[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 200510119370.8

[43] 公开日 2006 年 7 月 12 日

[11] 公开号 CN 1801147A

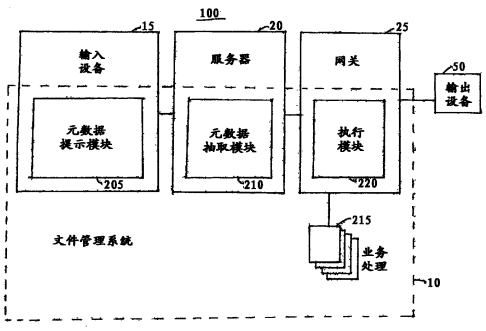
[22] 申请日 2005.11.2
[21] 申请号 200510119370.8
[30] 优先权
[32] 2004.11.3 [33] US [31] 10/980716
[71] 申请人 国际商业机器公司
地址 美国纽约
[72] 发明人 维卡斯·克里什纳
萨维塔·斯里尼瓦桑
[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 李德山
权利要求书 3 页 说明书 16 页 附图 8 页

[54] 发明名称

用于自动和动态地构建文件管理应用程序的方法和系统

[57] 摘要

一种文件管理系统，应用相关文件分析、元数据抽取、业务处理相关算法和方法来自动、动态地分类文件，以进行路由、处理和执行定制业务逻辑。 该文件管理系统从一个或多个通道接受文件，分类该文件，抽取元数据，执行定制的应用文档并触发与该处理相关的业务逻辑。 该文件管理系统包括一个规则引擎，用来检测和分类非结构化形式和结构化形式，这里属性的位置和视觉布局的位置是不固定的。 该文件管理系统提供管理文件的分立系统间的自动链接，用于完全执行业务处理。



1、一种自动和动态地为一个文件构建多个文件管理应用程序的方法，包括:

从输入设备接收该文件;

自动获取多个与该文件有关的用户提示的元数据;

自动从该文件抽取多个抽取的元数据;

对该文件、用户提示的元数据、抽取的元数据和分区数据元素中的任何一个或多个执行业务处理，以获得业务处理执行结果;

若需要，则自动地通知用户，该文件、用户提示的元数据和抽取的元数据中的任何一个或多个需要验证; 以及

将该文件、用户提示的元数据和抽取的元数据集成到输出设备。

2、如权利要求 1 所述的方法，还包括自动地对该文件执行选择性抽取，以产生分区数据元素。

3、如权利要求 1 所述的方法，还包括将该文件、用户提示的元数据和抽取的元数据中的任何一个或多个提供给用户进行验证。

4、如权利要求 1 所述的方法，还包括自动地提示用户提供所述多个用户提示的元数据。

5、如权利要求 1 所述的方法，其中该文件包括纸件文件、电子文件、视频记录、音频记录、照片和数字照片中的任何一个或多个。

6、如权利要求 1 所述的方法，其中验证包括多个用户提供的增加的数据。

7、如权利要求 2 所述的方法，其中自动地抽取所抽取的元数据的步骤包括对文件执行光学字符识别。

8、如权利要求 7 所述的方法，其中自动地执行选择性抽取的步骤包括在该文件的特定部分执行光学字符识别。

9、如权利要求 8 所述的方法，其中该文件的特定部分是所述业务处理确定的。

10、如权利要求 9 所述的方法，其中业务部分被用户改动，以改

变文件的特定部分的位置。

11、如权利要求 1 所述的方法，其中该输入设备包括扫描仪、计算机、打印机和其中可以在本地阅览器及应用程序中的任何一个或多个中阅览文件的设备中的任何一个或多个。

12、如权利要求 1 所述的方法，其中集成的步骤包括创建、更新、删除和查询中的任何一个或多个。

13、如权利要求 1 所述的方法，其中集成的步骤包括与外部系统接口进行接口。

14、如权利要求 13 所述的方法，其中该外部系统接口是由所述业务处理指定的。

15、如权利要求 1 所述的方法，其中用户通过所述业务处理指定用于验证的验证接口。

16、如权利要求 1 所述的方法，其中用户通过所述业务处理指定用于通知的通知接口。

17、如权利要求 1 所述的方法，其中用户通过所述业务处理指定对输出设备的访问。

18、如权利要求 1 所述的方法，其中所述业务处理在分布环境中操作。

19、如权利要求 1 所述的方法，其中所述业务处理在包括结构化格式的文件中规定。

20、如权利要求 1 所述的方法，其中所述业务处理在包括半结构化格式的文件中规定。

21、一种自动和动态地为一个文件构建多个文件管理应用程序的方法，包括:

为用户提供识别和修改业务处理的装置;

调用自动文件管理实用程序，其中所述文件和所述业务处理是所述自动文件管理实用程序可用的;

为用户提供验证文件和相关数据的装置;

产生验证的文件和多个验证的数据;

为用户提供增加文件和相关数据，和产生多个增加的数据的装置；并且

其中验证的文件、验证的数据和增加的数据由自动文件管理实用程序根据所述业务处理的指示来处理。

22、一种计算机程序产品，具有多个可执行指令码，用于自动和动态地为一个文件构建多个文件管理应用程序，包括：

第一组指令码，用于从输入设备接收该文件；

第二组指令码，用于自动地获取多个与该文件相关的用户提示的元数据；

第三组指令码，用于自动地从该文件抽取多个抽取的元数据；

第四组指令码，用于对所述文件、用户提示的元数据、抽取的元数据和分区数据元素中的任何一个或多个执行业务处理，以获得业务处理执行结果；

若需要，还包括第五组指令码，用于自动地向用户通知，所述文件、用户提示的元数据和抽取的元数据中的任何一个或多个需要验证；和

第六组指令码，用于将所述文件、用户提示的元数据和抽取的元数据集成到输出设备。

23、一种文件管理系统，用于自动和动态地为一个文件构建多个文件管理应用程序，包括：

产生该文件的输入设备；

用于自动地获取与该文件相关的多个用户提示的元数据的模块；

用于自动地从该文件抽取多个抽取的元数据的元数据抽取模块；

执行模块，用于对所述文件、用户提示的元数据、抽取的元数据和分区数据元素中的任何一个或多个执行业务处理以获得业务处理执行结果；

若需要，还包括通知模块，用于自动地通知用户，所述文件、用户提示的元数据和抽取的元数据中的任何一个或多个需要验证；

输出设备，用于将所述文件、用户提示的元数据和抽取的元数据集成到输出设备。

用于自动和动态地构建文件管理应用程序的方法和系统

技术领域

本发明总的涉及内容管理。更具体地，本发明涉及一种内容管理应用程序，其应用相关文件分析、元数据抽取和业务处理相关算法及方法来自动和动态地对文件进行分类，以便路由、处理和执行定制的业务逻辑。

背景技术

内容管理被定义为在任何介质中或以任何格式建立、组织、管理和存储数字作品集合的软件。内容管理是指处理各种类型的结构化和非结构化的信息的过程，该结构化和非结构化信息包括图像和文件，可包括帐单数据、用户服务信息和其它类型的内容。内容管理还指的是捕获、存储、分类、编码、集成、更新和保护任何和全部信息的过程。研究估算超过 75％的企业数据是非结构化和与文件相关的（Lyman Peter 等人著“多少信息，2000”,http://www.sims.berkeley.edu/how-much- info.）。

内容管理市场的关键技术包括文件管理、网络内容管理、数字资产管理和记录管理。内容管理的典型用户在文件量大的产业中，其中文件管理是基本需求，通常由于管理和服从的原因。内容包括许多不同形式的需要管理的非结构化数据：业务文件、动态网络内容、记录管理和丰富媒体。业务文件包括合同、发货单、表格和电子邮件。举例来说，业务文件能方便内部后－办公室处理及与用户、合作伙伴和供应商直接外部通信。动态网络内容包括有关数据库中的业务数据和个人化信息。记录管理典型地由政府和工业规范来驱动以便有效进行文件处理、审计索引和数据保留。丰富媒体包括数字音频和视频。丰

富媒体是许多产业中培训、教育、营销和用户关系管理中的快速变化领域。

将文件管理与工作流相关的概念已经通用了几十年，许多文件管理系统包括该特征。一个传统的方法用涉及为一个机器工具公司提供处理的案例研究来对集成文件和工作流管理的问题提供工具和方法（Morschheuser,S.,等著的"应用到机器工具公司的提供处理的集成文件和工作流管理"，有组织的计算系统会议公报，1995）。该传统方法为一种过程定义语言，使得带有工作流引擎的面向文件的工具更加高效。

另一传统方法将活动文件特性的思想利用到文件管理应用程序（Dourish, P., 等人著"利用用户特定活动特性延伸文件管理系统"，信息系统的 ACM 学报（TOIS），第 18 卷，第 2 期，2000）。该传统方法避免以前的分层存储机制，而反映对用户任务很有意义的文件分类，提供统一交互架构中一个或多个个体的想法的集成手段。基于特性的文件管理系统增加了活动特性的概念，以便在特性基础结构上提供基于文件的服务，该活动特性载有可执行码。

