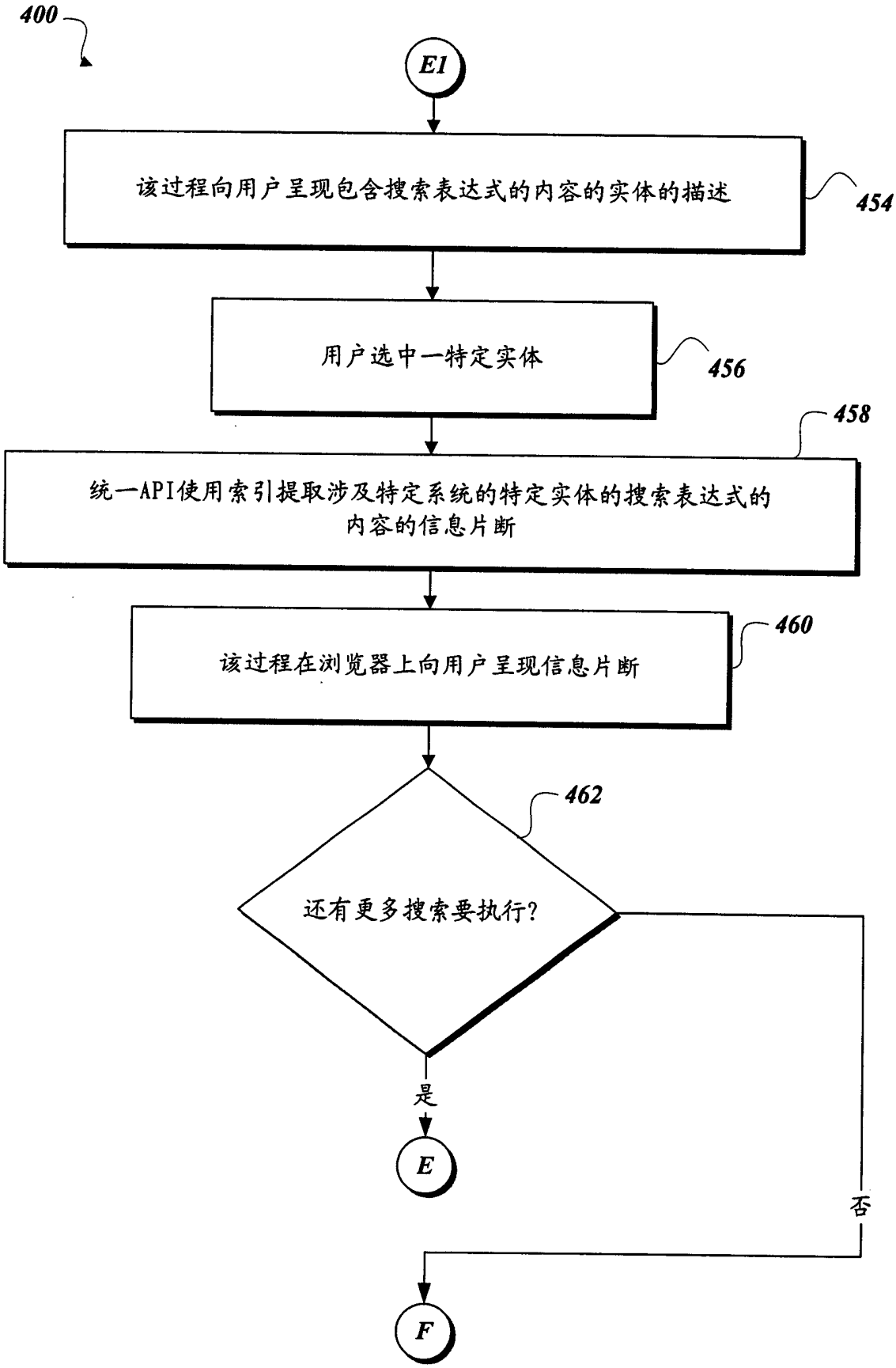

400
E1
该过程向用户呈现包含搜索表达式的内容的实体的描述
454
用户选中一特定实体
456
458
统一API使用索引提取涉及特定系统的特定实体的搜索表达式的内容的信息片断
该过程在浏览器上向用户呈现信息片断
460
462
还有更多搜索要执行?
是
E
否
F

图 4G