

[19] 中华人民共和国国家知识产权局

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



[12] 发明专利说明书

专利号 ZL 200410036668.8

[45] 授权公告日 2009 年 2 月 4 日

[11] 授权公告号 CN 100458773C

[22] 申请日 2004.4.29

[21] 申请号 200410036668.8

[30] 优先权

[32] 2003.4.30 [33] JP [31] 125816/2003

[32] 2004.3.10 [33] JP [31] 067779/2004

[73] 专利权人 佳能株式会社

地址 日本东京

[72] 发明人 江口贵巳 鹫泽充

[56] 参考文献

WO0049526A1 2000.8.24

US6397213B1 2002.5.28

审查员 李 琰

[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 曲 瑞

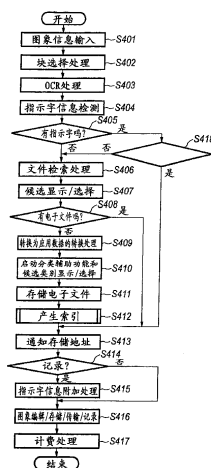
权利要求书 5 页 说明书 24 页 附图 19 页

[54] 发明名称

信息处理装置和信息处理方法

[57] 摘要

本发明提供一种信息处理装置和信息处理装置、以及存储媒体、程序。通过根据布局信息而对各属性的检索结果适应性估计检索结果，从而更正确地检索原始文档。为了实现该目的，用于检索类似于所输入的文档图象之图象数据的信息处理方法，其特征在于，包含：按每种属性将所述输入的文档图象分割为多个区域的步骤（步骤 S402）；为每个所述被分割的区域，利用适合于所述属性的检索步骤，计算出类似度的步骤；以及，对每个所述被分割的区域，对所计算出的类似度进行加权，从而计算出总类似度的步骤（步骤 S406）。



1.一种信息处理装置，用于检索类似于所输入的文档图象的图象数据，其特征在于具有：

分割单元，用于根据每种属性将所述输入的文档图象分割为多个区域；

类似度计算单元，对每个所述分割而成的区域，使用适于所述属性的检索单元来计算类似度；

总类似度计算单元，根据所述输入的文档图象中的所述分割而成的各区域的占有率，对按照每个所述分割而成的区域计算出的类似度进行加权，从而计算出总类似度；以及

布局类似度计算单元，用于计算所述输入的文档图象与比较对象的图象数据的布局类似度，该布局类似度计算单元还用于执行控制，以便对于所计算出的布局类似度等于或大于规定阈值的所述比较对象图象数据，执行由所述类似度计算单元以及所述总类似度计算单元执行的处理，而对于所计算出的布局类似度低于规定阈值的所述比较对象图象数据，不执行由所述类似度计算单元以及所述总类似度计算单元执行的处理。

2.如权利要求1所述的信息处理装置，其特征在于所述检索单元包括：

根据通过文字识别而从由所述分割单元分割的区域中的文本区域中提取出的文本数据来进行全文检索的单元；

根据所述文本数据进行概念检索的单元；以及

根据从由所述分割单元分割的区域中的图象区域中提取出的特征量来进行图象检索的单元。

3.如权利要求1所述的信息处理装置，其特征在于，还包括与用于存储与每个所述图象数据的版权版税有关的信息的服务器进行通信的通信单元。

4.如权利要求3所述的信息处理装置，其特征在于，当存在与所

述版权版税有关的信息的情况下，只要没有收到来自所述服务器的许可，就不执行对所述图象数据的处理。

5.如权利要求3所述的信息处理装置，其特征在于，还包括根据所述版权版税来执行针对用户的计费处理的计费单元。

6.如权利要求1所述的信息处理装置，其特征在于，还包括：

计费单元，用于对用户计费收取将所述输入的文档图象转换为矢量数据时的费用；

矢量数据转换单元，在执行了对所述计费单元计算出的费用的支付手续的情况下，将所述输入的文档图象转换为矢量数据；以及

存储单元，用于存储被转换为所述矢量数据的文档图象。

7.如权利要求6所述的信息处理装置，其特征在于，所述计费单元对管理被转换为矢量数据的文档图象的版权版税时的费用进行计费。

8.如权利要求1所述的信息处理装置，其特征在于，还包括：

矢量数据转换单元，用于将所述输入的文档图象转换为矢量数据；以及

候选类别输出单元，用于根据被转换为所述矢量数据的文档图象的特征，输出存储目标的候选类别。

9.如权利要求8所述的信息处理装置，其特征在于，还包括：

类别选择单元，由用户从所述候选类别输出单元所输出的候选类别中选择出存储目标的类别；以及

存储单元，按所述类别选择单元所选择的类别存储被转换为所述矢量数据的文档图象。

10.如权利要求8所述的信息处理装置，其特征在于，还包括存储单元，用于针对由所述候选类别输出单元输出的第1位候选类别，自动存储被转换为所述矢量数据的文档图象。

11.如权利要求8所述的信息处理装置，其特征在于，所述矢量数据转换单元，在根据所述总类似度计算单元的计算结果未发现与所述文档图象对应的原始数据文件的情况下，将所述文档图象转换为矢

量数据。

12.如权利要求 8 所述的信息处理装置，其特征在于，所述矢量数据转换单元，在与所述文档图象对应的原始数据文件是图象文件的情况下，将所述文档图象转换为矢量数据。

13.如权利要求 1 所述的信息处理装置，其特征在于，还包括选择装置，用于从经所述分割单元分割的多个区域中选择出用于类似度计算的区域，

其中所述类似度计算单元针对由所述选择装置选出的区域，使用适合于所述属性的检索单元，计算类似度。

14.一种信息处理方法，用于对类似于所输入的文档图象的图象数据进行检索，其特征在于，包括：

分割步骤，根据每种属性，将所述输入的文档图象分割为多个区域；

类似度计算步骤，对每个所述分割而成的区域，使用适于所述属性的检索步骤来计算类似度；

总类似度计算步骤，根据所述输入的文档图象中的所述分割而成的各区域的占有率，对按照每个所述分割后的区域计算出的类似度进行加权，从而计算出总类似度；以及

布局类似度计算步骤，计算所述输入的文档与作为比较对象的图象数据的布局类似度，该布局类似度计算步骤还用于执行控制，以便对于所计算出的布局类似度等于或大于规定阈值的所述比较对象图象数据，执行在所述类似度计算步骤以及所述总类似度计算步骤中执行的处理，而对于所计算出的布局类似度低于规定阈值的所述比较对象图象数据，不执行在所述类似度计算步骤以及所述总类似度计算步骤中执行的处理。

15.如权利要求 14 所述的信息处理方法，其特征在于，所述检索步骤包含：

根据通过文字识别而从所述分割步骤分割出的区域中的文本区域中提取出的文本数据，来执行全文检索的步骤；

根据所述文本数据执行概念检索的步骤；以及

根据从所述分割步骤所分割的区域中的图象区域中提取出的特征量来执行图象检索的步骤。

16.如权利要求 14 所述的信息处理方法，其特征在于，还包括与用于存储与每个所述图象数据的版权版税有关的信息的服务器进行通信的通信步骤。

17.如权利要求 16 所述的信息处理方法，其特征在于，当存在与所述版权版税有关的信息的情况下，只要没有收到来自所述服务器的许可，就不执行对所述图象数据的处理。

18.如权利要求 16 所述的信息处理方法，其特征在于，还包括根据所述版权版税来执行针对用户的计费处理的计费步骤。

19.如权利要求 14 所述的信息处理方法，其特征在于，还包括：
计费步骤，用于对用户计费收取将所述输入的文档图象转换为矢量数据时的费用，

矢量数据转换步骤，在执行了对所述计费步骤计算出的费用的支付手续的情况下，将所述输入的文档图象转换为矢量数据；以及

存储步骤，用于存储被转换为所述矢量数据的文档图象。

20.如权利要求 19 所述的信息处理方法，其特征在于，所述计费步骤对管理被转换为矢量数据的文档图象的版权版税时的费用进行计费。

21.如权利要求 14 所述的信息处理方法，其特征在于，还包括：
矢量数据转换步骤，用于将所述输入的文档图象转换为矢量数据；以及

候选类别输出步骤，用于根据被转换为所述矢量数据的文档图象的特征，输出存储目标的候选类别。

22.如权利要求 21 所述的信息处理方法，其特征在于，还包括：
类别选择步骤，由用户从所述候选类别输出步骤所输出的候选类别中选择出存储目标的类别；以及

存储步骤，按所述类别选择步骤所选择的类别存储被转换为所述

矢量数据的文档图象。

23.如权利要求 21 所述的信息处理方法，其特征在于，还包括存储步骤，用于针对由所述候选类别输出步骤输出的第 1 位候选类别，自动存储被转换为所述矢量数据的文档图象。

24.如权利要求 21 所述的信息处理方法，其特征在于，所述矢量数据转换步骤，在根据所述总类似度计算步骤的计算结果未发现与所述文档图象对应的原始数据文件的情况下，将所述文档图象转换为矢量数据。