而另一传统的系统捕获基本的自由结构化文件，诸如典型地用于办公室领域中的文件（Mattos, N.M.,等人著"集成办公室文件处理和管理的方法"，ACM SIGOIS 公报，办公室信息系统会议公报，第 11 卷，第 2－3 期，1990）。该传统系统易于处理包含信息。分析过的文件存储在文件管理系统中，该文件管理系统连接到几个不同的后续服务并用作基本工作流。

FileNet 提供一种结合了文件技术的工作流引擎来分别自动操作制造和特别业务处理（Whelan, D "FileNet 集成文件管理数据库使用和问题"，ACM SIGMOD 记录，数据管理 1998ACM SIGMOD 国际会议学报，第 27 卷，第 2 期，1998）。

大多数传统文件管理系统由一相关模型支持。关于有关的关系模型研究，关系方案的正式模型化源自对运行时间方面的强调，诸如查询表达（Andries M 等人著"用于延伸的实体关系模型的混合查询语

言”，视觉语言和计算期刊，8（1），1997，视觉查询系统特刊；和 Angelaccio M 等人著“QBD*:完全视觉查询系统”，视觉语言和计算期刊，1（2），255－273，1990）、查询结果显示和对存储数据的导航。总的来说，这些任务称作视觉查询系统（VQS）(Catarci, T.,等人著“数据库的视觉查询系统：一个调查”技术报告 SI/RR-95/17, Dipartimento di Scienze dell'Informazione, Universita'di Roma “La Sapienza” 1995)。

对比来说，在用于定义和操作数据模型和数据库方案的工具所提供的接口方面，传统系统投入的注意力较少。商用数据库建模产品（例如 Ration 工具）提供视觉数据建模文档，其集成到更广泛的软件开发周期中（Gornik D, “UML 数据建模文档”，IBM Rational 软件白纸 TP 162 05/02,2003）。这些文档通常适应于关系数据库的 UML（统一建模语言）建模。由 Wisconsin 大学开发的 OPOSSUM 系统允许数据库方案通过方案可视化操作来编辑（Haber, E.M.等人著“OPOSSUM:灵活的方案可视化和编辑工具”，1994ACM CHI 会议公报，MA 波士顿，1994 年 4 月；Haber, E.M.等人著“Opossum:通过可定制的可视化的桌面方案管理”，于第 21 次国际 VLDB 会议公报，第 527－538 页，瑞典 Zurich,1995 年 9 月）。

文件管理系统典型地包括文件理解和分类的某些方面来支持业务处理。已经有人探索了分类机器打印的文件的通常问题，其中视觉布局是识别精细粒化类别的一个关键因素，这是因为文件内容特征相似。文件管理的一个传统方法利用从文件页的扫描二进制图像检测的布局结构，而不利用光学字符识别（OCR）结果，而是利用属性关系图（Bagdanov,A.D.,等人著“利用一阶随机图形的精细粒化文件分类”，ICDAR01 学报）。

另一传统系统在布局上利用基于“逻辑近似性”的学习技术，其中定向的权重图用于代表文件布局（Li,X.,等人著“带有学习能力的文件分类和抽取系统”，ICDAR99 学报）。而另一传统系统利用基于视觉相似性的文件分类（Hu, J 等人著“文件图像布局比较和分类”，

ICDAR99 学报）。在该传统系统中，引入间隔编码法来捕获空间布局的元素。这些传统系统提出基于隐马尔可夫模型的页面布局分类系统，该系统是可以基于空间特征可训练和延伸的。

另一传统系统利用面向用户的扫描图像部分的“快速捕获”，其包括易于访问、编辑、和分配到需要的目的地（如档案、应用程序和网页等）的工具（Simske,S.J 等人著“编辑和创作：面向用户的扫描图像的分析”，文件工程 2003ACM 论坛公报，2003）。这些工具利用面向用户的分区分析（公知为“点击与选择”）和基于统计的区域分类。“点击与选择”包含从下向上的分区分析引擎。基于统计的区域分类允许区域的快速重构。

虽然这些传统技术被证明是有用的，但还需要进一步的改进。文件管理应用程序的生命周期典型地包括以下阶段：

a) 内容的摄取（ingest）或捕获；

b) 管理（包括搜索、检取和工作流）；

c) 在业务过程结束时完成；以及

d) 由于服从和规定的原因而建档。

摄取或捕获阶段典型地产生关于进入文件的元数据，并将该文件与内容管理系统中定义的方案关联起来。相关于一个方案的元数据使得管理阶段能在业务处理和工作流的上下文中有效地搜索数据库。在完成了相关于处理的所有管理和事务之后，可以触发完成（fulfillment）动作，诸如通知、与其它系统（如记帐、支付、记录等）的集成。如果文件需要保留一固定的时间段用于审计，可以在断线存储器中建档。

传统的文件管理系统在分开的捕获子系统中管理摄取阶段，这些子系统使得元数据在分开的环境中规定。传统文件管理系统中应该管理的数据放置在不同的位置，如不同的业务分支、相对于主办公室的现场办公室等。随后文件被“释放”到文件管理系统。由于这些捕获子系统经常是从总的内容管理系统中分离出来的，所抽取的元数据被松散地连到方案和业务处理。其结果是，经常有相关于元数据的实际分配和相关于具体方案和处理的一个人工步骤，这导致总的上下文的效

率降低。例如，一个业务需要的数据典型地通常由人工成批地收集和处理。此外，在管理阶段之后，摄取阶段总是与业务处理的完成或触发没有联系。

因此需要一种系统、一种服务、一种计算机程序产品和相关的方法，用来自动、动态和有选择地构建（compose）和管理数据和文件。这种需求目前尚未得到满足。

发明内容

本发明满足这一需要提供一个系统、一种服务、一个计算机程序产品和一相关方法（这里统称为"该系统"或"本系统"），用于应用相关文件分析、元数据抽取、业务处理相关算法和方法来自动、动态和选择性地分类文件，以进行路由、处理和执行定制的业务逻辑。

本发明提供一种智能文件管理架构，具有相关文件分析、元数据抽取和业务处理相关算法和方法。本系统从一个或多个通道接受文件－扫描纸件、打印数据流、来自桌上电脑的电子文件，分类这些文件并抽取元数据，执行定制的应用文档并触发与该处理有关的业务逻辑。

本发明包括一个元数据提示模块、一个元数据抽取模块、业务处理过程、一个验证模块和一个执行模块。元数据提示模块安装在诸如扫描仪或打印机的输入设备中。当用户通过输入设备将一个文件输入到本系统中时，元数据提示模块通过一个或多个提示从用户请求关于该文件的信息。这些提示的形式可以是选择、按钮点击、文本输入等。在一个实施例中，元数据提示模块安装在具有元数据抽取模块的服务器上。元数据抽取模块自动从文件中抽取元数据。

执行模块安装在网关上。在一个实施例中，执行模块安装在带有元数据抽取模块的服务器上。执行模块恢复文件和来自服务器的相关元数据。执行模块如确定的文件和相关元数据那样，选择性地、自动地执行业务处理中的指令。

业务处理包括由执行模块执行的指令。这些指令逐文件地选择地被执行，逐文件基础是从文件分类确定的。用户可以对于每个文件类

型选择执行业务处理的哪个指令。进一步，用户可以在本系统操作时修正指令的选择而不改变执行模块的任何部分并且不关闭本系统或重启本系统。如相关元数据和业务处理确定的，执行模块将文件和相关元数据发送到一个或多个输出设备。

传统的内容管理系统构成一个单一的架构，其利用一个共有基础结构紧密地将收取阶段和管理阶段及完成阶段连在一起。相比较而言，本系统利用动态和灵活的架构，该架构使得相关于文件管理处理的周期次数显著减少，提供了处理中的总体效率。

传统内容管理系统依赖具有特征的可预测位置的结构化的形式，通常仅仅在视觉特征上操作。本系统包括业务处理形式的一个规则引擎，来检测和分类非结构化形式和结构化形式，这里属性和视觉布局的位置不是固定的。本系统使用规则谓语中的文件布局及布局内的文本内容来检测和分类文件。由本系统管理的文件流可动态配置到一个应用，这是传统工作流和文件管理产品不能提供的。本系统在动态配置性能方面可有效定制，并适用于真实世界的文件，如发货单和航运帐单。

本系统可以做成一个实用程序，如自动文件管理实用程序。本系统向用户提供识别自动文件管理实用程序的一个或多个业务处理、然后调用该自动文件管理实用程序来接收作为输入的文件、从该文件抽取元数据、分析该文件的元数据并分类该文件的手段。本系统向用户提供接收文件和相关元数据需要验证的通知的手段。本系统向用户提供验证或增加文件和相关元数据的手段。本系统还发出一个更新内容到输出设备，该更新内容包括文件、相关元数据、文件的分类、用户提供的增加的数据、用户采取的行动及业务处理的执行结果。本系统还提供当本系统处于操作中用户修改业务处理的手段。

