



(12) 发明专利

(10) 授权公告号 CN 102414677 B

(45) 授权公告日 2016. 04. 13

(21) 申请号 201080018349. 8

代理人 高见

(22) 申请日 2010. 04. 14

(51) Int. Cl.

(30) 优先权数据

12/427, 755 2009. 04. 22 US

G06F 17/00(2006. 01)

G06F 7/24(2006. 01)

(85) PCT国际申请进入国家阶段日

2011. 10. 21

(56) 对比文件

US 7610285 B1, 2009. 10. 27,

WO 2007/062254 A3, 2007. 05. 31,

US 2008/0071908 A1, 2008. 03. 20,

US 2008/0071813 A1, 2008. 03. 20,

US 2008/0077682 A1, 2008. 03. 27,

(86) PCT国际申请的申请数据

PCT/US2010/031106 2010. 04. 14

(87) PCT国际申请的公布数据

W02010/123737 EN 2010. 10. 28

审查员 王金萍

(73) 专利权人 微软技术许可有限责任公司

地址 美国华盛顿州

(72) 发明人 P·A·奥尔泰安 C·劳 J·哈迪

N·本-茨维 R·卡拉赫

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

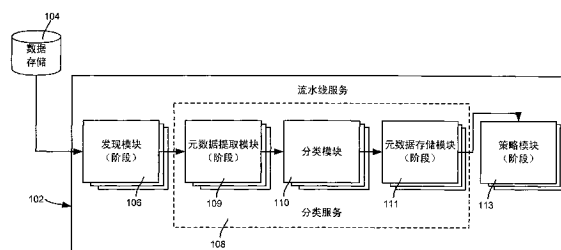
权利要求书3页 说明书11页 附图6页

(54) 发明名称

包括自动分类规则的数据分类流水线

(57) 摘要

描述了通过可扩展的数据处理流水线（包括分类流水线）来处理数据项（例如，文件）以基于数据项的分类来促进对数据项进行管理的技术。发现模块定位要处理的数据项。独立分类流水线获取与每一个被发现的数据项相关联的元数据（属性），而一个或多个分类器基于元数据来分类数据项。独立的策略模块基于其分类将策略应用到每一个数据项。可以基于各种准则来调用多个分类器。预定义排序的分类器、权威性分类器和/或聚合机制处理任何分类冲突。可以提供不同类型的分类器，而每一个分类器都可以对应于自动分类规则；分类器可以直接改变属性（例如，设置分类）或将结果返回到用于改变属性的相对应的规则机制。



1. 在计算环境中,一种用于对数据项进行分类的系统,所述系统包括:

一个或多个处理器;以及

耦合至所述一个或多个处理器的存储器,所述存储器存储指令,所述指令在由所述一个或多个处理器执行时使得所述一个或多个处理器:

提供分类流水线,其包括获取与数据项相关联的元数据以及与所述数据项相关联的现有分类元数据的组件,其中所述现有分类元数据包括所述数据项的当前分类值,并且其中所述数据项的所述当前分类值被存储在所述数据项中;

提供多个分类器模块,其中所述多个分类器模块中的每个分类器模块具有相关联的分类规则,并且其中所述分类规则中的每一个在被调用时,使用与数据项相关联的元数据以及与数据项相关联的现有分类元数据来确定所述数据项的分类元数据,并且其中所述多个分类器模块是使用以下中的至少一者来进行调停的:聚合分类、权威性分类和高度分类;以及

提供将所述分类元数据与所述数据项进行关联以便用于将策略应用到所述数据项的组件。

2. 如权利要求 1 所述的系统,其特征在于,所述分类流水线被结合到数据项处理流水线中,并且其中所述数据项处理流水线包括发现所述数据项的发现模块。

3. 如权利要求 2 所述的系统,其特征在于,所述数据项对应于文件,并且其中所述发现模块被配置成执行以下各项中的至少一项:(i) 扫描文件系统以发现其中的文件,以及(ii) 扫描文件以检测对文件的改变。

4. 如权利要求 1 所述的系统,其特征在于,所述分类流水线被结合到数据项处理流水线中,并且其中所述数据项处理流水线包括评估所述分类元数据以将策略应用到所述数据项的策略模块。

5. 如权利要求 1 所述的系统,其特征在于,还包括用于基于以下各项中的至少一项来确定是否调用所述多个分类器模块中的一个分类器模块的确定模块:(i) 任何现有分类数据,以及(ii) 指示以前的对数据文件的改变的时间戳或其他标识符。

6. 如权利要求 1 所述的系统,其特征在于,还包括,用于与所述分类流水线进行交互以外部地设置分类元数据的接口。

7. 如权利要求 1 所述的系统,其特征在于,还包括,用于与所述分类流水线进行交互以外部地获得分类元数据的接口。

8. 如权利要求 1 所述的系统,其特征在于,获取与数据项相关联的元数据以及与所述数据项相关联的现有分类元数据的组件是可扩展的;并且其中所述多个分类器模块中的每个分类器模块是可扩展的;并且进一步的,其中与分类器元数据相关联的组件是可扩展的。

9. 如权利要求 1 所述的系统,其特征在于,所述多个分类器模块包括执行以下各项中的至少一项的分类器:(i) 返回真或假结果,(ii) 显式地设置与分类器元数据相对应的至少一个属性,以及(iii) 返回真或假结果并显式地设置与分类器元数据相对应的至少一个属性。

10. 如权利要求 1 所述的系统,其特征在于,所述多个分类器模块包括基于以下各项中的至少一项来分类数据项的分类器:(i) 数据项的位置,(ii) 基于全局储存库的分类器,以及(iii) 基于项内包含的内容来分类项的基于内容的分类器。

11. 如权利要求 1 所述的系统,其特征在于,所述多个分类器模块包括替代所述分类器组中的另一个分类器的分类元数据的权威性分类器。

12. 在一计算环境中,一种用于对数据项进行分类的方法,所述方法由一个或多个处理器来执行并且包括:

发现数据项;

由一个或多个处理器使用与所述数据项相关联的一个或多个属性来分类所述数据项以生成相关联分类属性集,其中所述一个或多个属性包括与所述数据项相关联的现有分类属性,其中所述现有分类属性包括所述数据项的当前分类值,其中所述数据项的所述当前分类值由以下中的至少一者确定:聚合分类、权威性分类和高度分类,并被存储在所述数据项中,并且其中所述数据项是由多个分类组件来分类的;以及

基于以下各项中的至少一项来将策略应用到所述数据项:(i) 分类属性集,以及(ii) 由以下中的至少一者分配的值:聚合分类、权威性分类和高度分类。

13. 如权利要求 12 所述的方法,其特征在于,使用与所述数据项相关联的一个或多个属性包括使用来自包括多个分类器的分类器集的分类结果来自动应用分类规则。

14. 如权利要求 12 所述的方法,其特征在于,还包括按预定义顺序调用多个分类组件,并使得所述多个分类组件中的一个能将属性集传递到所述多个分类组件中的第二个。

15. 如权利要求 12 所述的方法,其特征在于,还包括按预定义顺序调用多个分类组件,并使得所述预定义顺序中的后继分类组件能改变所述预定义顺序中的在前分类组件的属性集。

16. 如权利要求 12 所述的方法,其特征在于,还包括基于一个或多个当前定义的分类属性来确定是否调用多个分类组件。

17. 如权利要求 12 所述的方法,其特征在于,还包括确定经聚合的分类属性集的评估顺序。

18. 一种用于对数据项进行分类的方法,所述方法包括:

发现一个或多个数据项;

获取与所述数据项相关联的属性的属性集,其中所述属性集包括与所述数据项相关联的现有分类属性,其中所述现有分类属性包括所述数据项的当前分类值,并且其中所述数据项的所述当前分类值被存储在所述数据项中;

确定是否使用分类器集中的多个分类器来分类数据项,包括:

当所述分类器集中的所述多个分类器被调用时,聚合来自所述多个分类器的分类结果;

执行权威性分类,由此所述分类器集中的所述多个分类器之一对其结果作标志,使得它在与任何其他分类器的任何冲突中胜出;

执行高度分类,由此基于预定义的高度在流水线中对所述多个分类器进行排序,使得较低高度的每一分类器在较高高度的每一分类器之前在所述流水线中执行;

基于所述多个分类器所产生的任何改变来更新所述属性集;以及

基于经更新的属性集,将策略应用到所述数据项。

19. 如权利要求 18 所述的方法,其特征在于,获得所述属性集包括以下中的至少一个:

(i) 提取与数据项相对应的元数据,以及(ii) 检索与数据项相关联的现有属性集。

20. 如权利要求 18 所述的方法, 其特征在于, 更新所述属性集包括使用以下中的至少一个来更新所述属性集: (i) 所述多个分类器, 以及 (ii) 被配置成基于从所述多个分类器提供的结果来更新所述属性集的规则机制。