25.如权利要求 21 所述的信息处理方法，其特征在于，所述矢量数据转换步骤，在与所述文档图象对应的原始数据文件是图象文件的情况下，将所述文档图象转换为矢量数据。

26.如权利要求 14 所述的信息处理方法，其特征在于，还包括选择步骤，用于从经所述分割单元分割的多个区域中选择出用于类似度计算的区域，

其中所述类似度计算步骤针对由所述选择步骤选出的区域，使用适合于所述属性的检索步骤，计算类似度。

信息处理装置和信息处理方法

发明领域

本发明涉及一种检索技术，用于从数据库中检索出与由扫描仪等输入装置读取的图象类似的图象数据。

背景技术

近年来，在呼吁环境问题的进程中，办公室中的无纸化快速发展。其中，提出了一种装置，用于通过扫描来读取在文件夹(binder)中收集的纸文档或散发资料(例如请参见专利第3017851号公报)。

但是，在上述检索装置中，由于一律要对所有图象执行所谓利用位图进行比较的方法，因此，要考虑检索效率或检索精度随着文档图象内容的不同而恶化的情况。因此，本申请人考虑了具有以下检索方法(将这种检索称为“复合检索”)的一种文件管理系统：当检索原始文档时，在从原始文档和扫描读取的图象中分别计算出文本、照片、线条等每个属性的特征量之后，综合判断文本的一致程度、照片图象的一致程度、使用各属性的布局(layout)信息的布局一致程度等多种一致程度。

在这种文件管理系统中处理的文件，从文本属性多的文件到照片、线条属性多的文件，遍及各个方面，每个文件的布局(文件内容)大不相同。于是，在上述文件管理系统中，由于在执行综合判断时，一律要估计每个属性的检索结果，因此，存在在混杂有布局大不相同的文件的环境中，不能得到充分的检索精度的问题。为了整理多种文件，提高文件的再利用性，不仅要全部文件进行综合判断，而且在文件仅有一部分不一致时也必须进行检索。即便在引用照片等的情况中，对某一部分一致的文档进行检索的需求也是显著的。例如，在同一文档的作者不同时，在必须针对各个部分适当征收版税等情况下这种检索是必不可少的。

发明内容

本发明是鉴于上述问题而作出的，目的在于通过适应性地估计各属性的检索结果，来实现更正确的原始文档的检索。

本发明的目的在于即便在仅有文件的一部分不一致的情况下也进行检索，使得重复数据的整理容易，从而实现有效的文件管理。

为了实现上述目的，本发明的信息处理装置具有以下结构，即，一种信息处理装置，用于检索类似于所输入的文档图象的图象数据，其特征在于具有：分割单元，用于根据每种属性将所述输入的文档图象分割为多个区域；类似度计算单元，对每个所述分割成的区域，使用适于所述属性的检索单元来计算类似度；总类似度计算单元，根据所述输入的文档图象中的所述分割而成的各区域的占有率，对按照每个所述分割成的区域计算出的类似度进行加权，从而计算出总类似度；以及布局类似度计算单元，计算所述输入的文档图象与比较对象的图象数据的布局类似度，该布局类似度计算单元还用于执行控制，以便对于所计算出的布局类似度等于或大于规定阈值的所述比较对象图象数据，执行在所述类似度计算单元以及所述总类似度计算单元中执行的处理，而对于所计算出的布局类似度低于规定阈值的所述比较对象图象数据，不执行在所述类似度计算单元以及所述总类似度计算单元中执行的处理。

根据本发明，通过适应性地估计各属性的检索结果，从而能够实现更正确的原始文档的检索。

即便在仅有文件的一部分不一致的情况下也进行检索，使得重复数据的整理容易，从而能够实现有效的文件管理。

通过以下结合附图所作的说明，本发明的其它特点以及优点将更加明显，在整个附图中，相同的字符表示相同或相似的部分。

附图说明

插入说明书并构成说明书一部分的附图说明了本发明的实施例，并与描述部分一起用于解释本发明的原理。

图1图示了实现有关本发明的信息处理方法的图象处理系统的结构。

图 2 是 MFP 100 的结构图。

图 3 是表示检索用索引生成处理的流程的流程图。

图 4 是显示类似原始检索处理以及登录处理的流程的流程图。

图 5A 和 5B 图示了输入原稿的一个例子。

图 6 是显示指示字信息检测处理的流程的流程图。

图 7 是检测指示字信息时使用的输入原稿的一个例子。

图 8 是一张流程图，显示了根据指示字信息来执行文件检索情况下的处理流程。

图 9 图示了对图 5 所示的输入原稿进行块选择的结果。

图 10 是一张流程图，显示了根据类似度来执行文件检索处理情况下的处理流程。

图 11 是一张流程图，显示了根据类似度来执行文件检索处理的块比较处理的细节。

图 12 是一张流程图，显示了块比较处理的块内比较处理的细节。

图 13 图示了 DAOF 的数据结构。

图 14 是一张流程图，显示了变换为应用数据的变换处理。

图 15 是一张流程图，显示了文档结构树生成处理。

图 16A 和 16B 图示了文档结构树的结构。

图 17 是一张流程图，显示了指示字信息嵌入处理。

图 18 是一张流程图，显示了第 4 实施例的文件检索处理。

图 19 是一张用于指定检索选项的用户界面画面。

具体实施方式

以下，将按照附图来详细说明本发明的最佳实施例。

第 1 实施例

以下，说明本发明的第 1 实施例。图 1 图示了实现本发明的信息处理方法的图象处理系统的结构。如该图所示，在本实施例中，构筑在办公室 10 和办公室 20 通过因特网 104 而连接的环境下、实现本发明的信息处理方法的图象处理系统。

在办公室 10 内构筑的 LAN 107 中, 分别连接有 MFP 100、控制 MFP 100 的管理 PC 101、客户 PC(具有外部存储单元) 102-1、文件管理服务器 106-1 及其数据库 105-1、计费服务器 110、以及代理服务器 103-1。同样, 在办公室 20 内构筑的 LAN 108 上, 分别连接有文件管理服务器 106-2 及其数据库 105-2、计费服务器 110-2。办公室 10 内的 LAN 107 和办公室 20 内的 LAN 108 通过代理服务器 103-1、103-2 而连接到因特网 104 上。

在具有这种结构的图象处理系统中, 没有特别限定存储作为检索对象的登记图像的场所。既可以存储在 MFP 100 内的存储装置(后述)内、管理 PC 101 的硬盘(未图示)内、也可以存储在文件管理服务器 106-1 的数据库(105-1)内。但是, 为了能够检索该存储的登录图象, 存储场所处于可经 LAN 107 或 108 从外部对其进行访问的状态。

根据本发明的信息处理方法(检索功能)可在图 1 所示的图象处理系统内的任意装置上实现。例如, 既可以使 MFP 100 内的数据处理装置(后述)具有该检索功能、使管理 PC 101 具有检索功能, 也可以使文件管理服务器 106-1 具有该检索功能。

MFP 100 承担纸文档的图象读取处理以及针对读取的图象信号的图象处理的一部分, 利用 LAN 109 将图象信号输入到管理 PC 101 内。管理 PC 是普通 PC, 其内部具有图象存储单元、图象处理单元、显示单元、以及输入单元, 但是, 其一部分也可以集成在 MFP 100 上。

图 2 是 MFP 100 的结构图。图 2 中, 含有自动送稿器(以下称为 ADF)的图象读取部 200 利用图中未示的光源照射推叠的或是 1 张纸文档(原稿), 通过透镜将原稿的反射影像成像在固体摄像元件上, 并从固体摄像元件获得一个光栅状的图象信号作为 600 DPI 密度的图象信息。在通常的复印处理的情况下, 利用数据处理装置 205 将该图象信号转换为记录信号, 在需要复印多张的情况下, 将一页的记录数据暂时存储保持在记录装置 202 内之后, 顺序输出到记录装置 202 后在纸上形成图象。

从客户 PC 102 发送的打印数据经由 LAN 107 以及网络 IF 204,

在数据处理装置 205 内被转换为可记录的光栅数据后,由记录装置 202 在纸上形成记录图象。操作者通过 MPF 100 内装备的键操作部(输入装置 203)或是管理 PC 101 的键盘以及鼠标对 MPF 100 发出指示,这一系列操作是由数据处理装置 205 内的图中未示的控制部来控制的。

另一方面,操作输入的状态显示以及处理中的图象数据的显示是由显示装置 206 来执行的。存储装置 201 也可以由客户 PC 101 来控制,这些 MPF 100 和客户 PC 101 的数据交换和控制也可以通过网络 IF 207 或直接经 LAN 109 来执行。

处理概要

接下来,利用图 3 和图 4,来说明图 1 所示的图象处理系统中的、包含依据本发明的信息处理方法而执行的处理的图象处理的总体概要。

为了便于说明,以下,假定具有登录图象的数据库构筑在客户 PC 内的硬盘(图中未示)上,而且通过使管理 PC 101 和 MPF 100 一起动作来实现的本发明的信息处理方法。

(生成检索用索引)