附图说明

本发明的各种特征和获得方式将参考后续的说明书、权利要求书和附图做更详细的说明，其中标号适当地重复使用以指明有关项目的

相关性，其中：

图1是其中可以使用本发明的文件管理系统的示范操作环境的示意图；

图2是图1中的文件管理系统的高级体系结构的方框图；

图3是说明本发明的文件和元数据流的图1和图2中的文件管理系统的方框图；

图4是说明图1和图2的文件管理系统的操作方法的处理流程图；

图5是图1和图2的文件管理系统的示范性业务处理；

图6是说明图1和图2的文件管理系统的串行连接特性的方框图；

图7是说明图1和图2中的文件管理系统的可扩展性（scalability）和分布性质方框图。

具体实施方式

图1显示了示范性整体环境（“内容管理系统100”），其中可以使用本发明的一个系统、一种服务、一个计算机程序产品和相关方法（文件管理系统10或“系统10”），用于自动、动态地为电子商务主管服务构建文件管理应用程序。系统10包括典型地嵌入或安装于输入设备15或服务器20或网关25的软件编程码或计算机程序产品。可选地，系统10可以存储在合适的存储介质上，诸如盘、CD、硬驱等设备上。虽然系统10是关于文件提及的，其可以用于管理能电子地传送、处理、存储的任何类型或形式的内容，例如纸件或电子文件、照片、视频记录、音频记录等。

输入设备15可以由多种设备表示，诸如计算机30、扫描仪35或打印机40。输入设备15是能将内容输入到内容管理系统100的任何类型的内容捕获设备。用户可以通过输入设备15输入文件、图像、视频、音频等到内容管理系统100。输入设备15可以通过网络45访问服务器20。网关25通过网络45访问服务器20和输出设备50。

输入设备15、服务器20、网关25和输出设备50的每一个都包括允许通过网络45安全接口的软件。服务器20、网关25和输出设备

50 分别经通信链路 55、60、65 连接到网络 45。通信链路 55、60、65 包括诸如电话、电缆和卫星链路等链路。输入设备 15 可以经诸如电话、电缆或卫星链路的通信链路连接到网络 45。计算机 30、扫描仪 35 和打印机 40 经通信链路 70、75、80 连接到网络 45。

虽然系统 10 是关于网络 45 描述的，输入设备 15、服务器 20、网关 25 和输出设备 50 也可以经局域网、广域网或其它任何允许输入设备 15、服务器 20、网关 25 和输出设备 50 之间通信的网络来通信。此外，输入设备 15、服务器 20、网关 25 或输出设备 50 中的任何一个或多个可以共同定位，经过诸如局域网的网络来通信，而输入设备 15、服务器 20、网关 25 和输出设备 50 中的其它设备可以远地定位，经过诸如因特网的网络来连接。

计算机 30 在内容管理系统 100 中的功能为输入设备。计算机 30 可以用作其它功能，例如作为到内容管理系统 100 的用户接口。用户可以从计算机或计算机 30 所代表的其它设备访问文件以验证或浏览。

图 2 说明了系统 10 的高级层次结构。系统 10 包括元数据提示模块 205、元数据抽取模块 210、业务处理 215、执行模块 220。元数据提示模块 205 安装在输入设备 15 上。当用户经输入设备 15 输入一文件到内容管理系统 100 时，元数据提示模块 205 通过一个或多个提示（prompts）向用户请求关于该文件的信息。这些提示可以采用文本、音频、视频等形式。在一个实施例中，元数据提示模块 205 安装在服务器 20 上。

元数据抽取模块 210 安装在服务器 20 上。元数据抽取模块 210 自动地从该文件抽取元数据。执行模块 220 安装在网关 25 上。业务处理 215 也安装在网关 25 上，它包括由执行模块 220 执行的指令。执行模块 220 从服务器 20 检取该文件和相关元数据。执行模块 220 分析该文件和相关元数据来确定文件类型并分类该文件。执行模块 220 于是逐个文件地、选择性地、自动地执行业务处理 215 中的指令，确定文件类型和文件分类。

用户可以针对每个文件类型选择业务处理 215 中的哪个指令被执

行。此外用户可以在系统 10 操作时修改指令的选择，而不改变执行模块 220 的任何部分、关掉系统 10 或重启系统 10。执行模块 220 发出外部系统更新到输出设备 50 来将该文件、相关元数据和执行模块 220 的输出集成到输出设备 50。外部系统更新包括生成、更新、删除或查询。虽然输出设备 50 仅为说明的目的表示为一个设备，应该清楚系统 10 也可以应用于例如作为输出设备 50 操作的附加设备上。此外，附加设备和输出设备 50 可以操作多个不同的应用程序，诸如数据库、数据存储、内容管理系统等。

图 3 更详细地显示了内容管理系统 100A 的例子。图 4（图 4A 和 4B）显示了内容管理系统 100A 中的操作系统 10 的方法 400。操作中，结合参考图 3 和 4，用户经输入设备 15 通过例如扫描一个文件、经打印机驱动器直接打印一个文件等输入一个文件（步骤 405）。元数据提示模块 205 针对关于该文件的信息提示用户（步骤 410）。元数据提示模块 205 允许系统 10 与用户接口并请求关于与该文件相关的用户的信息，例如用户名称、用户 ID 或用户意见。元数据提示模块 205 还允许系统 10 与用户接口并请求不能从该文件识别的关于该文件的信息。用户提供的关于用户的信息和关于该文件的信息称为用户提示（user-prompted）元数据。

例如，对于发货单，元数据提示模块 205 可以请求交易日期、批发商等。对于保险索赔，元数据提示模块 205 可以请求单据号、客户等。元数据提示模块 205 检测正输入文件的文件类型并根据文件类型调整提供给用户的提示。元数据提示模块 205 通常针对文件中未提供的关于该文件的信息提示用户。在对于保险公司的内容管理系统 100A 的一个例子中，对于产生的不同类型的文件，诸如发货单、索赔、估算、损害图片、证言的视频、音频采访、修理投标等，提示是不同的。元数据提示模块 205 的输出是文件和用户提示元数据。

文件和与该文件相关的用户提示元数据被发送到服务器 20 和元数据抽取模块 210（步骤 415）。服务器 20 暂时存储该文件和用户提示元数据（步骤 420）。元数据抽取模块 210 处理该文件以便获得抽

取的元数据（步骤 425）；即，通过从该文件自动抽取元数据发现的关于该文件的数据。从文件自动抽取元数据可以使用任何方法，例如光学字符识别（OCR），逻辑 OCR，命名的实体抽取等等。该文件、用户提示元数据和抽取的元数据总称为文件/元数据包。

执行模块 220 从服务器 20 检取（retrieve）文件/元数据包（步骤 430）。执行模块 220 选择性地自动执行业务处理 215 中的指令。执行模块 220 基于用户提示的元数据和抽取的元数据自动分类该文件（步骤 435）。执行模块 220 自动确定文件是例如发货单、保险索赔中的证据、一个申请表等。基于文件分类，执行模块 220 从文件的相关部分有选择地抽取关键数据字段（步骤 440）。例如，执行模块 220 可以根据文件分类从文件内的已知位置抽取交易号、文件 ID 号等。选择性抽取的结果称作分区（zonal）数据元素。业务处理 215 确定关键数据字段及在文件中的位置。

执行模块 220 执行的具体抽取是从业务处理 215 确定的。对于每个文件类型，业务处理 215 确定分类要求、要抽取的数据、OCR 要求等。如业务处理 215 指示的，执行模块 220 可以选择性地仅 OCR 文件中特定的区域，这里称为分区 OCR（zonal OCR）。例如，用到保险索赔处理上，分区 OCR 可以抽取关于索赔的信息而不是索赔者的地址。

如业务处理 215 指示的，执行模块 220 发送一个通知给用户，通知需要验证文件/元数据包及分区数据元素（步骤 445）。这个通知可以通过任何方式提供，如邮件、电子邮件、即时消息、语音邮件、蜂窝电话、无线、电话或任何其它机制，通知适当的人来验证文件。执行模块 220 可以从文件分类确定通知的接收者。例如，可以通知一个人来验证保险索赔，同时通知另一个人来验证发货单。业务处理 215 提供验证通知的指示到一个特定的人或组织。

执行模块 220 将文件/元数据包、分区数据元素、业务处理 215 确定的分类结果输出到验证模块。用户验证（步骤 450）包括浏览和校正数据、增加（augment）数据及执行任何需要的动作。在一个实

施例中，通过验证接口（例如基于网络的验证接口）向用户提供验证页面。执行模块 220 从用户提示的元数据和抽取的元数据提供的信息中及业务处理 215 提供的指令中生成一个或多个定制的验证页面“on the fly”。

用户浏览用户提示的元数据、抽取的元数据和分区数据元素来检查 OCR 和印刷错误。用户可以浏览文件的分类以便更精确。用户还可以在需要时增加数据。此外，用户可以执行文件到达后需要的任何操作，例如支付发货单。在浏览和修改之后，验证模块将验证的文件/元数据包、验证的分区数据元素、验证的分类结果、任何增加的数据、用户执行的任何操作的记录返回到执行模块。