包括自动分类规则的数据分类流水线

[0001] 背景

[0002] 在典型的企业环境中维护和处理的数据量是巨大的,并迅速地增大。例如,信息技术(IT)部门必须处理数十种格式的数百万或者甚至数十亿的文件是常见的。此外,现有的数量以很大的速率增长(例如,每年两位数的增长)。这些数据中大多数没有被积极地管理,并以非结构化的表单保留在文件共享中。

[0003] 现有数据管理工具和实践不能很好地跟上可能存在的各种和复杂情况。这样的情况包括顺应性、安全性,以及存储,并适用于非结构化数据(例如,文件),半结构化数据(例如,文件加额外的属性/元数据)和结构化数据(例如,在数据库中)。因此,需要降低管理成本和风险的任何技术。

[0004] 概述

[0005] 提供本概述是为了以简化的形式介绍将在以下详细描述中进一步描述的一些代表性的概念。本概述并不旨在标识出所要求保护的主题的关键因素或必要特征,也不旨在以将限定所要求保护的主题的范围的任何方式被使用。

[0006] 简言之,此处所描述的主题的各个方面涉及这样的技术:通过该技术经由数据处理流水线(包括分类流水线)来处理数据项(例如,文件),以促进基于它们的分类对数据项的管理。一方面,分类流水线获取与每一个已发现的数据项相关联的元数据(例如,业务影响、隐私级别等等)。一组一个或多个分类器若被调用,则将数据项分类为分类元数据(例如,一个或多个属性),然后将这些属性与数据项进行关联(相关联地保存)。然后,可以基于其相关联的分类元数据来将策略应用于每一个数据项,例如,以基于每一个文件的元数据,使文件期满,改变文件的保护/访问级别等等。

[0007] 一方面,数据项处理流水线包括用于项发现、分类和策略应用的独立阶段的模块化组件。每一个阶段都是可扩展的,并可包括在该阶段起作用的一个或多个模块(或没有模块)。可以在外部设置或分别通过 set(设置)或 get(获得)接口获取每一个项的分类元数据/属性。

[0008] 一方面,在分类阶段,可以调用多个分类器模块。可以基于各种准则,如先前是否和/或先前何时分类了数据项,作出是否要调用每一个分类器的决定。在分类数据项时,分类器可以使用与数据项相关联的属性中的任一个,和/或数据项本身的内容。预定义排序的分类器、权威性分类器和/或聚合机制是可以被用来处理关于不同的分类器如何分类相同项的任何冲突的技术。

[0009] 可以提供不同类型的分类器,包括基于数据项的位置来分类数据项的分类器,基于全局储存库的分类器(基于所有者和/或作者),和/或基于项内包含的内容来分类项的基于内容的分类器。每一个分类器都可以对应于自动分类规则;分类器可以直接改变属性值,或将结果返回到相对应的规则机制,以使得相对应的规则机制可以改变属性。

[0010] 通过下面的结合附图对本发明进行的详细说明,其他优点可以变得显而易见。

[0011] 附图简述

[0012] 本发明是作为示例说明的,而不仅限于附图,在附图中,类似的参考编号表示类似

的元件,其中:

[0013] 图 1 是示出了用于自动处理用于数据管理的数据项(包括发现数据项、分类那些数据项,并基于分类来应用策略)的流水线服务中的示例模块的框图。

[0014] 图 2 是示出了当将文件服务器的文件处理为与文件相关联的属性时由流水线服务执行的示例步骤的表示。

[0015] 图 3 是示范了如何通过分类运行时在多个模块之间传递数据项的属性以便进行处理的示例分类服务体系结构的表示。

[0016] 图 4A 和 4B 包括示出了处理数据项所采取的示例步骤的流程图,包括为策略应用而分类项的步骤。

[0017] 图 5 示出了可以将本发明的各个方面集成到其中的计算环境的说明性示例。

[0018] 详细描述

[0019] 此处所描述的技术的各个方面一般涉及通过将数据项(对象)分类为一个类别,并基于分类应用数据管理策略来管理数据(例如,文件服务器上的文件等等)。在一方面,这是通过用于启用数据分类的解决方案的模块化方法,基于分类流水线来完成的。一般而言,流水线包括通过公用接口进行通信的一系列模块化软件组件。在各时间点,数据被发现和分类,且基于数据分类,将策略应用于数据。

[0020] 尽管此处使用了各种示例,如用于分类文件服务器上维护的文件/数据的不同的文件分类类型,但是,应该理解,此处所描述的示例中的任何一个都是非限制性示例。例如,不仅可以分类文件,而且也可以将其他数据结构分类到相关的分类“类型”,例如,可以分类任何结构化的数据(例如,遵循描述如何表示数据的抽象模型并可以被访问的任何数据片段),例如,电子邮件项、数据库表、网络数据等等。此外,可以使用其他的存储数据的方式,例如,作为代替或补充的文件服务器,数据可以保留在本地存储、分布式存储、存储区域网络、因特网存储等等中。如此,本发明不仅限于此处所描述的任何特定实施例、方面、概念、结构、功能或示例。相反,此处所描述的任何实施例、方面、概念、结构、功能或示例都是非限制性的,可以以一般而言在计算和进行数据管理时提供好处和优点的各种方式来使用本发明。

[0021] 图 1 示出了涉及此处所描述的技术的各个方面,包括用于处理数据项的流水线,该流水线如此处所示范的可以被用来处理文件,但是,如可理解的可以被用来处理诸如电子邮件项之类的一个或多个其他数据结构。在图 1 的示例中,流水线被实现为对如由数据存储 104 所表示的任何数据集进行操作的服务 102。

[0022] 一般而言,流水线服务 102 包括发现模块 106、分类服务 108,以及策略模块 113。注意:术语“服务”不一定与单一机器相关联,而是协调流水线的某一执行的机制。在此示例中,分类服务 108 包括其他模块,即,元数据提取模块(诸元数据提取模块)109、分类模块(诸分类模)110,以及元数据存储模块(诸元数据存储模块)111。下面所描述的模块中的每一个都可以被视为阶段,实际上,每一个操作的时间线不必是相邻的,即,每一个阶段都可以被相对独立地执行,而不必紧接着前一阶段。例如,发现阶段可以发现和维护分类阶段以后分类的项。作为另一个示例,可以以每日为基础分类数据,且每周一次地运行数据管理应用程序(例如,备份)。任何一个阶段都可以被独立地执行,实时地在线处理或离线处理,在前台或在后台(例如,懒惰(lazy))操作中,或在分开的机器上以分布式方式执行。

[0023] 一般而言,发现模块(诸发现模块)106发现要分类的项(例如,文件),并可以使用一个以上的机制来执行此项操作。作为示例,可以有两种方式来发现文件服务器上的文件,一种方式是通过扫描文件系统来操作,而另一种方式是从远程文件访问协议检测新的对文件的修改。一般而言,被发现的数据作为项被提供到分类阶段/服务108,以便进行分类,无论是直接还是通过中间存储。以此方式,发现在逻辑上可以与分类分离。

[0024] 可以以许多方式来启动发现。一种方式是按需的,其中,项是应请求发现的。另一种方式是实时的,其中,对一个或多个项的更改触发发现操作。再一种方式是调度的发现,例如,一天一次,如在正常工作时间之后。再一种方式是懒惰发现,其中,后台进程等等以低优先级操作以发现项,例如,当网络或服务器利用率相对较低时。此外,还要注意,发现可以在在线操作中运行,即,对真实数据,或对数据的离线副本,如原始数据的时间点快照;(注意,一般而言,快照副本是指当特定数据项处于某个定义的时间点时的该特定数据项的副本,由此,对快照副本进行处理会在数据项正在被处理时有助于将数据项维持在恒定状态,这与其中数据项可以实时地变化的实况系统不同)。

[0025] 在分类阶段/服务108(下面所描述的)之后,策略模块113基于每一个项的分类来应用策略。作为示例,信息泄露保护产品可以将某些文件分类为具有“个人可标识信息”等等。可以利用一个策略配置文件备份产品,以便将被分类为具有“个人可标识信息”的任何文件备份到加密存储中。

[0026] 转向涉及分类的各个方面,如在图1中所表示的,元数据提取模块(诸元数据提取模块)109发现与数据项相关联的元数据。例如,文件系统具有它将其与文件进行关联的许多属性,并可以以已知方式提取这些属性。元数据提取模块(诸元数据提取模块)109还提取分类元数据的当前值,以便它可以被用作对分类阶段的输入。注意,分类可以对实时数据或备份数据进行。