图 3 是一张流程图,显示了文件检索用索引生成处理的流程。首先,输入索引生成对象文件作为原稿(步骤 S301)。该原稿可以是图象读取部 200 输入印刷物并将其位图化后形成的(位图图象),可以是利用文档编辑应用程序做成的电子数据,也可以是按照客户 PC 102 的硬盘或 CD 驱动器、DVD 驱动器、FDD 驱动器等存储媒体内存储的应用程序的固有格式的电子数据。假定该原稿通过用户的操作而被存储在数据库的不同类别的文件夹内。例如,通过按照每个体育项目、产品小册子或商业文件等种类(类别),将原稿存储到各个文件夹内,从而容易执行文件管理。

在不是位图图象的应用数据的情况下(步骤 S302 中“NO”的情况下),分支到步骤 S303,在管理 PC 101 的 CPU 中执行光栅化。所谓

光栅化,是对应用数据进行分析并产生图象数据的处理,一般而言,这是生成打印图象时广泛执行的处理。

接下来,执行块选择(区域分割处理)(步骤 S304)。在步骤 S304 中,将区域分离成文字/艺术线条部分以及半色调图象部分。通过将文字部分分离为基于段落将文字分组而成的块,或者,分离为由线构成的表或图形,从而执行相应的分段化。另一方面,将用半色调表现的图象部分分割为针对每个数据块独立的所谓对象(object),诸如被分离为矩形的块的图象部分、背景部分等。而且,对文字块执行 OCR 处理(步骤 S305)。

接下来,在输入图象中,检测对应于作为附加信息记录的 2 维条码或 URL 的对象,通过对 URL 利用 OCR 进行文字识别,通过对 2 维条码解读该标记,从而检测出分别存储该原稿的原始电子文件的存储装置内的指示字信息(步骤 S306)。作为附加指示字信息的方法也可以是不直接可视的所谓的电子水印方法等,例如像在文字与文字之间嵌入信息的方法,或嵌入半色调图象内的方法等。

在步骤 S306 中,在没有检测出指示字信息的情况下,分支到步骤 S307,使用上述任意一种方法来嵌入指示字信息。但是,在检索处理中,指示字信息的嵌入不是必须的处理。

接下来,利用作为图象部分分割的对象来产生图象检索用索引(步骤 S308)。在这里,记录图象检索中所必需的图象特征量,例如是平均色。还可将该图象的色分布等作为特征量来使用。

接着,利用作为文字部分分割的对象来产生概念检索用索引(步骤 S309)。在这里,在后述的概念检索中,分析所输入文章的意思、文理,并找到与该内容相关的文档。由此,与以往的检索不同,即便在不了解明确的关键词的情况下,也能检索文件。在概念检索索引产生处理中使用的文字在原稿为位图数据的情况下,输入在步骤 S305 中执行文字识别所得到的文字,而在应用数据的情况下,输入存储在应用程序内的文字。

接下来,产生全文检索用索引(步骤 S310)。全文检索用索引产生

处理中使用的文字，是与上述概念检索索引产生处理中使用的文字相同的文字。

接着，在步骤 S311 中，学习各个种类的特征(单词的出现频率等)、产生分类用特征矢量。学习各种文件夹内存储的索引产生对象文件的特征(单词的出现频率等)，产生、更新各种文件夹的分类用特征矢量。如此，通过预先提取各类文件夹的特征，当在步骤 S411 中存储在后续图 4 的 S409 中被矢量化了的输入图象时，能够在步骤 S410 中，提示作为存储目标的推荐文件夹候选单。

之后，存储产生的索引信息(步骤 S312)。存储目标可以是连接在图 1 中的办公室 10 或 20 的 LAN 107、108 上的文件管理服务器 106-1、106-2 内的数据库 105-1、105-2 内、或 MFP 100 自身具有的存储装置 201 等。在本实施例中，假定是存储在客户 PC 101 内的硬盘上的。

(类似原稿检索或登录)

接下来，使用图 4 来说明类似原稿检索处理以及登录处理。首先，操作图 2 的 MFP 100 的图象读取部 200，光栅状扫描一张原稿，从而得到 600 DPI-24 比特的图象信号。之后，对于该图象信号利用数据处理装置 205 执行前处理，并作为 1 页的图象数据保存在存储装置 201 内(步骤 S401)。或者是，对作为应用数据存储的原始电子文件进行光栅化，从而得到 600 DPI-24 比特的图象信号。对于该图象信号在数据处理装置 205 中实施前处理，并作为 1 页的图象数据保存在存储装置 201 中(步骤 S401)。

管理 PC 101 的 CPU 将所存储的图象信号区域分离为文字/线条部分和半色调的图象部分。将文字/线条部分分离为进一步按段落分块而成的每块，或者，分离为由线构成的表、图形，从而执行相应的分段化。另一方面，将用半色调表现的图象部分分割为所谓按每个数据块独立的对象(步骤 S402)，诸如像按照分离为矩形的块的图象部分、背景部分等。

对文件块执行 OCR 处理(步骤 S403)。此时，在输入图象中，检

测出适用于作为附加信息记录的 2 维条码或 URL 的对象，通过利用 OCR 对 URL 进行文字识别，通过对 2 维条码解读该标记，从而检测出存储各个原稿的原始电子文件的存储装置内之指示字信息(步骤 S404)。另外，作为附加指示字信息的方法，也可以是不直接可视的所谓的电子水印方法等，例如像在文字与文字之间嵌入信息的方法，或嵌入半色调图象内的方法等。

在检测出指示字信息的情况下(步骤 S405 中的“YES”的情况下)，分支到步骤 S418，从指示字所指示的地址中检索原始电子文件。在本实施例中，原始电子文件存储在管理 PC 101 内的硬盘内，按照步骤 S404 中得到的地址信息对该硬盘进行检索。原始电子文件的存储场所并不仅限于此，也可以是以下任意一种：图 1 中的客户 PC 102 内的硬盘内；或连接在办公室 10 或 20 的 LAN 107、108 上的文件管理服务器 106-1、106-2 内的数据库 105-1、105-2；或是 MFP 100 自身具有的存储装置 201。在步骤 S418 中，在没有看到原始电子文件的情况下、在看到代表 PDF 或 tiff 的所谓的图象文件的情况下、或是在指示字信息自身不存在的情况下(步骤 S405 中的“NO”的情况下)，分支到步骤 S406。

在步骤 S406 中，为了检索数据库上的原始电子文件，首先，在步骤 S403 中将经受了 OCR 的文字块分割为单词，并提取关键词。作为关键词提取是这样一种功能：分析文字部分，根据单词的出现频率等仅仅提取预定数目的所确定的单词，作为关键词。该关键词作为全文检索的检索词进行输入。从文字块中提取文字，并利用文章开头对文章全体的说明等性质来产生摘要。所产生的摘要文作为概念检索的检索词进行输入。另外，对于图象块，作为作为图象数据的个别的图象文件执行图象检索。接下来，研究与数据库上的各原始电子文件的相似度，以检索原始电子文件。本发明的特征在于：为每个对象求取相似度，按照每个对象的文件内占有率而向文件全体的相似度反映该对象的相似度。即，通过较大地向文件全体的相似度反映文件内占有比例大的对象的相似度，从而有可能自适应地应对各种格式的文件。

对于能够对全体区域计算类似度，因此，也可以看出部分一致的文件。

步骤 S406 中的检索处理结果，在发现类似度高的原始电子文件的情况下，显示简图(步骤 S407)，如果必须由操作这从多个原始电子文件中选择的情况下，通过操作者的输入操作，来执行原始电子文件的指定。

另外，在候选文件是 1 个文件的情况下，也可以自动从步骤 S408 分支到步骤 S413，并通知存储地址。在步骤 S406 的检索处理中没有看到原始电子文件的情况下，或是，在看到的是代表 PDF 或 tiff 的所谓的图象文件的情况下，从步骤 S408 分支到步骤 S409。

在步骤 S409 中，执行从图象转换为矢量数据的变换处理。例如，在步骤 S403 中，对于经过 OCR 处理的文字块，进一步识别文字的大小、种类、字体，并使用预先为每种文字类别(字体类别、种类)准备的布局数据，对所述 OCR 结果的文字码进行矢量化。就文字块而言，识别文字图象的轮廓，通过沿着文字图象的轮廓提取布局而对文字区域进行矢量化，且，此时，也可以存储 OCR 处理结果，作为与各文字的矢量数据对应的文字码。对于由线构成的表、图形块进行矢量化，能够识别表等图形形状的话，则能识别该图形。另外，出于图象块是图象数据的原因，将其作为单独的 JPEG 文件来进行处理。这些矢量化处理是针对逐个对象执行的，并进一步保存各对象的布局信息，将其转换为例如是嵌入了矢量数据的 rtf (Rich Text Format)形式。通过在可以使用如此产生的 rtf 形式的文件的应用中执行读入，不用说，能够将文章恢复为能够编辑的状态，而且，也能够将图形或图象数据恢复为能够编辑的状态。变换的应用用数据文件并不仅限于可嵌入矢量数据的对象之 rtf 形式，例如，也可以是转换为 SVG (Scarable Vector Grafics)格式等其它的文件形式。