验证模块 305 获得的结果被返回到执行模块 220（步骤 455）。执行模块 220 选择性地自动执行来自业务处理 215 的任何附加指令（步骤 460）。执行模块 220 将文件/元数据包与输出设备 50 关联起来（步骤 465）。输出设备可以是数据库、内容管理系统、内容存储器等。执行模块 220 将文件/元数据包、分区数据元素、增加的数据、业务处理 215 的执行结果、用户执行的任何动作的记录及任何需要的外部系统更新输出到输出设备（步骤 470）。执行模块 220 的输出还包括与输出设备的外部系统集成，如生成、更新、删除和查询。

执行模块 220 根据相关于用户提示的元数据和抽取的元数据中的信息的业务处理 215 处理文件/元数据包。在一个实施例中，业务处理 215 以结构化或半结构化的表述存储，如可扩展的标识语言（XML）、网络服务的业务处理执行语言（BPEL）等。业务处理 215 将系统 10 定制到某特定业务发展和某具体业务处理。业务处理 215 是动态可适应的；业务处理 215 中编码的逻辑业务处理可简单地通过改变一个文件（例如 XML 文件）来改变，而无需改变系统 10 的任何其它部分、安装新软件、重启内容管理系统 100A、中断内容管理系统 100A 的操作。

业务处理 215 的示范说明作为 XML 文件 500 显示在图 5 中。虽然为说明目的业务处理 215 仅相对于 XML 文件做了说明，很显然系

统 10 也可以应用到例如任何结构化或半结构化编程语言。业务处理 215 包括分类说明 505、分区 OCR 说明 510、通知说明 515。根据需要可以增加另外的说明到业务处理 215。

对业务处理 215 的每个构成元件，使用说明 520 可以设定为开（on）（如图 5 所示）或关（<USAGE>Off</USAGE>）。如图 5 所示，使用说明 520 对于分类说明 505、分区 OCR 说明 510、通知说明 515 设定为“开”。分类说明 505、分区 OCR 说明 510、通知说明 515 之中的一个或多个的使用说明 520 可在操作内容管理系统 100 的任何时间改变。

分类说明 505 和分区 OCR 说明 510 还包括验证说明 525。验证说明 525 指定文件自动处理的人工验证。可以为分类说明 505 和分区 OCR 说明 510 指定验证说明 525。验证说明 525 可以设定为“开”（如图 5 所示）或关（<VERIFICATION> Off</VERIFICATION>）。分类说明 505 和分区 OCR 说明 510 中的一个或多个的验证说明 525 可以在内容管理系统 100 的操作期间的任意时间改变。

通知说明 515 包括通知接口说明 530、通知接触说明 535、通知文本 540。虽然在图 5 中示为电子邮件通知，该通知接口说明 530 可以制作为其它形式的通知，例如邮件、即时消息、语音消息（如蜂窝电话）、无线、电话等。由通知接口说明 530、通知接触说明 535 和通知文本 540 指定的任何一个或多个形式的通知可以在内容管理系统 100 操作期间的任意时间改变。

图 6 示出一个实施例，其中附加版本的内容管理系统 100 作为串行内容管理系统 600 的节点来操作。内容管理系统 100B 包括具有元数据提示模块 205（未示出）的输入设备 15B、元数据抽取模块 210B、执行模块 220B 和输出设备 50B。类似地，内容管理系统 100C 包括具有元数据提示模块 205（未示出）的输入设备 15C、元数据抽取模块 210C、执行模块 220C 和输出设备 50C。可以增加内容管理系统 100 的附加版本，如内容管理系统 100N 所示的。内容管理系统 100N 包括具有元数据提示模块 205（未示出）的输入设备 15N、元数据抽取模

块 210N、执行模块 220N 和输出设备 50N。

内容管理系统 100B、内容管理系统 100C 和内容管理系统 100N 中的每一个在工作流中作为节点运行。执行模块 220B 的输出发送到内容管理系统 100B 的输出设备 50B 及内容管理系统 100C 的元数据抽取模块 210C。以类似的方式，执行模块 605 的输出发送到串行内容管理系统 600 的总工作流的下一个元数据抽取模块 610。

例如，串行内容管理系统 600 可以表示一个发明的专利申请发展过程的工作流。内容管理系统 100B 代表专利披露（disclosure）节点。内容管理系统 100C 代表专利评估（review）节点。内容管理系统 100N 代表专利申请提交节点。输入设备 15B 代表从一个大公司的世界各地的发明人收集信息的许多输入设备。输入设备 15B 包括发明人使用的计算机、扫描仪、打印机、实验设备或任何其它捕获可以用于专利申请发展过程的信息的设备。来自输入设备 15B 的信息发送到元数据抽取模块 210B 和执行模块 220B，用于如前所述地处理。执行模块的输出如前所述地进行验证并存储在输出设备 50B 中。

执行模块 220B 的选择输出由执行模块 220B 自动输入到元数据抽取模块 210C 并加到专利评估节点的信息流上。专利评估节点需要的进一步信息由输入设备 15C 收集。专利评估节点的验证过程包括管理者和同伴对专利申请的认同。

执行模块 220C 的选择输出自动输入到元数据抽取模块 210N 并加到专利申请提交节点的信息流上。到元数据抽取模块 210N 的输入包括来自专利评审节点、专利代理人的输入、专利申请写作者的输入、起草者的输入及发明人的附加输入的选择的文件和信息。执行模块 50N 的输出包括专利申请和申请文件。

图 7 显示分布式文件管理系统 700，该系统说明了系统 10 的分布式能力并说明系统 10 的可扩展性。例如，一个公司可以包括北美分部、亚太分部和欧洲分部。北美分部包括北美内容管理系统 705。亚太分部包括亚太内容管理系统 710。欧洲分部包括欧洲内容管理系统 715。

北美内容管理系统 705 包括诸如输入设备 15AA 到 15AN 中的任

何一个或多个输入设备、诸如元数据抽取模块 210AA 到 210AN 中的任何一个或多个的元数据抽取模块、诸如执行模块 220AA 到 220AN 中的任何一个或多个的执行模块。输入设备 15AA 到 15AN、元数据抽取模块 210AA 到 210AN 和执行模块 220AA 到 220AN 中的任何一个或多个可以在同一房间、同一建筑或整个北美的不同位置。此外，可以按照需要将输入设备 15AA 到 15AN，元数据抽取模块 210AA 到 210AN，或执行模块 220AA 到 220AN 中的适当数量的单元加入到北美内容管理系统 705 之中，以便充分管理文件流。

亚太内容管理系统 710 包括输入设备 15BB、元数据抽取模块 210BB 和执行模块 220BB。输入设备 15BB、元数据抽取模块 210BB 和执行模块 220BB 中的任何一个或多个可以在同一房间、同一建筑或整个亚太地区的不同位置。虽然输入设备 15BB、元数据抽取模块 210BB 和执行模块 220BB 的每一个都在图 7 中示出，可以按照需要将输入设备 15AA，元数据抽取模块 210BB 和执行模块 220BB 中的适当数量的设备加入到亚太内容管理系统 710 之中，以便充分管理文件流。

欧洲内容管理系统 715 包括输入设备 15CC、元数据抽取模块 210CC 和执行模块 220CC。输入设备 15CC、元数据抽取模块 210CC 和执行模块 220CC 中的任何一个或多个可以在同一房间、同一建筑或整个欧洲的不同位置。虽然输入设备 15CC、元数据抽取模块 210CC 和执行模块 220CC 的每一个都在图 7 中示出，可以按照需要将输入设备 15CC，元数据抽取模块 210CC 和执行模块 220CC 加入到欧洲内容管理系统 715 之中，以便充分管理文件流。

如图 7 所示，北美内容管理系统 705、亚太内容管理系统 710 和欧洲内容管理系统 715 的输出发送到一个输出设备 50AA。输出设备 50AA 可以放置在北美、亚太地区、欧洲或其他任何位置。因此，使用系统 10 的内容管理系统 100 可以在世界范围串行（图 6）或分布式（图 7）管理文件流，或者以结合了串行和分布式特征的方式来管理文件流。例如，亚太内容管理系统 710 可以由串行内容管理系统 600 代替，输出设备 50N 的功能由输出设备 50AA 代替。

本内容管理系统可以应用的一个例子为信用卡争端管理。例如，一个客户关系管理公司处理客户和销售商之间因为信用卡收费引起的争端。信用卡争端管理的传统内容管理系统的争端处理流典型地如下：

1、 客户打电话给客户服务代表（CSR），并接收到一个唯一的案件 ID 和客户争端表；

2、 争端管理系统接收销售商争端文件并自动存储该销售商争端文件到传统的文件管理系统中；

3、 客户利用多种输入通道（例如邮件、电子邮件或传真）将该争端表和证明文件邮回到客户关系管理公司；

4、 邮件室工作人员扫描客户文件；客户文件放在停放区域直到客户服务代表浏览该客户文件及将客户文件与争端记录联系起来；

5、 客户还通过电子邮件发送一个证明该争端的收据；此电子邮件要求在该电子邮件能与争端记录联系起来之前客户服务代表对其进行阅览。

利用传统的信用卡争端管理内容管理系统，当客户已经发送了争端文件时会在步骤 3、4 之间发生多达一周的延迟，直到客户服务代表评估争端文件夹。与将客户文件与争端文件夹链接相关的人工步骤是由争端处理中的不同人员参与的，由此引起延迟。