[0027] 元数据的一些示例包括具有各种元素的分类属性定义,如属性名称(或标识符)、属性值类型(标识实际值的数据类型,例如,简单数据类型,如字符串、日期、布尔值、值的有序集或多集),以及复杂数据类型,如通过分层分类法(文档类型、组织单元,或地理位置)所描述的数据类型)。分类属性值(叫做“属性值”或简称为“属性”)是可以被指派到数据项以达到分类该数据项的某一值。此值与分类属性相关联,并一般遵守由相关联的属性定义施加的限制。

[0028] 其他示例包括属性模式(描述对可能的值的更多限制),以及描述如何将多个值聚合在单个值中的聚合策略(在流水线执行过程中需要这样的聚合的情况下)。更进一步,元数据可以包括与属性相关联的附加属性,如语言依存信息、额外标识符,等等。

[0029] 作为示例,考虑类型“ordered value set(有序值集合)”的名为“Business impact(业务影响)”的属性,其只限于值HBI(高业务影响)、MBI(中等业务影响)和LBI(低业务影响),聚合策略是:HBI优先于MBI,而MBI又优先于LBI。注意,在分类过程中,属性值与数据项的关联将自动将该文档“绑定到”文档的类(即,类别)。例如,通过将属性“BusinessImpact = HBI”附连到数据项,此数据项被隐式地指派到文档“类别”BusinessImpact = HBI。

[0030] 元数据也可以保留在外部数据源或其他高速缓存中。一个示例包括允许用户或客户端和/或一个或多个其他机制设置分类元数据,或分类本身,并将它维护在诸如数据库

之类的数据存储中。如此,例如,用户可以手动将文件设置为包含“个人可标识信息”等等。自动化过程可以执行类似的操作,如通过基于什么文件夹包含文件来确定元数据,例如,当一个文件被添加到敏感的文件夹时,进程可以自动设置该文件的相关联元数据。

[0031] 此外,可以从先前的提取和 / 或分类操作维护 (高速缓存) 项的元数据。如此,元数据提取可以分多个部分,例如,提取现有元数据 (检索) 和提取新元数据。如可以轻松理解的,检索现有元数据可以提高分类效率,如对于很少改变的文件。更进一步,效率机制可以基于分类器元数据被更新的最后的时间,例如,基于从分类器接收到的时间戳,来确定是否要调用分类器。分类服务 108 的配置中的变化,如规则变化或分类器变化,也可以触发新的分类。

[0032] 一旦获得关于项的元数据,分类模块 110 基于其元数据来分类项。也可以评估项的内容,例如,查找可以被用来对其进行分类的某些关键字 (例如,“confidential(秘密)”), 标签或关于文件的属性的其他指示符。有各种方式来分类数据。例如,当分类文件时,文件可能已经由用户针对分类手动设置,和 / 或由控制文件的业务线 (LOB) 应用程序 (例如,人力资源应用程序) 来分类。文件可以被设置为通过运行管理员脚本来分类,和 / 或使用一组分类规则来自动分类。

[0033] 一般而言,自动分类规则提供作为分类流水线阶段 108 的一部分的通用、可扩展的机制。这允许管理员等等来定义被应用于数据项以分类那些项的自动分类规则。每一个自动分类规则都激活可以确定某一数据对象集的分类并设置分类属性的分类模块 (分类器)。注意,一个分类器模块可包括确定同一数据项 (或不同的数据项) 的不同的分类属性的若干规则。此外,可以将多个分类器应用于同一数据项 ; 例如,两个不同的分类器可以各自确定文件是否具有“个人可标识信息”。可以部署两个分类器来评估同一文件,从而即使只有一个分类器确定文件包含“个人可标识信息”,文件也被如此分类。

[0034] 作为示例,一个规则可以包含的一些元素包括规则管理信息 (规则名称、标识符等等)、规则范围 (将由规则管理的数据项集的描述,如“c:\folder1 中的所有文件”), 以及描述规则在流水线中期间如何执行的规则评估选项。其他元素包括分类器模块 (对此规则用来实际指派属性值的分类器的引用)、属性 (定义由此规则指派的属性集的任选描述), 以及诸如附加执行策略之类的附加规则参数 (如额外的过滤器,如用于分类文件的内容的正则表达式,等等)。

[0035] 示例分类器模块包括 : (1) 基于数据项的位置 (例如,文件目录) 来对项进行分类的分类器, (2) 基于数据项的一些特征, (例如,基于文件所有者,查找 Active Directory ® 或 AD 中的组织单元), 通过使用全局储存库来分类的分类器, 以及 (3) 基于数据内容和数据特征 (例如,查找项的数据中的模式) 来分类的分类器。注意,这些只是示例而已,且那些熟悉本技术的人可以认识到,项的其他特征也可以用于分类不同的项,即,实际上项之间的任何相对差异可以用于分类目的。

[0036] 在一种实现中,分类器可以以各种模式操作。例如,一个“显式分类器”操作模式使分类器设置一个或多个实际属性,例如,当在文件中发现个人信息时,分类器将对应的属性“PII”设置为“Exists(存在)”等等。另一个合适的模式是“非显式分类器”,这种模式可能使分类器返回 TRUE (真) 或 FALSE (假), 例如,关于文件是否处于诸如 c:\debugger 之类的某一目录中。在 TRUE 或 FALSE 模式下,自动分类规则与每当分类器返回 TRUE 时将被

设置的属性和值相关联。如此,分类器可以设置一个或多个属性值,或调用分类器的规则可以这样做。注意,可以使用除 TRUE 或 FALSE 类型以外的分类器,例如,返回数值(例如,概率值)以提供更加细粒度的分类和分类规则的分类器。

[0037] 在进行分类之后,分类结果,以及可能其他提取的元数据,可任选地与项相关联地保存。如在图 1 中所表示的,元数据存储模块 111 执行此操作。存储允许策略以后基于分类被应用。

[0038] 注意,分类流水线模块中的每一个都是可扩展的,以使得各种企业可以定制给定实现。可扩展性允许一个以上的模块被插入到流水线的同一阶段。此外,任何一个阶段都可以并行地,或按顺序,例如,以分布式方式(跨多个机器)执行。例如,如果分类从计算上来说说是昂贵的,那么,项可以被分布到(例如,使用负载平衡技术)在不同的机器上运行的并行的分类器集,每一个并行路径的结果都被提供到策略模块。

[0039] 相对于策略,应用程序(包括那些没有被直接插入到流水线中的)可以评估分类元数据,以便对如何处理项作出策略决定。这样的应用程序包括那些执行检查项期满、审计、备份、保留、搜索、安全性、顺应性、优化等等的操作的应用程序。注意,在数据还没有被分类,或者相对于待决操作没有被分类的情况下,任何这样的待决操作都可以触发对数据的分类。

[0040] 如可以轻松理解的,不同的分类器可以导致不同且可能冲突的分类。一方面,执行属性的分类值的聚合。为此,对于每一个数据项,评估所定义的分类规则(例如,由管理员或进程)以确定分类属性。如果两个分类规则能够为一个特定分类属性设置相同的值,则聚合进程确定分类属性的最终值。如此,例如,如果一个规则导致其中属性被设置为“1”的结果而另一个规则导致该同一属性将被设置为“2”的结果,那么,在某些实施例中,所定义的聚合策略可以确定该属性的实际值应该是什么,即,“1”或“2”,或别的值。注意,在此特定情况下,一个规则不会覆盖另一个规则的属性设置,而是调用聚合策略以管理冲突。

[0041] 在另一种情况下,可以使用权威性分类器。权威性分类器是另一种类型的分类器,一般而言,它们是在不激活聚合规则的情况下可以替换其他分类器的分类器。这样的分类器可以对其结果作标志,例如,以使得它在任何冲突中胜出。

[0042] 在另一方面,提供了用于自动确定分类规则的评估顺序的机制。为此,规则评估顺序可以由管理员确定,和/或通过确定不同的规则和分类器之间的任何依赖关系来自动确定。例如,Rule-R1(规则-R1)设置分类属性 Property-P1(属性-P1),而 Rule-R2(规则-R2)使用 Classifier-C1(分类器-C1),该 Classifier-C1 使用 Property-P1 来确定 Property-P2(属性-P2)的值,那么,Rule-R1 需要在 Rule-R2 之前被评估。

[0043] 此外,是否要运行分类器可以随先前分类器的结果而定。如此,例如,可以使用很少具有假肯定的一个分类器,而每当“TRUE”时,都使用其结果。只有在权威性分类器不返回“TRUE”的情况下(例如,返回“FALSE”或可能地,指示不确定性的结果),才考虑辅助分类器(例如,被设计成能消除假否定)。另一个示例是基于预定义的“高度”在流水线中对某些分类器进行排序。例如,较低高度的分类器在较高高度分类器之前在流水线中执行。因此,在流水线中,按照递增的高度顺序,对分类器进行排序。