接下来，在步骤 S410 中启动了分类辅助功能。所谓分类辅助功能，是在存储文件时，辅助将其分类并存储到哪一个种类的文件夹内的功能。根据该功能，能够大幅减轻有关使用者的文件登录的作业量。从文件的文字部分来产生分类用特征矢量，并按照与索引内存储的分

类用特征矢量的类似度判断，从上级的类别文件夹开始顺序显示类似度。如此，通过显示登录目的单元的推荐的候选类别文件夹，使得使用者能够容易地执行登录处理。另外，使用者还能够有所显示的类别文件夹中进行选择，作为文件的登录目的单元，也可以产生未显示的类别文件夹、或是新的类别文件夹，并能够将作为电子文件而转换的数据存储在管理 PC 101 内的硬盘内(步骤 S411)。

在步骤 S409 中，经矢量化了的输入图象，在此后执行相同的检索处理时，被作为原始电子文件检索出来，在步骤 S412 中，执行与前述的 S308 - S311 相同的索引产生处理，产生索引信息，并追加到检索用索引文件内。

在能够通过步骤 S418 由指示字信息来指定原始电子文件的情况下，以及在 S408 的检索处理中能够指定原始电子文件的情况下，以及能够通过 S409 - 412 中进行矢量化和存储的文件来进行向电子文件的转换的情况下，为了操作者可以直接指定原始电子文件，在步骤 S413 中，将该原始电子文件的存储地址通知操作者。

再者，在步骤 S414 中，如果判断出操作者想执行的处理是“记录”，则分支到步骤 S415，将指示字信息作为图象数据追加到文件内。

使用如此得到的电子文件自身，从而能够在步骤 S416 中执行文件的加工、存储、传送、以及记录。与使用图象数据的情况相比，这些处理削减了信息量，存储效率高，传送时间短，另外，在记录显示时，能够非常优秀地作为高品质的数据。

在步骤 S417 中，根据计费服务器 110-1 或 110-2 中存储的、按每个文件定义的版权版税信息，来执行与检索结果的显示方法/访问方法、使用者处理方法相应的计费处理。

以下，将详细说明各处理块。首先，就步骤 S402 中显示的块选择处理进行说明。

(块选择处理(区域分割处理))

使用图 5A 和 5B 来说明步骤 S402 的块选择处理。图 5A 是在步

骤 S401 中读取的一页图象数据的一个例子, 图 5B 显示了将该图象数据逐对象地识别为象素块 (pixel cluster), 针对该每个数据块判断其文字/图形/照片/线/表/等属性, 并将其分割为具有不同属性的区域。

以下说明块选择处理的实施例。首先, 将输入图象进行黑白二值化, 执行轮廓线追踪, 提取由黑象素轮廓包含的象素块。对于面积大的黑象素块, 对内部存在的白象素执行轮廓追踪, 提取出白象素块, 又进一步从一定面积以上的白象素块内再次提取出黑象素块。

按大小以及形状对如此得到的黑象素块分类, 将其分到具有不同属性的区域中。例如, 假定: 将纵横比接近 1、大小位于一定范围内的块作为与文字相对应的象素块; 另外, 将相邻文字排列整齐的、被分组化的部分作为文字区域(TEXT); 将扁平的象素块作为线区域(LINE); 将一定大小以上的、四角形的、其中包含整齐排列的白象素块的黑象素块所占有的范围作为表区域(TABLE); 将形状不定的象素块散乱的区域作为照片区域(PHOTO); 将除此之外的任意形状的象素块成为图形区域(PICTURE)等。

将块选择处理中得到的各块的块信息用作后面将要说明的矢量化、或检索处理用信息。

(指示字信息的检测)

接下来, 将就用于从图象信息中提取出原始电子文件的存储位置的指示字信息检测处理(步骤 S404)进行说明。

图 6 是一张流程图, 显示了对输入图象中附加的 2 维条码(QR 码符号(code symbol))进行解码, 并输出数据文字列的过程。在图 7 中显示了附加 2 维条码的原稿的一个例子。

首先, 利用 CPU (未图示)来扫描表示存储于数据处理装置 205 内的页存储器中的原稿 301 的图象, 根据在先说明的块选择处理的结果来检测规定的 2 维条码符号 703 的位置(步骤 600)。QR 码的位置检测图案, 由位于符号 4 角中的 3 角上之相同的位置检测要素图案构成。

接下来, 恢复与位置检测图案相邻的格式信息, 并获取应用于该

符号的纠错水平以及主图案(S601)。

在确定符号的型号(S602)后,通过使用利用格式信息所得到的主图案,对编码区域的比特图案(bit pattern)进行 XOR 运算,从而消除了屏蔽(mask)处理(步骤 S603)。

另外,按照与该模型对应的配置规则,读取符号字符、并恢复消息的数据以及纠错码字(步骤 S604)。

执行在恢复的代码上是否存在错误的检测(步骤 S605),在检测出错误的情况下,分支到步骤 S606,并对其进行校正。

根据经过校正的数据的模式指示符和文字数指示符,将数据代码字分割为段(步骤 S607)。

最后,根据式样模式对数据文字进行解码,并输出结果(步骤 S608)。

另外,组合在 2 维条码内的数据,表示对应的原始电子文件的地址信息,例如,由文件服务名称以及由文件名称构成的路径信息,或者是由去往对应的原始电子文件的 URL 构成。

在本实施例中,尽管是对利用 2 维条码而添加了指示字信息的原稿 701 进行说明,但是,在直接用文字串记录指示字信息的情况下,利用先前的块选择处理来检测遵循规定规则的文字串的块,通过文字识别显示该指示字信息的文字串的各文字,从而能够直接得到原始电子文件的地址信息。

另外,也可以通过对于图 7 的原稿 701 的文字块 702、或 704 的文字串,在相邻文字和文字的间隔等上施加难以目视程度的调制,并在该文字间隔内嵌入信息,从而提供了指示字信息。如果在执行后述的文字识别处理时,在各文字间隔中检测出所谓的水印信息,则得到指示字信息。此外,也可以在自然图形 705 中添加指示字信息作为电子水印。

(利用指示字信息而执行的原始电子文件检索)

接下来,使用图 8 的流程图,来说明图 4 的步骤 S405 和步骤 S418

所示的、根据指示字信息的原始电子文件的检索处理。

首先，基于指示字信息中包含的地址信息来指定文件服务器(步骤 S800)。在本实施例中，尽管文件服务器指示的是管理 PC 101 内的硬盘，但是，在原始电子文件的存储目标是客户 PC102、文件管理服务器 106-1、106-2 内的数据库 105-1、105-2 或是 MFP 100 内的存储装置 201 的情况下，也可以指示这些部件。地址是由 URL、或由服务器名称和文件名称构成的路径信息。

如果能够指示文件服务器则对文件服务器传送地址(步骤 S801)。文件服务器在接收到地址后，检索该原始电子文件(步骤 S802)。在原始电子文件不存在的情况下(步骤 S803 中“No”的情况下)，将其通知给 MFP 100。

在原始电子文件存在的情况下(步骤 S803 中“YES”的情况下)，如图 4 所说明的那样，在通过原始电子文件的地址(步骤 S413)的同时，如果使用者所希望的处理是图象文件数据的获取，则将原始电子文件传送给 MFP 100(步骤 S808)。

(OCR 处理)

接下来，将就步骤 S403 中的 OCR 处理(文字识别处理)进行说明。

在文字识别部中，使用图案匹配方法来对按照文字单位切割出的图象进行识别，从而得到所对应的文字码。该识别处理，是将从文字图象所得到的特征转换为数十维的数值序列后的观测特征矢量，与每个字种事先所获取的辞典特征矢量进行比较，并将距离最近的字种作为识别结果进行处理。特征矢量的提取有各种公知的方法，例如，有这样一种方法：将文字分割为网格(mesh)状，并按照方向将各网格内的文字线作为线素，将如此计数后所得到的网格数目的维数矢量作为特征。

对在块选择处理(步骤 S402)中提取出的文字区域执行文字识别的情况下，首先对该区域判断是横向书写还是竖向书写，按各对应方向切割出行，之后，再切割出文字，从而得到文字图象。横写、竖写的判断，可以通过取得象素值在该区域内的水平/垂直投影，在水平投

影的方差大的情况下判断为横写区域，在垂直投影的方差大的情况下判断为竖写区域。在横写的情况下，通过利用水平方向的投影来切割出行，再根据相对于切出的行在垂直方向的投影切出文字，来进行文字串和文字的分解。对于竖写的文字区域，只要使水平和垂直相反即可。另外，能够检测出此时文字的大小。

(文件检索)

接下来，利用图 10 来说明图 4 的步骤 S406 所示的文件检索处理的细节。

文件检索，是利用由前述块选择处理(步骤 S402)所分割的各个块信息来执行的。具体而言，它使用了复合检索，复合了各块的属性以及文件中的块坐标信息的比较，即借助布局的比较，以及按照文件内的各块的属性而使用不同的比较方法的每块的内部信息的比较。