利用内容管理系统 100 和系统 10，自上述步骤 3 的改进的处理过程如下：

1、 邮件室工作人员使用输入设备 15 扫描客户文件，响应来自元数据提示模块 205 的提示输入案件 ID。系统 10 自动将客户文件与争端记录关联起来。

2、 收到客户的电子邮件后，客户服务代表通过响应来自元数据提示模块 205 的提示输入案件 ID，将该电子邮件直接从电子邮件应用程序插入到正确的争端文件夹。

3、 执行模块 220 自动地将争端文件夹移动从“悬置”状态移动到“准备好”状态以便争端办公人员阅览（即验证）。

内容管理系统 100 和系统 10 提供的改进的业务处理能减少争端

解决时间，从大约一周到大约两天，这为客户提供了有吸引力的业务价值。

内容管理系统 100 和系统 10 可以应用的另一例子可以是管理停车票据。大城市管理停车票据的过程包括数据中心、呼叫中心、支付系统和支付应用。美国的一个大城市每年拥有 300 万手写票据。

目前，停车票据通过晚上从分支办公室（全城大约 30 个分支办公室，平均每个位置有 1 万张票据）收集纸件文件来管理。在中心位置，由两个扫描仪操作人员和专门负责扫描后验证文件任务的十个验证人员，利用高容量扫描仪将文件成批扫描成图像。在建立票据的电子记录之前这个过程花费 3 个工作日；因此票据输入和验证是相关于票据的任何业务处理的关键因素。

内容管理系统 100 和系统 10 在票据事件的 1 个工作日内生成每个分支位置的 1 万张票据的电子记录。系统 10 还支持票据和相关数据的分布式验证，这样票据记录可以在两个工作日内触发与票据相关的业务处理 215。总体来说，在处理周期中，利用内容管理系统 100 和系统 10 可以获得高效率。

应当理解，已经说明的本发明的具体实施例只是本发明原理的特定应用。可以在不脱离本发明的精神和范围，对本系统、方法和服务做成很多变动，用于自动和动态地为电子商务主持服务构造文件管理应用程序。虽然本发明是针对文件的，很显然本发明也可以应用于可以电子传送、处理或存储的任何形式或类型的内容，例如纸件或电子文件、照片、视频记录、音频记录等。

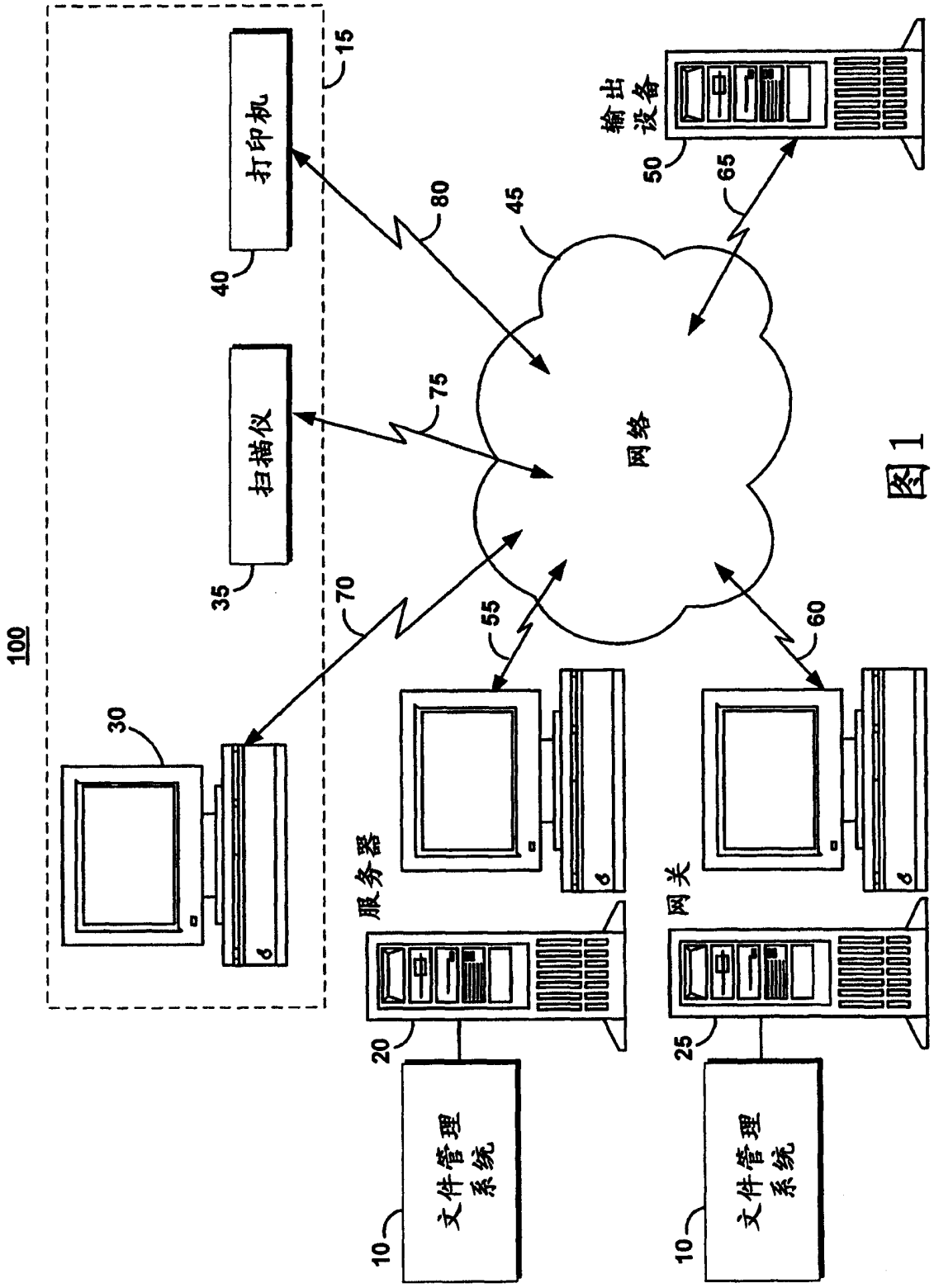

100
打印机
15
40
扫描仪
35
80
75
70
45
网络
输出设备
50
65
55
60
30
服务器
网关
20
25
文件管理系统
10
文件管理系统
10

图1

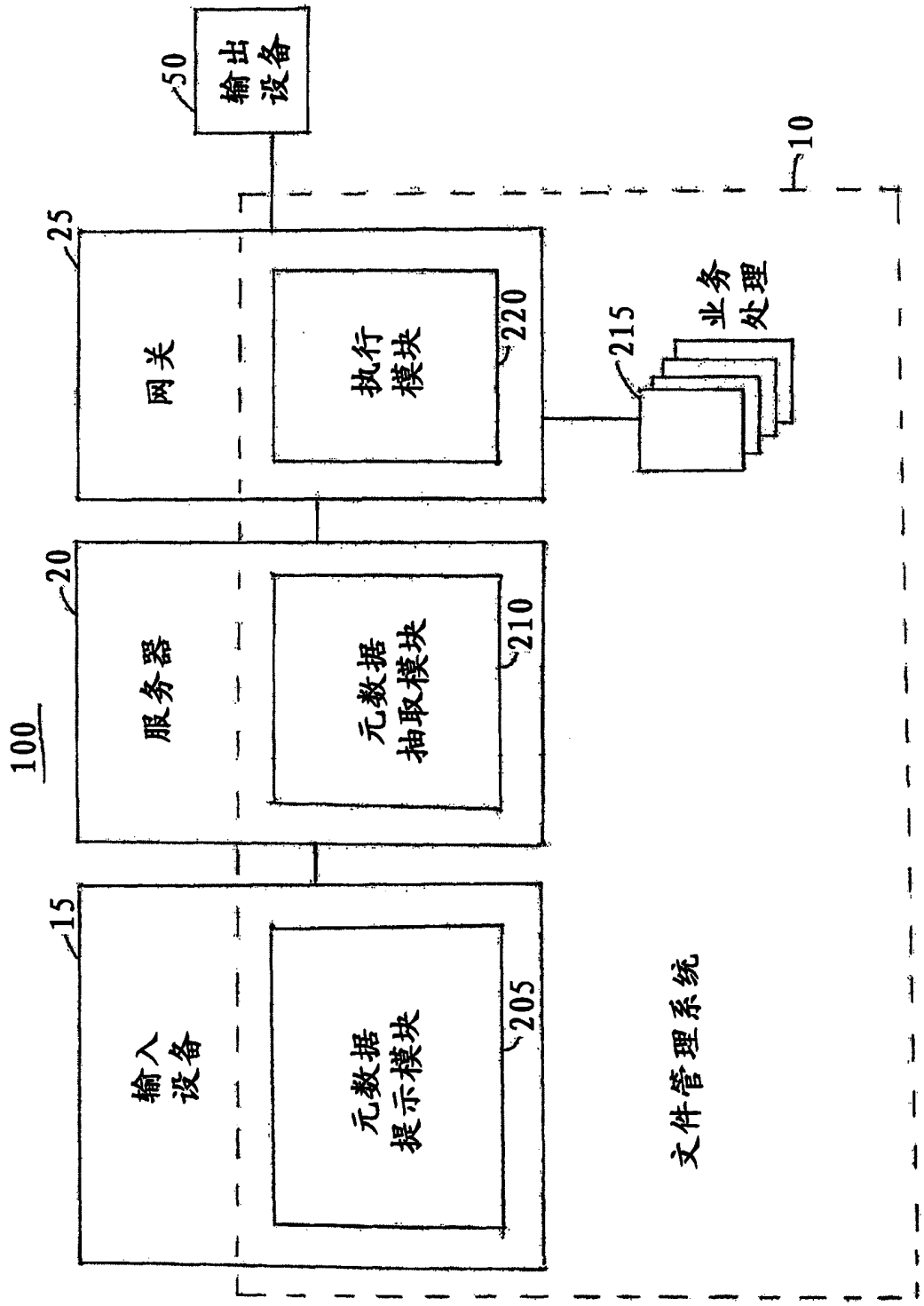

100
50
输出设备
10
25
网关
执行模块
220
215
业务处理
20
服务器
元数据抽取模块
210
15
输入设备
元数据提示模块
205
文件管理系统