[0044] 图 2 示出了涉及在文件服务器 220 上实现可扩展自动分类规则的更具体的示例。一般而言,图 2 表示流水线服务的各个步骤 221-225,而不是模块;如可以看出的,这些步骤

/ 模块 221-225 分别对应于图 1 的模块 106、109-111 和 113。如此分类规则被应用于分类流水线内,并包括一个或多个数据发现模块 221(例如,扫描仪)、一个或多个元数据读取模块 222(例如,提取器和检索器)、确定分类的一组一个或多个模块 223(分类器)、存储元数据的一个或多个模块 224(设置器)和基于分类应用策略的一个或多个模块 225(策略模块)。

[0045] 也如图 2 中所表示的,可以扩展任何给定步骤中的模块的数量。例如,分类步骤为分类器提供可扩展性模型;管理员可以注册新分类器,列举现有的分类器,并取消注册不再需要的分类器。

[0046] 如此处所一般性地描述的,用于管理文件服务器上的文件的步骤包括分类文件,以及基于每一个文件的分类来应用数据管理策略。注意,文件可以被分类,以使得没有策略被应用于它。

[0047] 在一种实现中,对于文件服务器 220 上的文件的自动分类过程由该服务器 220 上定义的分类规则来驱动。当文件被存储在其中分类是活动的文件服务器中时,它被自动分类,即,没有来自用户的对文件进行分类的显式请求。可以被用来分类该特定文件服务器上的文件的各种分类标准包括:(1) 在文件服务器上运行的分类规则和分类器,(2) 保持与文件相关联的任何先前的分类结果,和/或(3) 存储在文件本身中(或其属性中)的属性。当确定给定文件的分类以提供所得属性集 232 时,评估这些准则,这些属性集 232 被存储在属性存储 234 中(但是可以存储在文件本身中)。

[0048] 在一种实现中,每一个分类规则都可以具有诸如下面阐述的那些之类的评估选项:

[0049] 只有在文件还没有被分类的情况下才进行评估;

[0050] 即使文件已经被分类,也进行评估,并考虑先前的分类属性值(例如,来自先前在同一个文件上运行的分类过程——若存在的话);即使文件已经被分类,也进行评估,但是,不考虑任何先前的分类属性值。

[0051] 作为示例,考虑由用户作为文件保存到服务器上的文件夹的文档(没有属性被指派)。自动分类规则将文件分类为具有中等业务影响,即, $\text{BusinessImpact} = \text{MBI}$ (业务影响 = MBI)。此分类也可以存储在文档内(因为文件服务器具有为这种类型的文档安装的解析器)。

[0052] 考虑文档随后被复制到另一个服务器(以及不同的文件夹)。新文件夹落入这样的分类规则:如果运行该分类规则,则将该文件夹中的文件分类为具有高业务影响 $\text{BusinessImpact} = \text{HBI}$ (业务影响 = HBI)——若该文件还没有被分类。然而,由于此文件内的属性指示 BusinessImpact 分类已经被设置为 MBI,因此文件 BusinessImpact 属性保持 MBI。

[0053] 可以修改上述规则以便即使文件已经被分类也评估文件,并可以考虑或可以不考虑文件中的属性值。在随后的分类运行中,评估规则,并且由于 HBI 高于 MBI,因此,聚合策略确定文件属性将被设置为 HBI。

[0054] 如可以看出的,每一个分类规则都依赖于用于该规则的分类器。作为另一个示例,考虑包含 `<scope>`(范围)、`<classifier>`(分类器)、`<classification property>`(分类属性)、`<value>`(值)的分类规则,其中,分类器包含用于分类文件的特定实现。例如,“`classify by folder`(按文件夹分类)”分类器允许按文件的位置来对这些文件进行

分类。此分类器检查文件的当前路径,并将它与在分类规则的 <scope> 中指定的路径进行匹配。如果路径在 <scope> 内,那么,规则指示 <classification property> 可以具有规则中指定的 <value>;(属性不一定被设置,因为多个规则可能需要被聚合,以确定对于此分类属性,实际值是什么)。注意,这是显式的分类器,因为它要求指定 <value>。

[0055] 作为不同类型的文件分类器的示例,“Retrieve classification from AD by owner(按所有者从 AD(活动目录)检索分类)”分类器读取文件的所有者,并查询活动目录,以按照规则中提到的 <classification property> 的所有者,弄清什么是正确的值。注意,这是非显式分类器,因为它确定 <value>;因而,<value> 将不在规则中指定。

[0056] 每一个分类器都可以任选地指示它为分类逻辑使用哪些属性。此信息对确定分类过程调用分类器的顺序,以及指示在调用分类器之前需要从存储 234 中检索哪些属性是有用的。

[0057] 另外,每一个分类器都可以任选地指示哪些属性被用于设置。此信息可以被用在用户界面中,用以示出了哪些属性对于此分类器是相关的(如果没有提及任何属性,那么,所有属性都相关),并用在分类过程中,其中此信息指示在调用分类器之前将从存储中检索哪些属性。对于显式的和非显式的分类器,该信息是相关的。例如:“Classify by folder(按文件夹分来)”显式分类器没有指示特定属性,“Retrieve classification from AD by owner”非显式分类器也没有指示。然而,“Determine organizational unit(确定组织单元)”非显式分类器只知道如何设置“Organizational Unit(组织单元)”属性。

[0058] 为了附加标识,可以使用任选信息来描述诸如公司名称和版本标签之类的分类器。

[0059] 分类器也可能需要消费附加参数。例如,如果分类器被构建为基于一些粒度表达式来查找文件中的个人信息,那么,那些粒度表达式不必被硬编码为分类器,而是可以从诸如定期更新的 XML 文件之类的外部源来提供。在此情况下,分类器包括指向该 XML 文件的指针。基于文件服务器资源管理器(FSRM)的分类允许为分类器指定附加参数,且当其被调用时,这些参数作为输入被传递到分类器。

[0060] 此外,由于分类器用来运行的许可级别,在不同的分类器之间,分类器运行时行为可以是不同的。一个许可级别是“本地服务”,然而,可能需要较高或较低的许可级别,例如,“本地系统”或“网络服务”。

[0061] 另一方面是分类器是否需要访问文件内容。例如,上文所描述的文件夹分类器不需要访问文件内容,因为它基于包含文件夹来进行分类。相比之下,标识文件中的特定文本或模式(例如,信用卡号)的分类器需要处理文件内容。注意,需要访问文件内容的分类器不需要以提升的特权运行,因为 FSRM 分类为分类器流送文件内容。

[0062] 下表概述了分类器的一种实现的各种特征:

[0063]

名称（唯一）
启用/禁用（默认-启用）
显式/非显式
分类器需要 FSRM 分类来为它流送文件内容吗？（默认：否）
分类器的运行时特权（默认：本地服务）
它使用的属性（任选）
它设置的属性（任选）
描述（任选）
公司名称（任选）
版本（任选）
高度水平
附加参数（任选）

[0064] 图 2 还表示允许其他外部应用程序分别为数据项获得或设置属性的 API 240、242。一般而言, Get Properties API(获得属性 API) 240 被用来在任意时间“拉 (pull)”属性 (与当在运行时向策略模块推 (push) 属性的流水线不同)。注意, 此 API 240 被示为分别在分类和存储阶段 223 和 224 之后, 以便能够获得在分类数据阶段 223 设置的任何属性。

[0065] Set Properties API(设置属性 API) 242 被用来在任意时间将属性“推”到系统中, (虽然要注意, 此 API 242 被示为与分类数据阶段 223 一起操作, 以使得属性可以在稍后在“存储属性”阶段 224 被保存 ; 即, Set Properties(设置属性) 基本上是由用户引导的人工分类)。进一步请注意, 作为分类过程的一部分, 分类器可以访问从文件中提取的供分类时使用的附加预定义文件 (例如, File.CreationTime(文件 . 创建时间) …)。这些属性可以不通过分类 API 暴露为分类属性。

[0066] 转向图 3, 通过组合经由共同的流送接口——例如经由标记为 (1) 到 (10) 的操作——与分类运行时 370 进行通信的流水线模块 361-365, 来构建包括文件夹分类器 363 的分类服务 108 的一个示例体系结构 ; 例如, 实线箭头表示 DCOM 调用。在此示例中, 每一个流水线模块 361-365 都处理 PropertyBag(属性包) 对象的流 (每个文档 / 文件, 一个属性包), 其中每一个 PropertyBag 对象都保存从前一流水线模块 (如果有的话) 累加的属性的列表。一般而言, 每一个流水线模块 361-365 的任务是基于这些文件属性来执行一些动作 (例如, 添加更多属性), 并将相同的属性包传递回到运行时 370。运行时 370 将属性包的流传递到下一流水线模块, 直到完成。