图 10 是一张流程图，用于将经过图 4 的步骤 S406 中的块选择处理的输入文件顺序与已经存储的原始电子文件的检索用索引相比较，并检索出与其一致的原始电子文件。首先，利用数据库对原始电子文件进行访问(步骤 S101)。将输入文件的各块和原始电子文件的各块进行比较，对于输入文件的每块求取与原始电子文件的块的类似度(步骤 S1002)。

这里，使用图 5A、图 9、图 11 以及图 12，对步骤 S1002 的块比较进行说明。将图 5A 作为输入文件，在图 9 中显示了原始电子文件的例子。图 9 的原始电子文件被分割为块 B'1-B'9，并分别进行了矢量化处理。图 11 是一张流程图，用于将输入文件中的一个块与原始电子文件进行比较，并计算出类似度。当为每个块计算出类似度时，首先，选择出被推定为与输入文件的该块布局上一致的原始电子文件的对象块。在该处理中，对于输入文件的多个块，也可以重复选择原始电子文件的对象块。接下来，求出该块和对象块的布局信息的类似度。比较块的位置、大小、以及属性(步骤 S1102、1103、1104)、并从该误差中求出布局的类似度。接着，执行块内部的比较。由于在比较块内

部时是作为相同属性进行比较的,因此,在属性不同的情况下,要执行将单方的块再矢量化为一致的属性等前处理。对于通过前处理作为相同属性处理的输入文件的块以及原始电子文件的对象块,执行块内部比较(步骤 S1105)。在块内部比较中,为了采取最适合块属性的比较方法,随属性不同其比较方法也不同。图 12 是一张流程图,记述了每块的比较方法。例如,通过前述的块选择处理,将块分割为文本、照片、表、线条等属性。在处理的块是照片图象块的情况下(步骤 S1201 的“YES”的情况),则基于从图象中提取出的特征矢量,利用特征空间上的误差来计算类似度(步骤 S1202)。这里,所谓的特征矢量,可以举出由色直方图或色矩等有关颜色的特征量、同现矩阵(日文:共起行列)、对比度、熵(entropy)、Gabor 变换等所表现的结构特征量、傅利叶描述符等形状特征量等许多种,使用这多个特征量中最合适的组合。检索结果(按照类似度顺序排列的候选清单)存储于规定的存储装置内,例如是存储在管理 PC 101 的存储装置内(步骤 S1203)。

在文本块的情况下(步骤 S1204 的“YES”的情况下),由于利用 OCR 处理来判断文字码,因此,在步骤 S1205 中执行摘要文的产生,并执行概念检索。在步骤 S1206 中,将检索结果存储在存储装置内。接下来,在步骤 S1207 中,执行关键词的提取,并执行全文检索。将检索结果存储于存储装置内(步骤 S1208)。

如此,计算出块位置、大小、属性、块内部的类似度,并通过合计各类似度,从而可以对输入文件的该块计算出其类似度,并记录该块类似度。对于输入文件的所有块,反复执行一连串的处理。通过整体合并求出的块类似度,来求出输入文件的类似度(步骤 S1003)。

接下来,说明总类似度计算处理(步骤 S1003)。对于图 5A 的输入文件的块,设算出的每块的类似度为 n_1 、 n_2 、……。此时,文件整体的总类似度 N 用下式表示。

$$N = w_1 \times n_1 + w_2 \times n_2 + \dots + \gamma \quad \dots \quad (1)$$

这里,设 w_1 、 w_2 、……是估测各块的类似度的加权系数;设 γ 是校正项,例如,是作为对于图 9 的原始电子文件的输入文件之对象

块的未选出之块的估计值等。加权系数 w_1 、 w_2 、.....是利用块的文件内占有率来求出的。例如,当设块 1 的大小为 S_1 时,则块 1 的占有率 w_1 就这样算出:

$$w_1 = S_1 / (\text{所有块的大小之和}) \quad \dots\dots(2)$$

通过使用这种占有率的加权处理,文件内占有大区域的块的类似度被进一步反映在文件全体的类似度中。

以上,在类似度高于阈值 Th 的情况下,将检索结果作为候选而保存(步骤 S1005),并由简略图等来显示原始电子文件(步骤 S407)。如果要操作者从这么多中选择一个,则通过操作者的输入来指定原始电子文件。

(转换为应用数据的转换处理)

但是,对一页的图象数据进行块选择处理(步骤 S402),矢量化处理(步骤 S408)的结果被转换为图 3 所示样子的中间数据形式的文件,可以将这种数据形式称为文件分析输出格式(DAOF)。

图 13 图示了 DAOF 的数据结果。在图 13 中,1301 是头部,用于保持与处理对象的输入文件有关的信息。在布局描述数据部 1302 中,保持按照输入文件中的 TEXT(文字)、TITLE(题目)、CAPTION(字幕)、LINEART(线条)、EPICTURE(自然画)、FRAME(框)、TABLE(表)等属性而被识别的各块的属性信息和其矩形地址信息。在文字识别描述数据部 1303 中,保持文字识别 TEXT、TITLE、CAPTION 等的 TEXT 块所得到的文字识别结果。在表描述数据部 1304 中,存储了 TABLE 块的结构细节。图象描述数据部分 1305 从输入文件中切出 PICTURE 或 LINEART 等块的图象数据,并对其进行保存。

这种 DAOF 不仅仅作为中间数据,而且还存在其自身被文件化后保存的情况,在这种文件的状态下,在所谓的一般的文件产生应用中,不能再使用各个对象。因此,接下来,对从该 DAOF 转换为应用数据的处理(步骤 S409)进行说明。

图 14 是一张流程图,显示了转换为应用数据的处理(步骤 S409)

的详细流程。

在步骤 S1400 中, 执行 DAOF 数据的输入。在步骤 S1401 中, 产生作为应用数据基础的文档结构树。在步骤 S1402 中, 以文档结构树为基础, 读入 DAOF 内的实际数据, 并产生实际的应用数据。

图 15 图示了文档结构树产生处理(步骤 S1401)的详细流程。图 16A 和 B 是文档结构树的说明图。作为全体控制的基本规则, 处理的流程从微块(因此块)转向宏块(块的集合体)。

以后, 所谓的块是指全体微块以及宏块。在步骤 S1501 中, 以纵向方向的关联性为基础, 以块单位执行再分组。在开始后, 按照微块单位进行判断。

这里, 所谓的关联性, 是能够用距离近、块宽(横向的情况下指高)基本相等等来定义。

参照 DAOF 来提取距离、宽度、高度等信息。图 16A 是实际的页结构, 图 16B 是其文档结构树。步骤 S1501 的处理结果是, T3、T4、T5 作为一个组 V1 而产生, T6、T7 作为一个处于同阶层的组 V2 而首先产生。

在步骤 S1502 中, 检验是否有纵向的隔离物。隔离物例如是物理上具有 DAOF 中的线属性的对象。在逻辑意义上说, 它是应用中明确分割块的要素。这里, 在检测出隔离物的情况下, 在相同阶层中执行再分割。

在步骤 S1504 中, 利用组长来判断是否有大于它的分割。这里, 在纵向的组长成为页高度(该页中存在的多个块的最上端和最下端的距离)的情况下, 文档结构树的产生结束。

在图 16A 的情况下, 由于在组 V1、V2 中没有隔离物, 且组高度也不是页高度, 因此转到步骤 S1504。在步骤 S1504 中, 以横向的关联性为基础, 以块单位执行再分组。这里, 在开始后立即执行的第一个操作也是以微块单位来执行判断。关联性以及判断信息的定义与纵向的情况相同。

图 16A 的情况下, 产生 H1 作为由 T1、T2 组成的一个组, 产生

H2 作为由 V1、V2 组成的一个组，H1 和 H2 分别是位于 V1 和 V2 的上一层的同阶层组。

在步骤 S1505 中，校验横向是否有隔离物。在图 16A 中，由于存在 S1，因此，将其登录到树中，产生所谓的 H1、S1、H2 阶层。

在步骤 S1506 中，利用组长来判断是否存在大于它的分割。这里，在横向的组长为页宽的情况下，文档结构树的产生结束。不是这样的情况下，返回步骤 S1501，在更高一级的阶层中，从纵向关联性校验开始反复执行。

在图 16A 的情况下，由于分割宽度为页款，因此，在此结束，最后，在向文档结构树中添加表示整个页的最高阶层的 V0。在文档结构树完成后，以该信息为基础，在步骤 S1402 中产生应用数据。

在图 16A 的情况下，具体情况如下所述。即，由于 H1 在横方向上有 2 个块 T1 和 T2，因此，作为 2 列，在输出 T1 的内部信息(参见 DAOF，文字识别结果的文章、图象等)后，改变列，并在 T2 的内部信息输出后，输出 S1。

由于 H2 在横方向有 2 个块 V1 和 V2，因此，作为 2 列输出，V1 按照 T3、T4、T5 的顺序输出其内部信息，此后，改变列，并输出 V2 的 T6、T7 的内部信息。通过上述操作，执行了向应用数据的转换处理。