图2

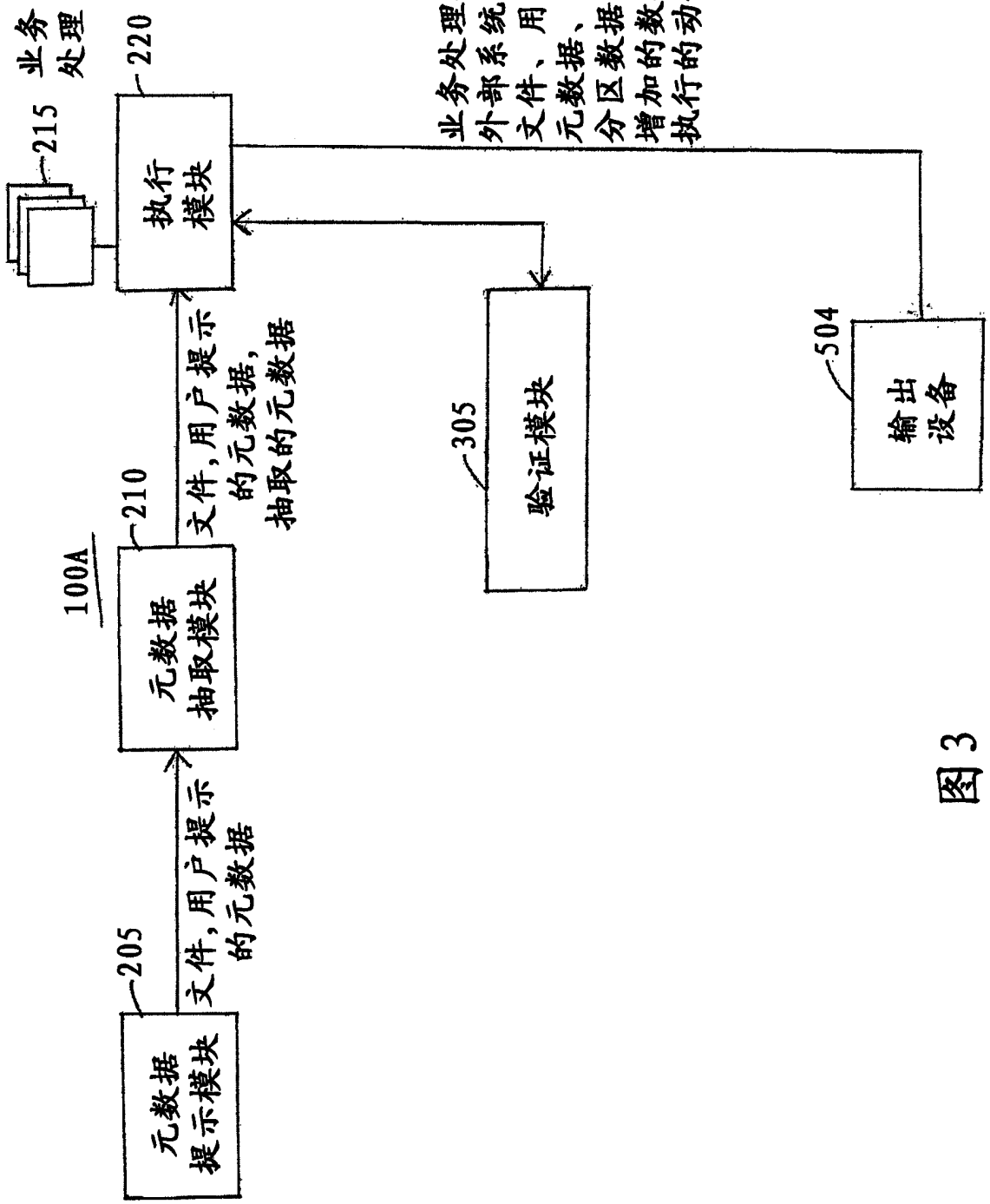

元数据
提示模块
205
文件，用户提示
的元数据
100A
元数据
抽取模块
210
文件，用户提示
的元数据，
抽取的元数据
执行
模块
220
业务
处理
215
业务处理执行结果、
外部系统更新、
文件、用户提示的
元数据、分类结果、
分区数据元素、
增加的数据
执行的动作的记录
验证模块
305
输出
设备
504

图 3

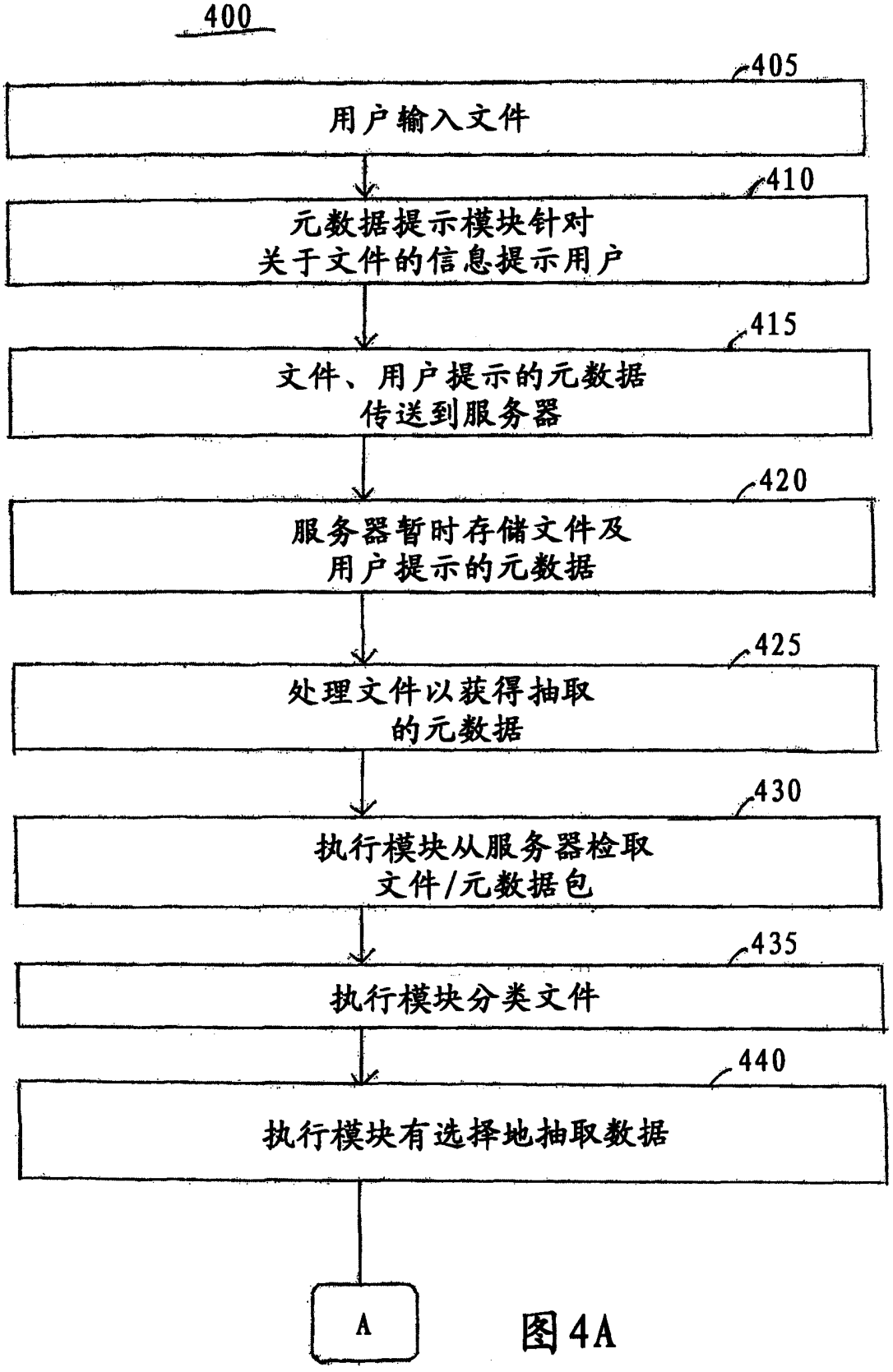

400
405
用户输入文件
410
元数据提示模块针对
关于文件的信息提示用户
415
文件、用户提示的元数据
传送到服务器
420
服务器暂时存储文件及
用户提示的元数据
425
处理文件以获得抽取
的元数据
430
执行模块从服务器检取
文件/元数据包
435
执行模块分类文件
440
执行模块有选择地抽取数据
A