[0067] 在一个基于 FSRM 的分类服务中, 取决于敏感性, 流水线模块被以不同的方式主存。更具体而言, 不解释 / 解析用户内容的流水线模块 (如解释文件系统元数据的示例性“文件夹”分类器或涉及 AD 属性的“AD”分类器) 可以被直接主存在 FSRM 分类服务中。流水线模块处理用户所提供的内容和 / 或第三方 / 外部模块, 如解析被主存在低特权主存进

程中的,在非管理员用户帐户下运行的 Word 文档。

[0068] 图 4A 和 4B 通过流程图的示例步骤概括了各种流水线操作,其始于表示发现项的步骤 402 开始可以作为步骤 402 来操作的步骤 404 提供每一个新项,或在步骤 402 之后的任何时间提供至少一个项,选择第一项。

[0069] 步骤 406 评估所选项是否被高速缓存,并在高速缓存中保持最新。倘若如此,在流水线的其余部分不必处理项,由此,分支到步骤 407,以根据需要,基于属性来应用任何策略;注意,相应地,将策略应用于高速缓存的 / 最新的文件。对于其他项,步骤 408 和 409 重复过程,直到没有项剩余。

[0070] 如果在流水线的其余部分将处理项,则步骤 406 分支到步骤 410,该步骤表示针对该项的基本属性对项进行扫描。这些可以是文件元数据,嵌入的属性等等。

[0071] 步骤 412 表示检索与项相关联的任何现有属性。这些可以来自如上文所描述的各种存储模块,例如,嵌入和数据库模块。

[0072] 步骤 414 聚合各种属性。注意,属性有可能会发生冲突,例如,在上面的示例中,文件的分类属性可以被嵌入在文件中,也可以在外部与文件相关联。时间戳或其他冲突解决规则可以确定获胜者,或者,如果由于冲突的属性值而导致否则跳过分类,则可以强制分类。步骤 416 表示解决任何这样的冲突,例如,基于存储模块管理机构。

[0073] 过程持续到图 4B 的步骤 420,该步骤表示基于如上文所描述的分类器排序来选择第一分类器;(注意,可以只有一个分类器)。步骤 422 表示确定是否要调用所选分类器。如上文所描述的,存在为什么特定分类器可以不运行的各种原因,例如,基于以前的分类的存在,基于时间戳或其他准则,等等。如果不被调用,则步骤 422 分支到步骤 426,以检查是否将考虑另一个分类器。

[0074] 如果在步骤 422 中将调用所选分类器,则执行步骤 424,该步骤表示调用分类器,如上文所描述的,传递任何参数,然后,执行分类。还如上面所描述的,如果分类器不直接设置属性,那么,基于分类器的结果,来使用对应的规则。

[0075] 对于任何其他分类器,步骤 426 和 427 重复步骤 422 和 424 的过程。如由高度或其他排序技术所规定的,根据评估等级,选择每个其他分类器。

[0076] 步骤 430 表示基于分类在适当时聚合属性。如上文所描述的,这包括处理任何冲突,虽然聚合不适用于任何权威性分类器的分类结果。

[0077] 步骤 432 表示保存与文件相关联的属性改变,如果有的话。注意,如果文件的属性没有改变,则策略模块可以跳过策略应用。然后,过程可以返回到图 4A 的步骤 405,以应用任何策略(步骤 407)选择和 / 处理下一项——如果有的话——直到没有项剩余。

[0078] 示例性操作环境

[0079] 图 5 示出了其上可实现图 1-4 的示例的合适的计算和联网环境 500 的示例。计算系统环境 500 只是合适计算环境的一个示例,而非意在暗示对本发明使用范围或功能有任何限制。也不应该将计算环境 500 解释为对示例性操作环境 500 中示出的任一组件或其组合有任何依赖性要求。

[0080] 本发明可用各种其他通用或专用计算系统环境或配置来操作。适用于本发明的公知计算系统、环境、和 / 或配置的示例包括但不限于:个人计算机、服务器计算机、手持式或膝上型设备、平板设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产

品、网络 PC、微型计算机、大型计算机、包括任何以上系统或设备的分布式计算环境等等。

[0081] 本发明可在诸如程序模块等由计算机执行的计算机可执行指令的通用上下文中描述。一般而言,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。本发明也可以在其中任务由通过通信网络链接的远程处理设备执行的分布式计算环境中实现。在分布式计算环境中,程序模块可以位于包括存储器存储设备在内的本地和 / 或远程计算机存储介质中。

[0082] 参考图 5,用于实现本发明的各方面的示例性系统可包括计算机 510 形式的通用计算设备。计算机 510 的组件可以包括但不限于:处理单元 520、系统存储器 530 和将包括系统存储器在内的各种系统组件耦合至处理单元 520 的系统总线 521。系统总线 521 可以是若干类型的总线结构中的任一种,包括使用各种总线体系结构中的任一种的存储器总线或存储器控制器、外围总线、以及局部总线。作为示例而非限制,这样的体系结构包括工业标准体系结构 (ISA) 总线、微通道体系结构 (MCA) 总线、增强型 ISA (EISA) 总线、视频电子标准协会 (VESA) 局部总线,以及也称为夹层总线的外围部件互连 (PCI) 总线。

[0083] 计算机 510 通常包括各种计算机可读介质。计算机可读介质可以是能由计算机 510 访问的任何可用介质,并包含易失性和非易失性介质以及可移动、不可移动介质。作为示例而非限制,计算机可读介质可包括计算机存储介质和通信介质。计算机存储介质包括以存储诸如计算机可读的指令、数据结构、程序模块或其他数据之类的信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括,但不仅限于, RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘 (DVD) 或其他光盘存储、磁带盒、磁带、磁盘存储或其他磁存储设备,或可以用来存储所需信息并可以被计算机 510 访问的任何其他介质。通信介质通常以诸如载波或其他传输机制的已调制数据信号来体现计算机可读指令、数据结构、程序模块或其他数据,并包括任意信息传送介质。术语“已调制数据信号”是指具有以在信号中编码信息的方式被设定或改变其一个或多个特征的信号。作为示例而非限制,通信介质包括诸如有线网络或直接线连接之类的有线介质,以及诸如声学、RF、红外及其他无线介质之类的无线介质。上面各项中的任何项的组合也包括在计算机可读介质的范围内。

[0084] 系统存储器 530 包括易失性和 / 或非易失性存储器形式的计算机存储介质,如只读存储器 (ROM) 531 和随机存取存储器 (RAM) 532。基本输入 / 输出系统 533 (BIOS) 包含诸如在启动期间帮助在计算机 510 内的元件之间传输信息的基本例程,基本输入 / 输出系统 533 (BIOS) 通常储存在 ROM 531 中。RAM 532 通常包含处理单元 520 可立即访问和 / 或目前正在操作的数据和 / 或程序模块。作为示例而非限制,图 5 示出了操作系统 534、应用程序 535、其他程序模块 536 和程序数据 537。

[0085] 计算机 510 也可以包括其他可移动 / 不可移动、易失性 / 非易失性计算机存储介质。仅作为示例,图 5 示出了从不可移动、非易失性磁介质中读取或向其写入的硬盘驱动器 541,从可移动、非易失性磁盘 552 中读取或向其写入的磁盘驱动器 551,以及从诸如 CD ROM 或其他光学介质等可移动、非易失性光盘 556 中读取或向其写入的光盘驱动器 555。可在示例性操作环境中使用的其他可移动 / 不可移动、易失性 / 非易失性计算机存储介质包括但不限于,磁带盒、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等。硬盘驱动器 541 通常由诸如接口 540 等不可移动存储器接口连接至系统总线 521,并且磁盘驱动器 551