(指示字信息的添加)

接下来，对步骤 S415 所示的、指示字信息附加处理进行说明。在由检索处理指定了应当处理的文件的情况下，或者在通过矢量化而能够再现原始电子文件的情况下，在记录处理该文件的情况中，通过在向纸上进行记录时添加指示字信息，从而在使用该文件再次执行各种处理时，能够简单获取原始电子文件。

图 17 是一张流程图，它显示了利用 2 维条码(QR 码符号：JIS X0510) 701 对作为指示字信息的数据文字串进行编码，并将其添加到图象中的过程。

引入到 2 维条码内的数据表示对应的原始电子文件的地址信息，它例如可以由文件服务器名称和文件名称构成的路径信息构成。或者是，由通往对应的原始电子文件的 URL、对应的原始电子文件的存储的数据库内或者是 MFP 100 自身具有的存储装置内所管理的文件 ID 等构成。

首先，为了识别进行编码的各种不同的文字，对输入数据串进行分析。另外，选择误差检测以及误差校正水平，选择输入数据能够收容的最小型号(步骤 S1700)。

接着，将输入数据串转换为规定的比特串，并根据需要添加表示数据类型(数字、西文数字、8 比特字节、汉字等)的标识符、终端图形。另外，还将其转换为规定的比特码字(步骤是 1701)。

此时，为了执行纠错，需要按照型号以及纠错水平将码字串分割为规定的块数，一个块一个块地产生纠错码字，并附加到数据代码字的后面(步骤 S1702)。

连接在步骤 S1702 中所得到的各块的数据代码字，并后续各块的纠错码字以及根据需要后续剩余的码字(步骤 S1703)。

接下来，在矩阵中配置了位置检测图案、分离图案、定时图案以及位置组合图案等，同时还配置了码字模块(步骤 S1704)。

另外，对于符号的编码化区域，选择最合适的主图案，通过 XOR 运算将屏蔽处理图案转换为在步骤 S1704 中所得到的模块(步骤 S1705)。

最后，在步骤 S1705 中得到的模块上，产生格式信息以及型号信息，完成 2 维码符号(步骤 S1706)。

如上所述，例如在将原始电子文件作为打印数据从客户 PC 102 送出，并通过记录装置 202 在纸上形成为记录图象的情况下，组合有地址信息的 2 维条码在数据处理装置 205 内被转换为可记录的光栅数据后，附着到光栅数据上的规定位置上而形成图象。这里，形成有图象的纸被分发给用户，用户通过利用图象读取部 200 进行读取，能够利用前述步骤 S404，从指示字信息中检测出原始电子文件的存储地

点。

另外，出于相同目的添加附加信息的装置能够使用各种方法，除了本实施例中说明的 2 维条码外，例如还能够使用直接利用文字串在文件中添加指示字信息的方法；调制文件内的文字串、特别是文件和文件的间隔，并嵌入信息的方法；以及将其嵌入文件中的半色调图象内的方法等一般称为电子水印的各种方法。

(计费时的版权校验处理)

接着，来说明步骤 S417 所示的计费处理。在本实施例中，是以能够自由访问此前存储在文件服务器内的所有原始电子文件并且原始电子文件全体、或其部分的对象全都能够再利用为前提进行说明的。但是，在所处理的输入文件中，应当限制由第三者导致的再利用。因此，对包含限制第三者的再利用的输入文件的情况下的步骤 S417 中的处理进行说明。

在步骤 S416 中，在执行文件的打印等处理前，针对在从指示字信息中检索出原始电子文件时能够指定的原始电子文件，询问在计费服务器 110-1、110-2 中存储的、按照每个文件/对象定义的版权版税信息。作为版权版税信息，是与检索结果的显示方法、访问方法、使用者处理方法相对应的费用的对应表。例如记载有图象区域的打印花费 100 日元，复印花费 200 日元、或者文本部分可以自由地再使用等内容。

计费服务器 110-1、110-2 调查该原始电子文件的版税信息，在有费用的情况下，向 MFP 100 请求发送用户 ID、口令。

MFP 100 提示操作者输入用户 ID、口令，并将输入的用户 ID、口令传送给计费服务器。

计费服务器 110-1、110-2 参照所传送的口令，在一致的情况下，许可步骤 S416 的所希望的处理(记录等的使用)，并执行相应的计费处理(步骤 S417)。用于执行计费的确认方法，并不仅限制为使用口令的方法，例如也可以使用指纹确认等一般广泛使用的活体确认、利用卡

确认等所有确认方法。

另外,尽管在本实施方式中,显示了通过在纸文档上添加指示字信息来指定原始电子文件的情况的实施例,但是,在利用图4的步骤S406-S407所示的所谓的检索处理来指定原始电子文件的情况中,也可以执行相同的处理。

即便对于图4的步骤S409-412所说明的矢量化处理,也能够执行计费处理。即,在根据扫描纸文档所得到的图象信息而执行矢量化处理时设定为有计费的情况下,在执行矢量化处理前,仅仅提示用户执行矢量化处理所需的费用。之后,执行支付的手续(输入用户ID、口令),在确认成功的情况下,执行矢量化处理,由此,能够从使用者处收取费用。在执行矢量化处理时,也可以组合仅执行矢量化处理所需的费用以及其后仅管理版权版税所需的费用而进行计费。

(第2实施例)

在上述第1实施例中,在文件检索中,要将摘要文字代入概念检索、将关键词代入全文检索,因此,在处理能力低下的环境中,对于摘要文字、关键词的产生中存在得不到资源的情况。这种情况下,也可以输入全部文本作为概念检索的检索词、并省略全文检索。

(第3实施例)

在上述第1实施例中,包含了检测、嵌入指示字信息的过程,但是,即便将其限制为不执行指示字信息的检测而仅仅是复合检索处理,本发明的有用性也没有丧失。

根据上述实施例,包含在块选择前执行光栅化的步骤,但这不是必需的。索引的产生也可以不在检索时执行,而是通过执行调度(scheduling)来执行。尽管在实施例中,每当输入文档时,就产生分类矢量,但是也可以在文档累积到某种程度时执行。

在实施例中,每当输入文档时,就产生分类用特征矢量,但是,也可以在文档累积到某种程度时就执行。分类用特征矢量的产生尽管

是使用文字部分执行的，但也可以就表、图象等其它属性来产生分类用特征矢量，与检索时相同，利用区域的面积比来综合地产生分类用特征矢量。在上述实施例1中，在步骤S410中，提示推荐候选文件夹作为存储目标，并允许用户进行选择，但是在设置了是否自动进行分类的设定按钮，并预先由用户设定为自动分类的情况下，也可以不显示选择项而自动登录到第1位的推荐候选文件夹内。

(第4实施例)

在第1实施例的文件检索中，在比较输入文档和数据库文件时，对于所有输入文档的全部块，就布局信息以及块的内部信息执行比较。在本实施例中，在输入文件的布局 and 数据库文件的布局全部不同的情况，可以省去块内部信息的比较处理，从而减轻了检索处理的计算量。图18是本实施例的文件检索的流程图。首先，从数据库中对原始电子文件进行访问(步骤S1801)。将输入文件的各块的位置、大小、属性和原始电子文件的各块进行比较，求出其类似度，从而求出文件整体的布局类似度(步骤S1802)。在步骤S523中，判断布局类似度是否高于规定阈值，在布局类似度低于规定阈值的情况下，不执行块内部信息的比较而转到步骤S1807。另一方面，当判断出布局类似信息高于规定阈值时，转到步骤S1804，并执行块内部信息的比较，按照先前求出的布局类似度以及块内部信息的类似度，来求出文件整体的总类似度。根据每块的类似度来求取总类似度的方法是与图10的步骤S1003相同的处理，因此省略对其的说明。在步骤S525中，判断总类似度是否高于规定阈值，如果高于规定阈值，则将该比较对象的文件作为候选进行保存(步骤S1806)。另一方面，如果在规定阈值以下，则不作为候选。在步骤S1807中，如果存在下一个比较对象的数据库文件，则转到步骤S1801，否则结束处理。

求取块内部信息的类似度的处理，由于执行照片块等信息量多的块的比较，因此，与布局信息的比较相比，一般处理负担较重。因此，通过缩小作为布局信息的程度文件，执行检索处理量的减轻、处理的

高速化，能够更有效地检索所希望的文件。

(第5实施例)

在第5实施例中，通过使用用户指定成为文件内特征的块区域，并指定从比较对象中除去的块，指定文件内的特征，从而更高精度地执行文件检索。

图19是用于令用户指定检索对象的用户界面画面(1901)的例子。输入文件通过块选择处理而被分割为多个块，在输入画面中显示文件上的文本、照片、表、线条等各块区域(1911-1917)。