图4A

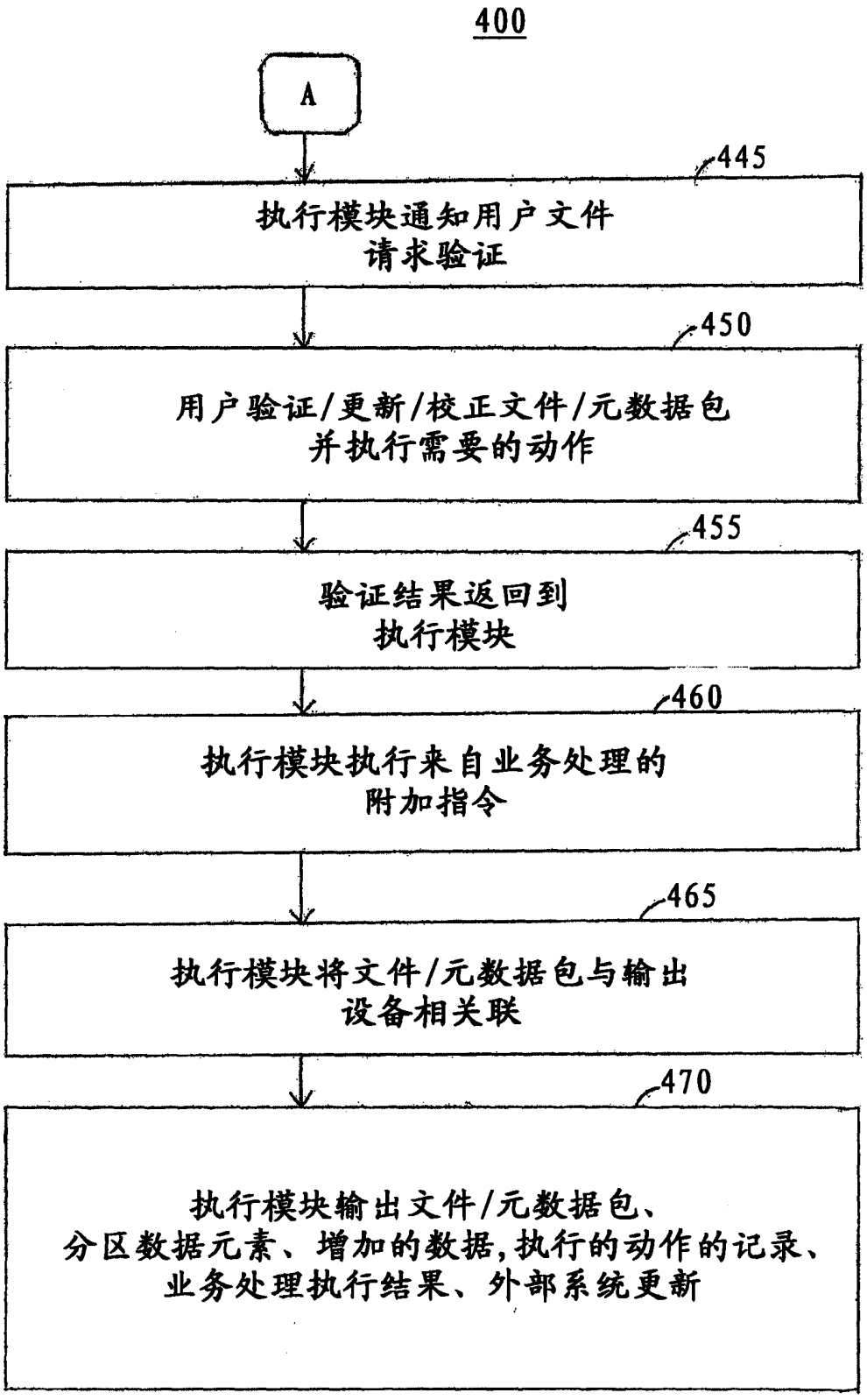

400
A
445
执行模块通知用户文件
请求验证
450
用户验证/更新/校正文件/元数据包
并执行需要的动作
455
验证结果返回到
执行模块
460
执行模块执行来自业务处理的
附加指令
465
执行模块将文件/元数据包与输出
设备相关联
470
执行模块输出文件/元数据包、
分区数据元素、增加的数据,执行的动作的记录、
业务处理执行结果、外部系统更新

图4B

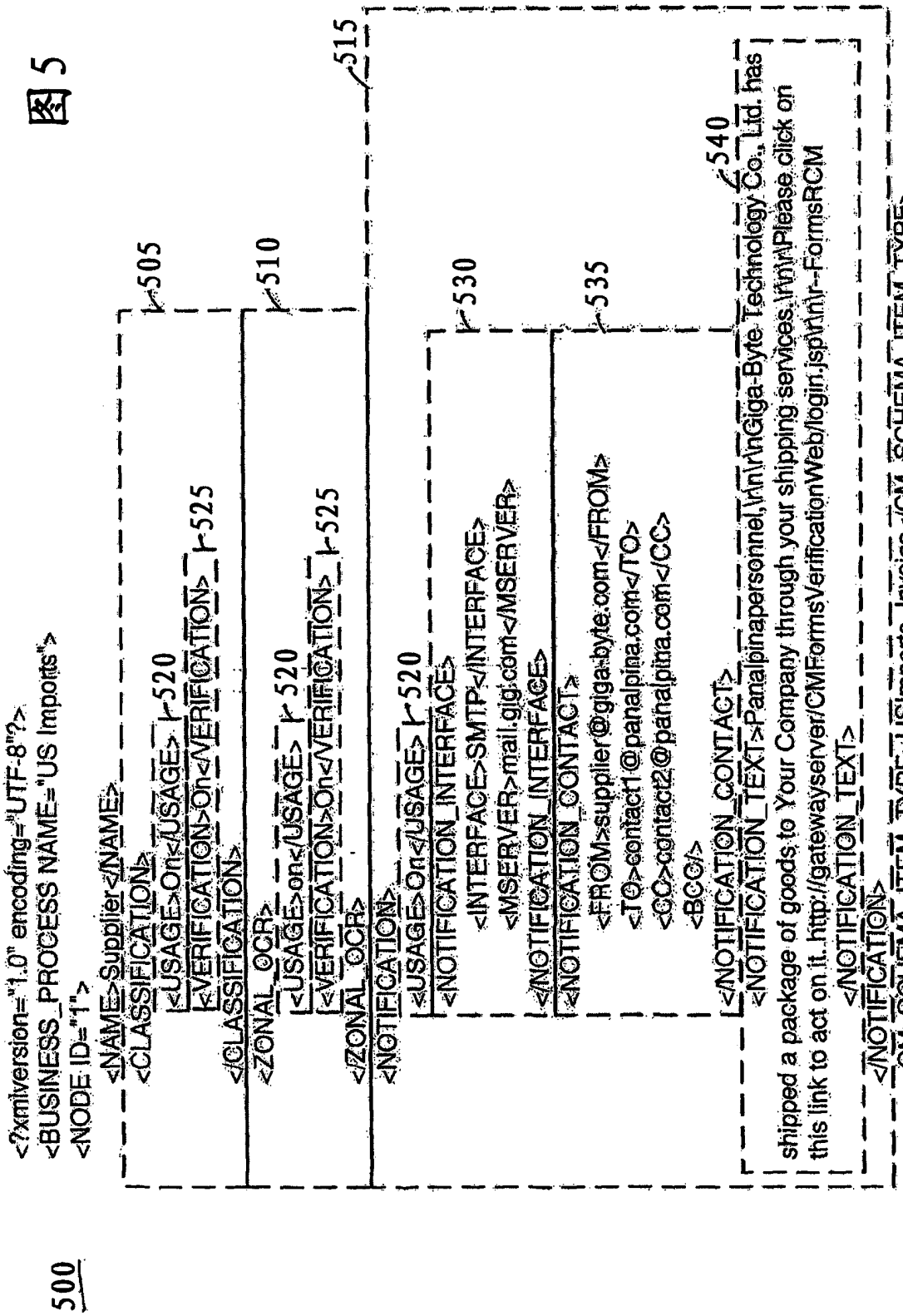

500
<?xmlversion="1.0" encoding="UTF-8"?>
<BUSINESS_PROCESS NAME="US Imports">
<NODE ID="1">
<NAME>Supplier</NAME>
<CLASSIFICATION>
<USAGE>On</USAGE>
<VERIFICATION>On</VERIFICATION>
</CLASSIFICATION>
<ZONAL_OCR>
<USAGE>on</USAGE>
<VERIFICATION>On</VERIFICATION>
</ZONAL_OCR>
<NOTIFICATION>
<USAGE>On</USAGE>
<NOTIFICATION_INTERFACE>
<INTERFACE>SMTP</INTERFACE>
<MSERVER>mail.gigi.com</MSERVER>
</NOTIFICATION_INTERFACE>
<NOTIFICATION_CONTACT>
<FROM>supplier@giga-byte.com</FROM>
<TO>contact1@panalpina.com</TO>
<CC>contact2@panalpina.com</CC>
<BCC/>
</NOTIFICATION_CONTACT>
<NOTIFICATION_TEXT>Panalpinapersonnel,\r\n\r\nGiga-Byte Technology Co., Ltd. has
shipped a package of goods to Your Company through your shipping services.\r\n\rPlease click on
this link to act on it..http://gatewayserver/CMFormsVerificationWeb/login.jsp\r\n\r--FormsRCM
</NOTIFICATION_TEXT>
</NOTIFICATION>
<CM_SCHEMA_ITEM_TYPE>USImports_Invoice</CM_SCHEMA_ITEM_TYPE>
</NODE>
505
510
515
520
520
520
525
525
530
535
540