和光盘驱动器 555 通常由诸如接口 550 等可移动存储器接口连接至系统总线 521。

[0086] 以上描述并在图 5 中示出的驱动器及其相关联的计算机存储介质为计算机 510 提供了对计算机可读指令、数据结构、程序模块和其他数据的存储。例如,在图 5 中,硬盘驱动器 541 被示为存储操作系统 544、应用程序 545、其他程序模块 546 和程序数据 547。注意,这些组件可以与操作系统 534、应用程序 535、其他程序模块 536 和程序数据 537 相同,也可以与它们不同。操作系统 544、应用程序 545、其他程序模块 546 和程序数据 547 在这里被标注了不同的附图标记是为了说明至少它们是不同的副本。用户可通过诸如平板或者电子数字化仪 564、话筒 563、键盘 562 和定点设备 561(通常指的是鼠标、跟踪球或触摸垫)等输入设备向计算机 510 输入命令和信息。图 5 中未示出的其他输入设备可以包括操纵杆、游戏手柄、圆盘式卫星天线、扫描仪等。这些和其他输入设备通常通过耦合至系统总线的用户输入接口 560 连接至处理单元 520,但也可以由其他接口和总线结构,例如并行端口、游戏端口或通用串行总线(USB)来连接。监视器 591 或其他类型的显示设备也通过诸如视频接口 590 之类的接口连接至系统总线 521。监视器 591 也可以与触摸屏面板等集成。注意到监视器和 / 或触摸屏面板可以在物理上耦合至其中包括计算设备 510 的外壳,诸如在平板型个人计算机中。此外,诸如计算设备 510 等计算机还可以包括其他外围输出设备,诸如扬声器 595 和打印机 596,它们可以通过输出外围接口 594 等连接。

[0087] 计算机 510 可以使用到一个或多个远程计算机(如远程计算机 580)的逻辑连接,以在联网环境中操作。远程计算机 580 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其他常见网络节点,并且通常包括许多或所有以上关于计算机 510 所描述的元件,但在图 5 中仅示出了存储器存储设备 581。图 5 中所示的逻辑连接包括一个或多个局域网(LAN)571 和一个或多个广域网(WAN)573,但也可以包括其他网络。这些联网环境在办公室、企业范围计算机网络、内联网和因特网中是常见的。

[0088] 当用于 LAN 网络环境中时,计算机 510 通过网络接口或适配器 570 连接到 LAN 571。当在 WAN 联网环境中使用时,计算机 510 通常包括调制解调器 572 或用于通过例如因特网等 WAN 573 建立通信的其他手段。可为内置或可为外置的调制解调器 572 可以经由用户输入接口 560 或其他合适的机制连接至系统总线 521。诸如包括接口和天线的无线联网组件 574 等无线网络可以通过诸如接入点或对等计算机等合适的设备耦合到 WAN 或 LAN。在联网环境中,相对于计算机 510 所描述的程序模块或其部分可被存储在远程存储器存储设备中。作为示例而非限制,图 5 示出了远程应用程序 585 驻留在存储器设备 581 上。可以理解,所示的网络连接是示例性的,也可以使用在计算机之间建立通信链路的其他手段。

[0089] 辅助子系统 599(例如,用于内容的辅助显示)可经由用户接口 560 连接,从而即使计算机系统的主要部分处于低功率状态中,也允许诸如程序内容、系统状态和事件通知等数据被提供给用户。辅助子系统 599 可连接至调制解调器 572 和 / 或网络接口 570,从而在主处理单元 520 处于低功率状态中时,也允许在这些系统之间进行通信。

[0090] 结论

[0091] 尽管本发明易于作出各种修改和替换构造,但其某些说明性实施例在附图中示出并在上面被详细地描述。然而应当了解,这不旨在将本发明限于所公开的具体形式,而是相反地,旨在覆盖落入本发明的精神和范围之内所有修改、替换构造和等效方案。