用户可以从所显示的块中选择成为该输入文件特征的块，可以由用户指定要强调的块。此时选择的块也可以是多个。例如，在选择块1914的状态下，当按下按键“强调”(1902)时，执行强调块1914的检索处理。在根据每块的类似度来求出文件整体的类似度的计算式(1)中，通过增大所指定的块1914的加权系数并减小除此之外的块的加权系数而实现该检索处理。如果多次按压“强调”按钮(1902)，则所选择的块的加权系数进一步增大，从而执行更加强调块的检索。

如果按压排除按钮(1903)，则在省略了所选块1914的状态下实施检索。例如，在自动提取了错误的块的情况下，用户通过将该块排除在外，可以省略无用的检索处理，且能够防止出现错误的检索结果。在排除的情况下，也可以将算式(1)的该块的加权系数设为零。

另外，可以利用细节设置按钮(1904)来实现块的属性变更，通过将块的属性修正为正确的属性，能够更正确地进行检索。利用细节设置按钮1904，可以由用户精细地调节用于优先进行块的检索的权重。如此，在检索时，通过用户指定和设置作为特征的块，能够执行检索的最优化。

另一方面，还要考虑随着文件的不同其布局特殊的情况。有关这种文件，通过选择布局优先按钮(1905)，从而可以执行强调布局的文件检索。这种情况下，可以通过更加强调布局的类似度的结果的方式添加加权值，从而实现这一点。

如果选择了文本优先按钮(1906),则仅仅使用文本块来进行检索,并实现了处理的减轻。

如此,通过由用户选择图象的特征,从而执行强调文件特征的检索。

(其它实施例)

本发明既适用于由多种设备(例如主机、接口设备、读取器、打印机等)构成的系统,也适用于由一个设备构成的装置(例如复印机、传真机装置等)。

不用说,本发明的目的是将记录了实现前述实施例的功能的软件的程序代码提供给系统或装置,通过该系统或装置的计算机(或 CPU、MPU)读出存储媒体内存储的程序代码,从而实现其功能。

这种情况,从存储媒体中读出的程序代码自身实现所述的实施例的功能,从而存储有该程序代码的存储媒体构成了本发明。

作为用于提供程序代码的存储媒体,例如能够使用软(注册商标)盘、硬盘、光盘、磁光盘、CD-ROM、CD-R、磁盘、非易失性存储卡、ROM 等。

通过执行计算机读出的程序代码,不仅实现了前述实施例的功能,而且不用说还包括以下情况:在计算机上运行的 OS(操作系统)等根据该程序代码的指示,执行实际处理的一部分或全部,并借助于该处理而实现了前述实施例的功能。

另外,不用说,还包括以下情况:从存储媒体中读出的程序代码,在写入插入计算机的功能扩充板或连接在计算机上的功能扩展单元内具有的存储器内之后,该功能扩充板或功能扩展单元内具有的 CPU 等根据该程序代码的指示,执行实际处理的一部分或全部,并借助于该处理来显示前述实施例的功能。