图 5

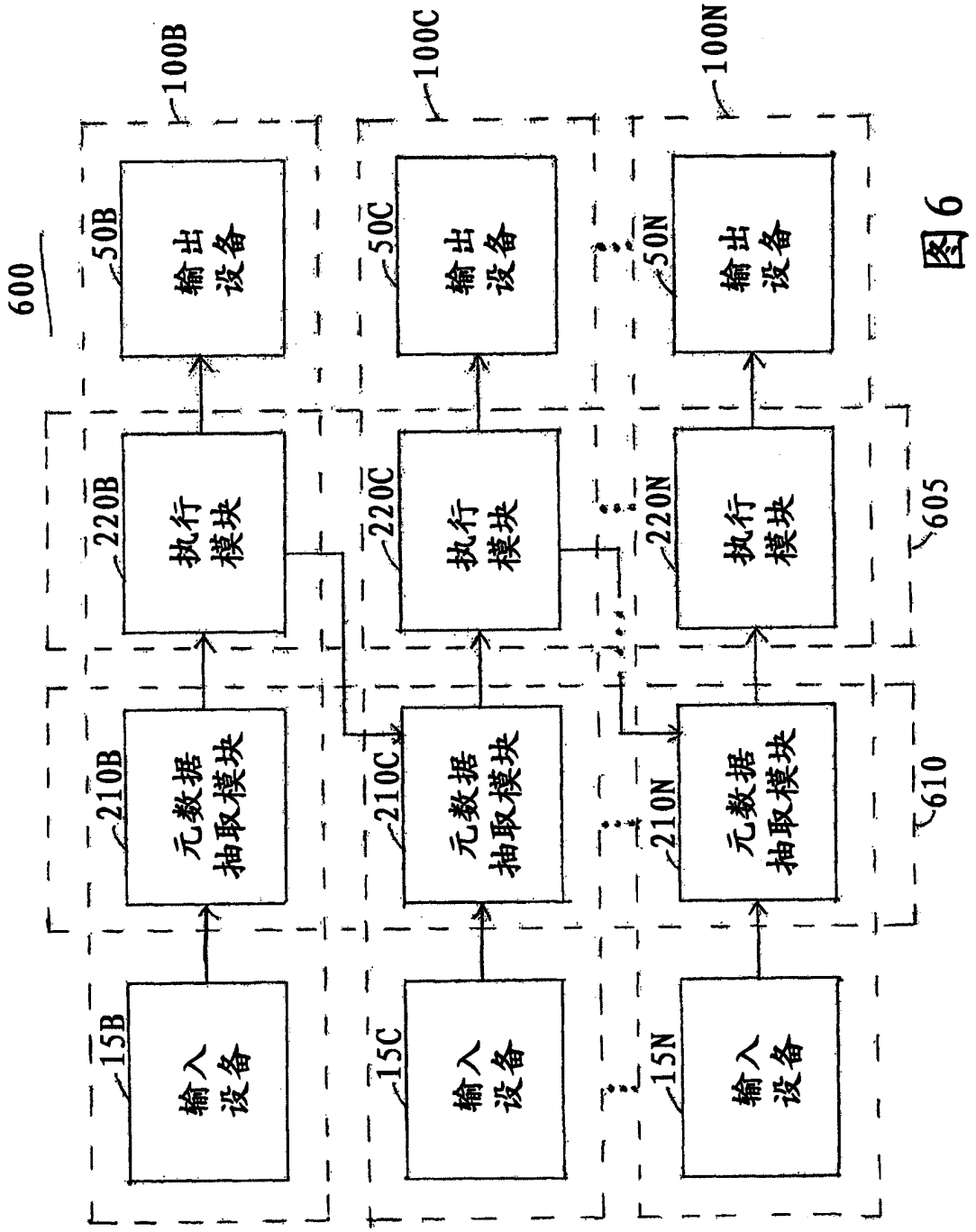

600
100B
100C
100N
50B
50C
50N
输出设备
220B
220C
220N
执行模块
605
210B
210C
210N
元数据抽取模块
610
15B
15C
15N
输入设备

图6

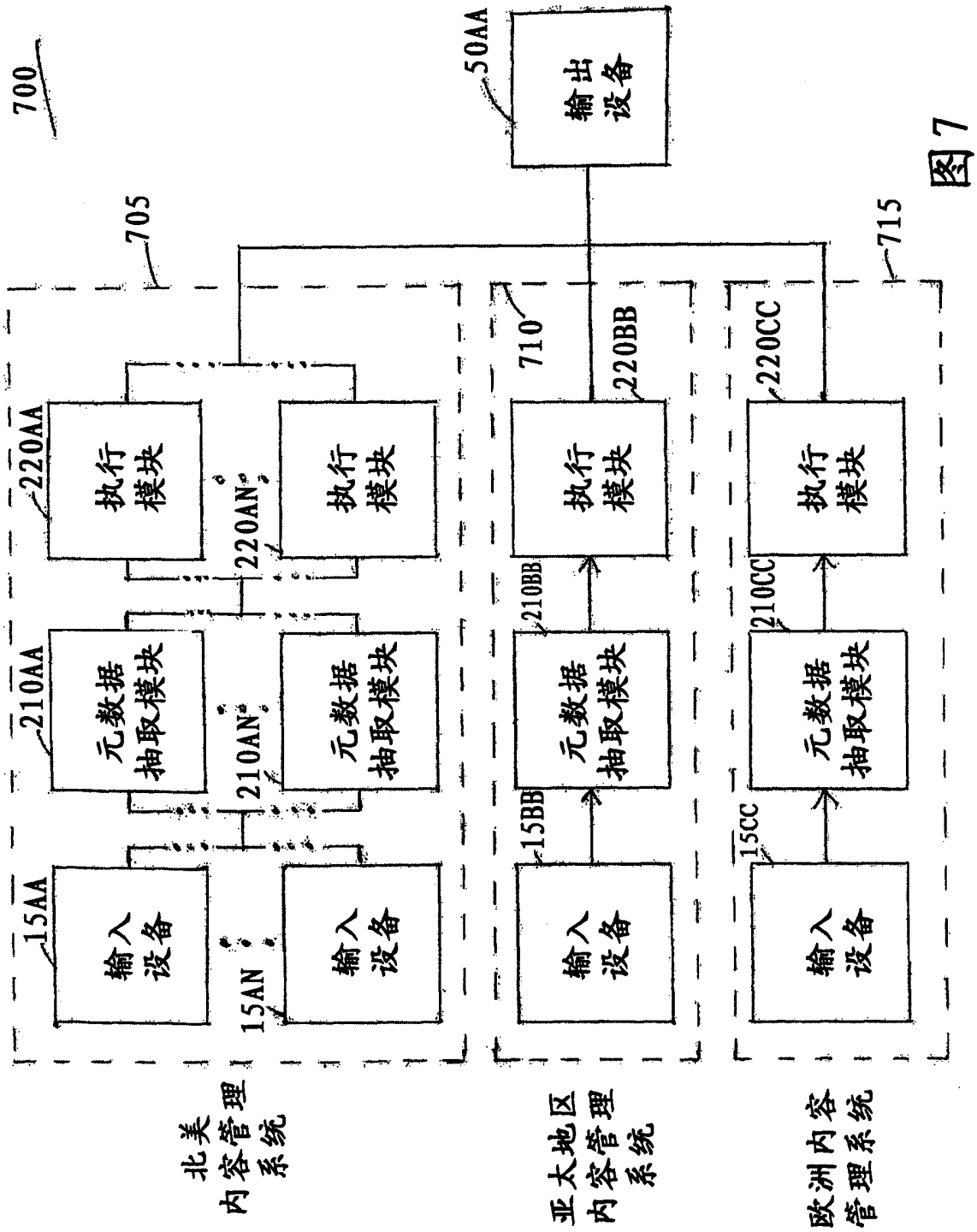

700
50AA
输出设备
705
715
710
220AA
220AN
220BB
220CC
执行模块
210AA
210AN
210BB
210CC
元数据抽取模块
15AA
15AN
15BB
15CC
输入设备
北美内容管理系统
亚太地区内容管理系统
欧洲内容管理系统

图7