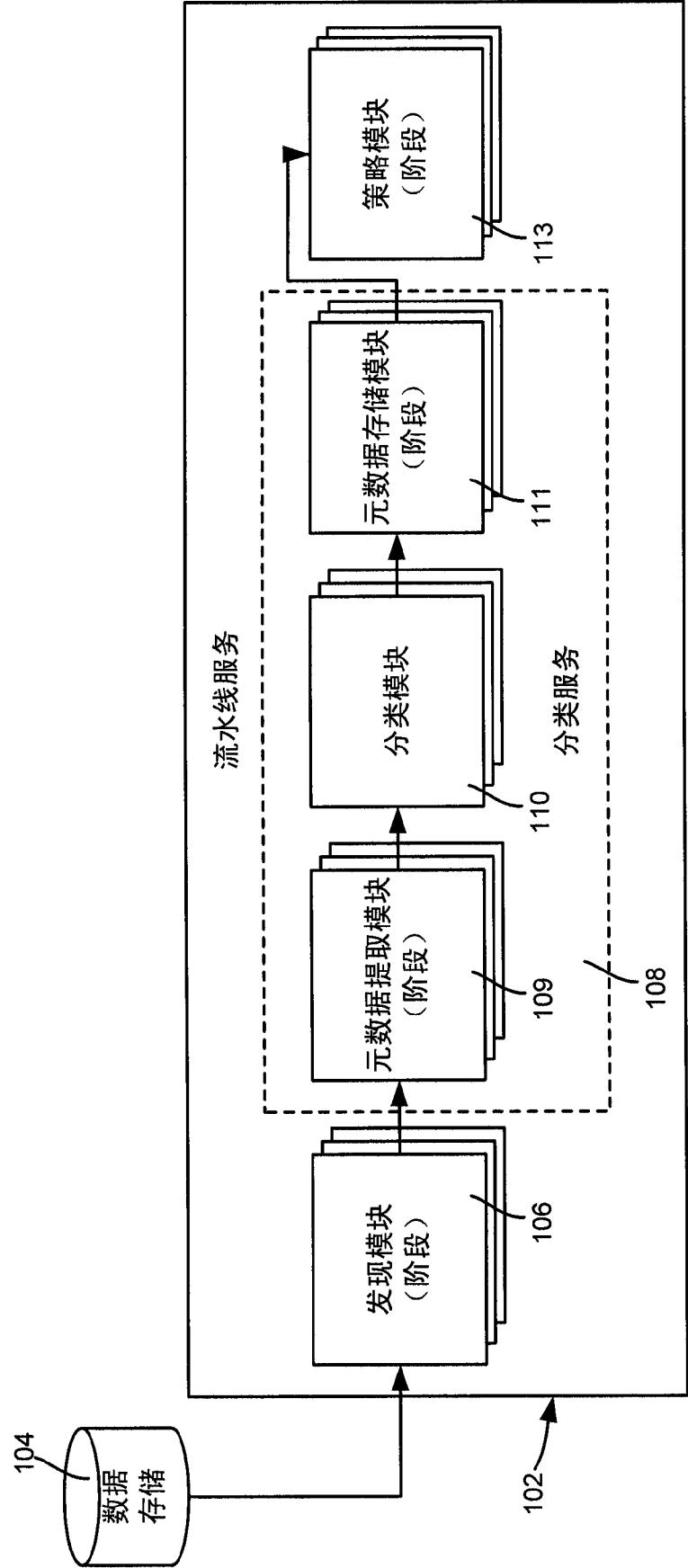


图 1

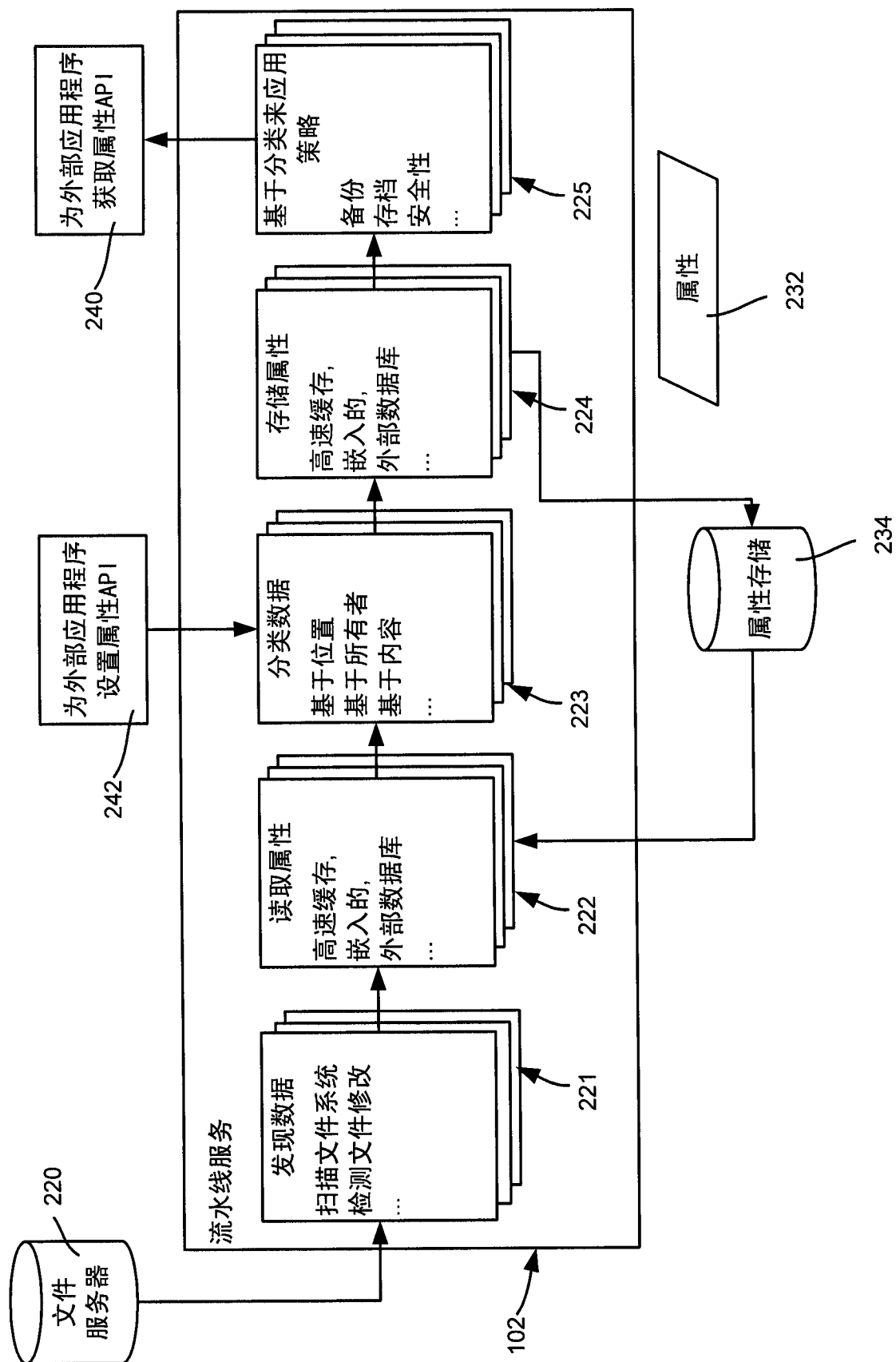


图 2

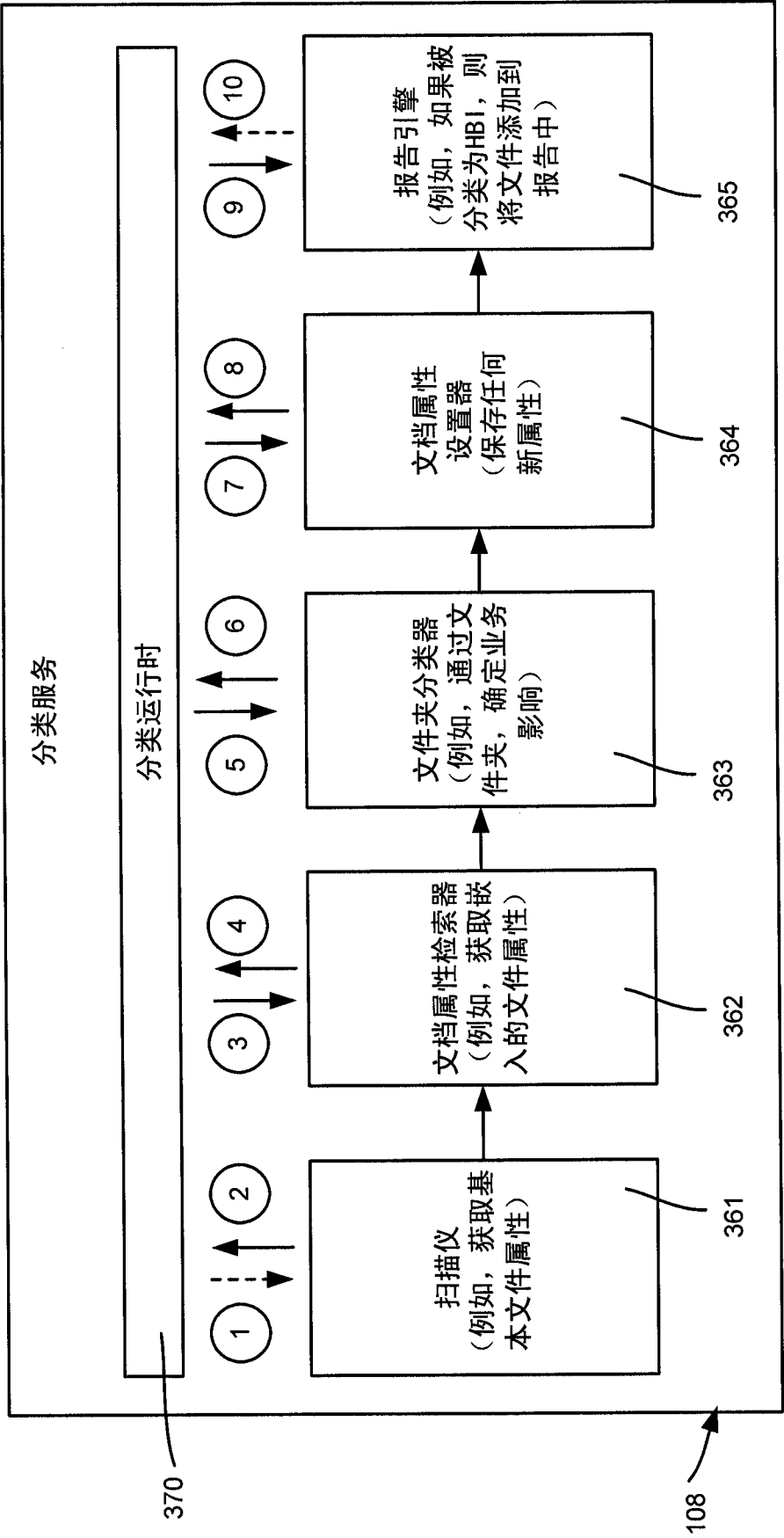


图 3

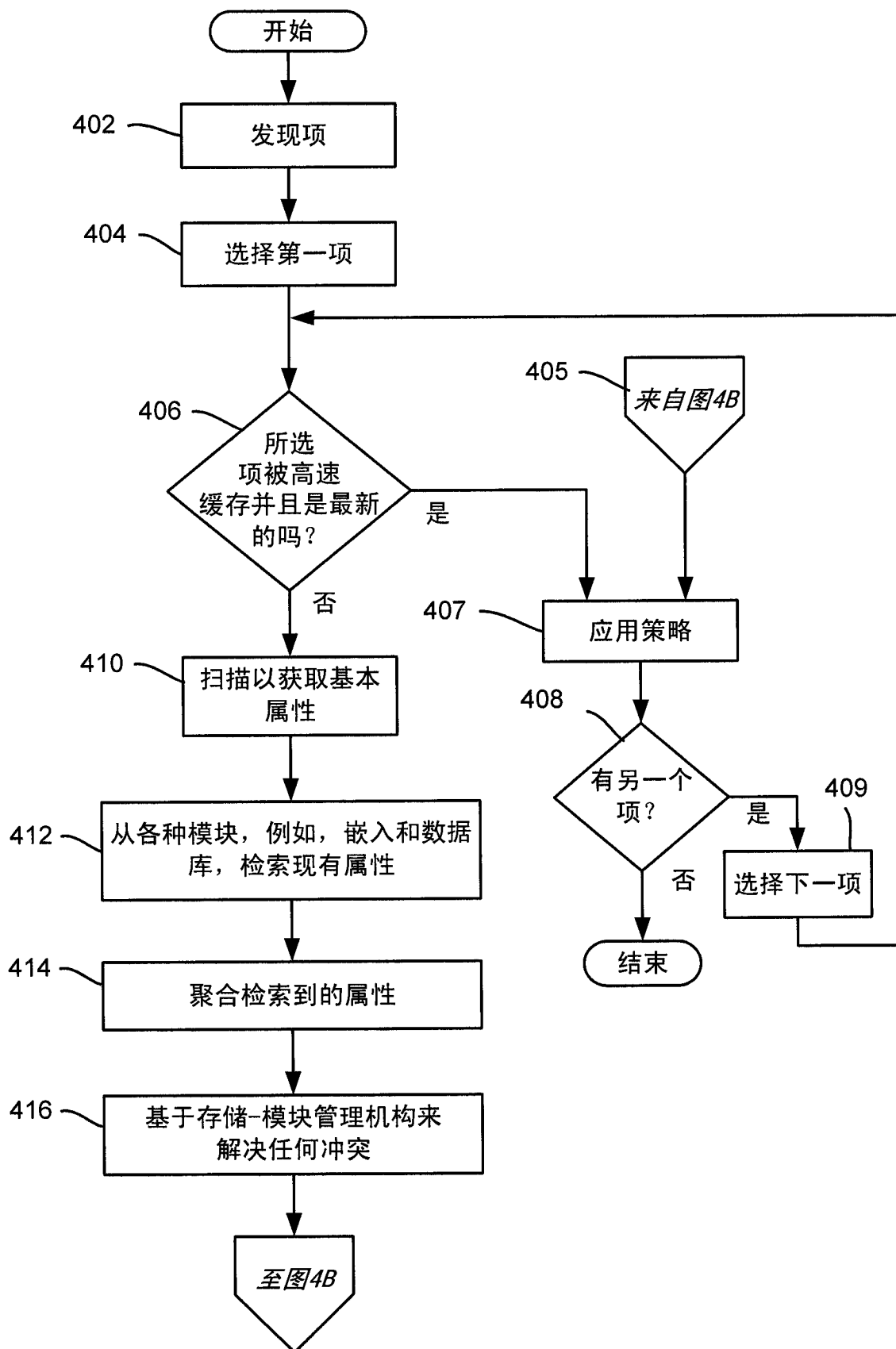


图 4A

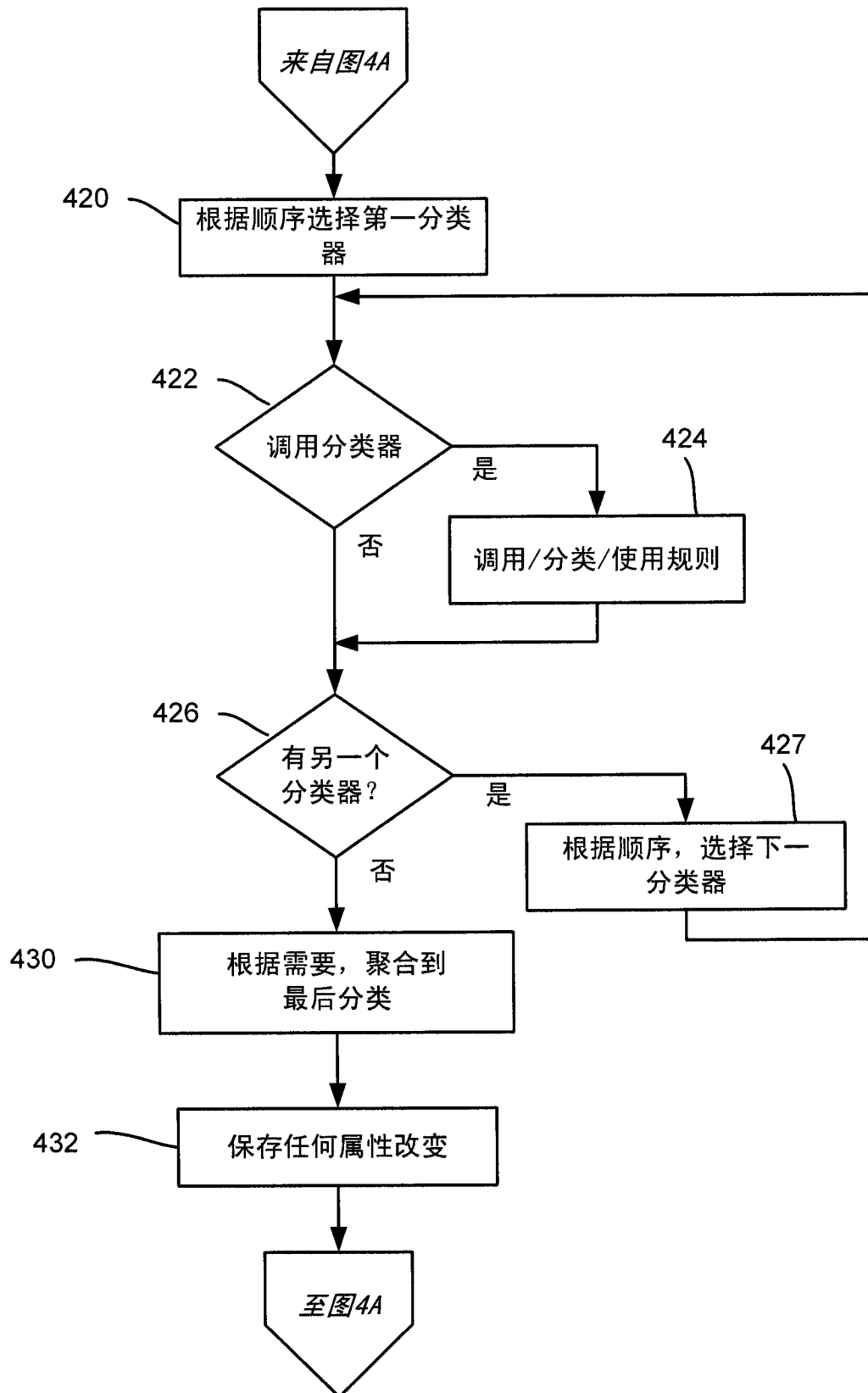


图 4B

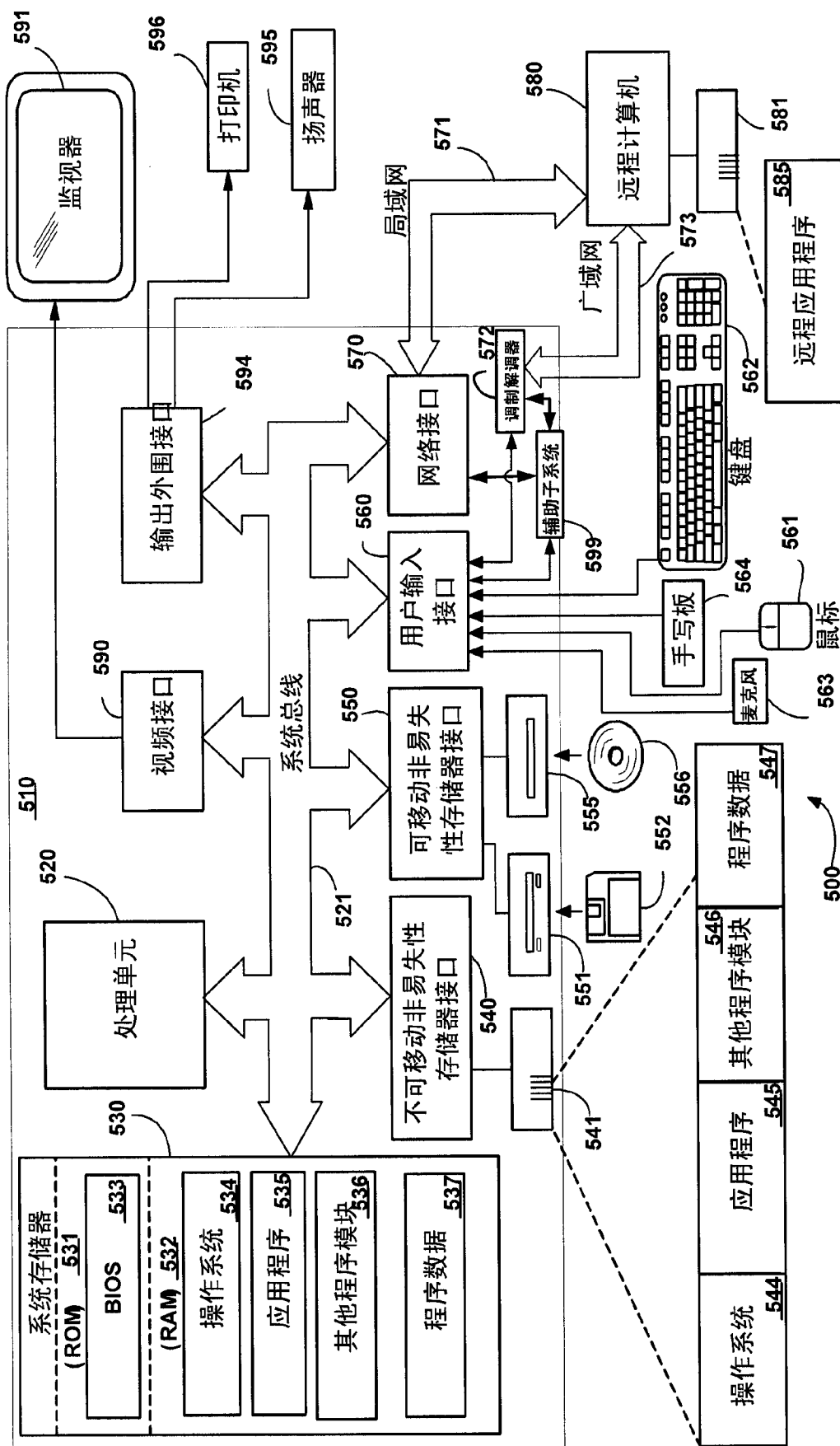


图 5