本发明并不仅限于上述实施例,在本发明的主旨和范围内可以有可中改变和修改。因此,为了通知公众本发明的范围而制作了后续的权利要求书。

图1

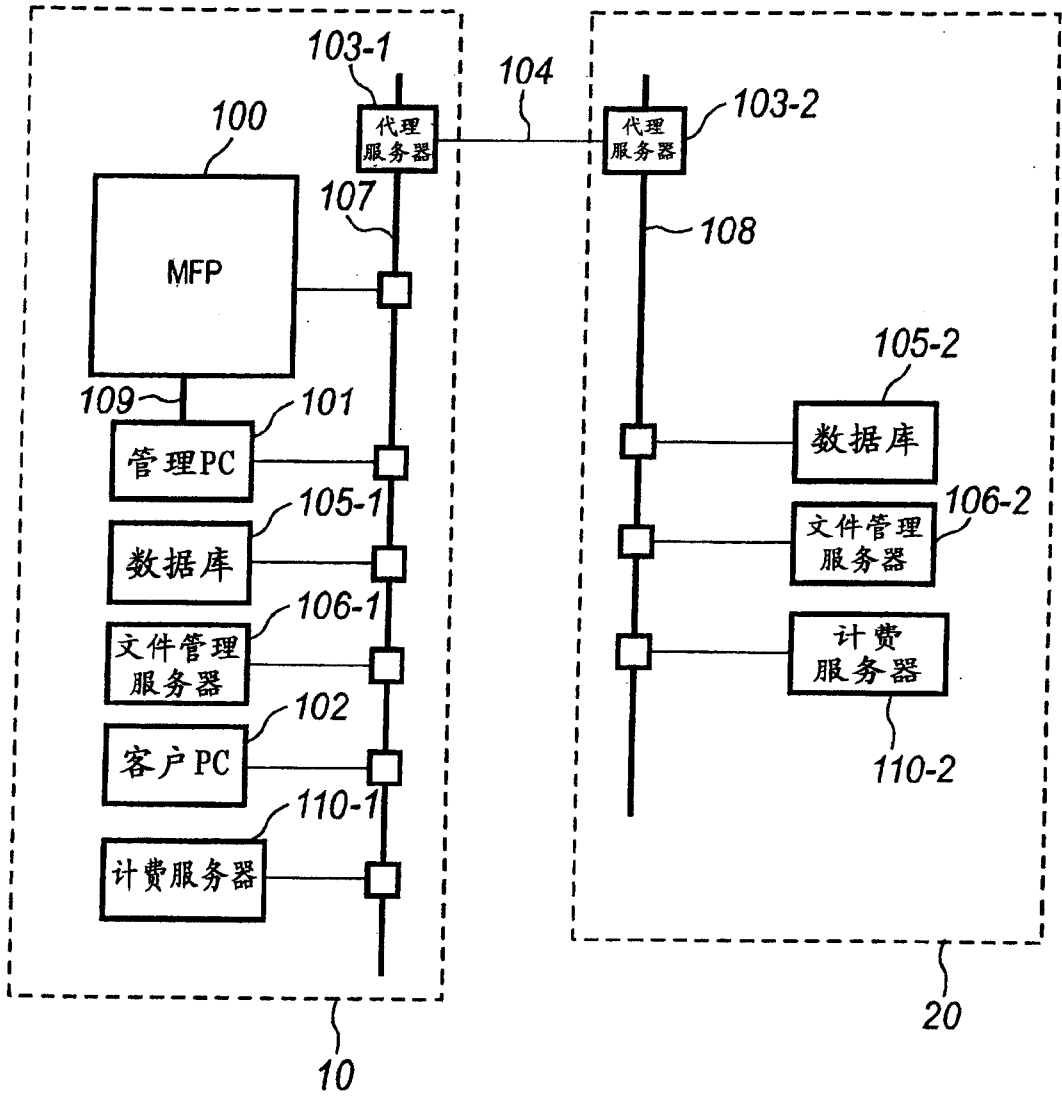


图 2

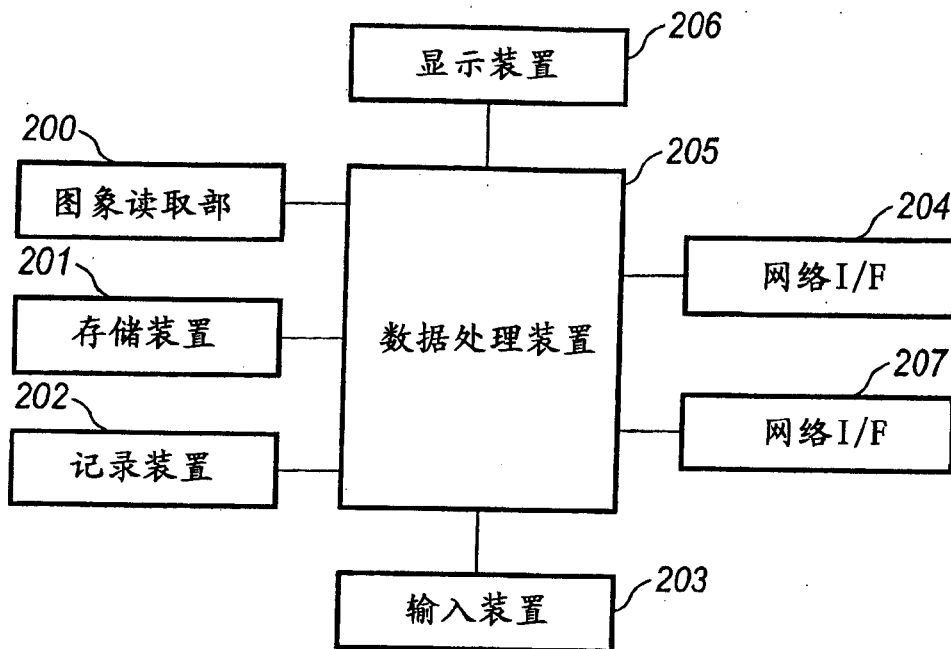


图3

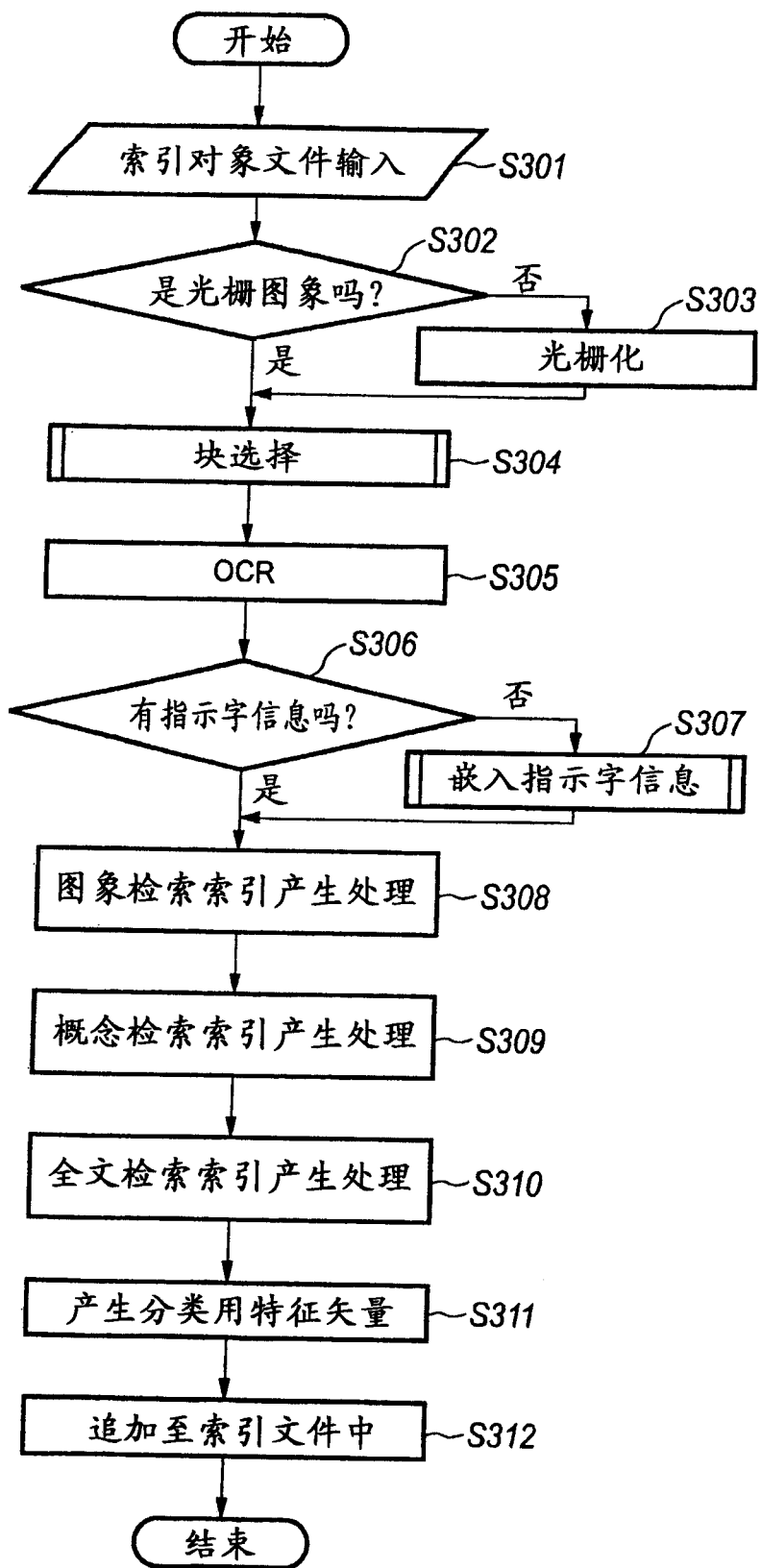
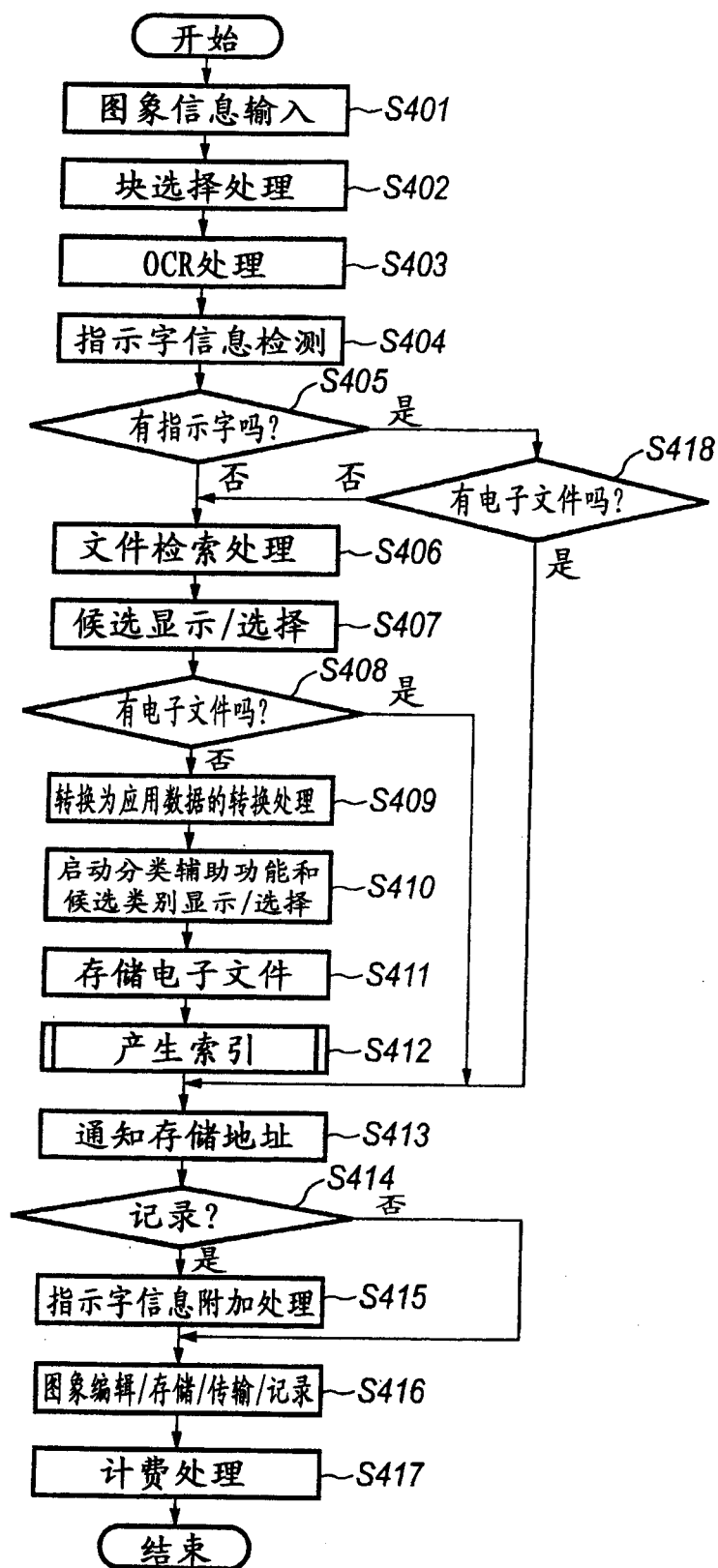


图4



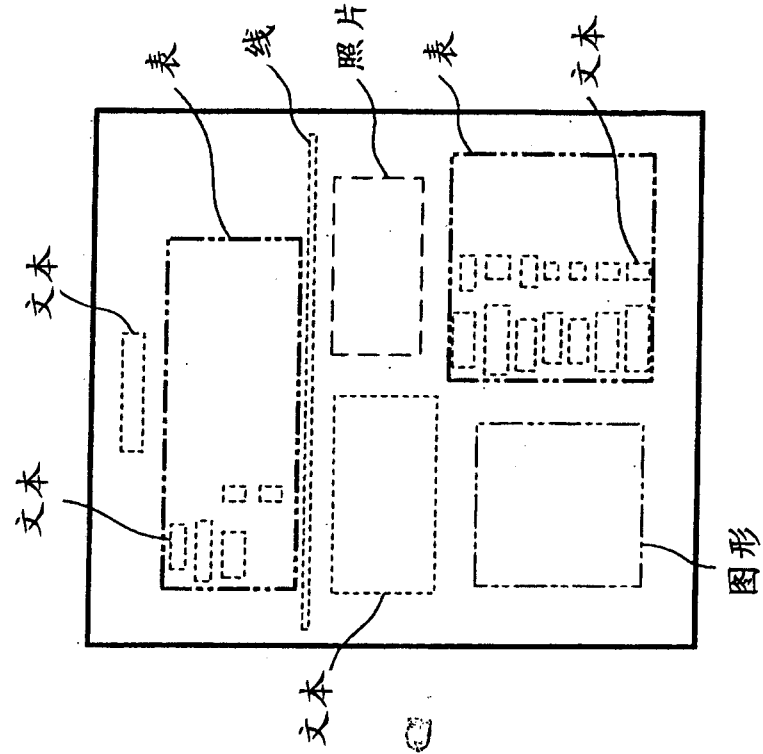


图 5B

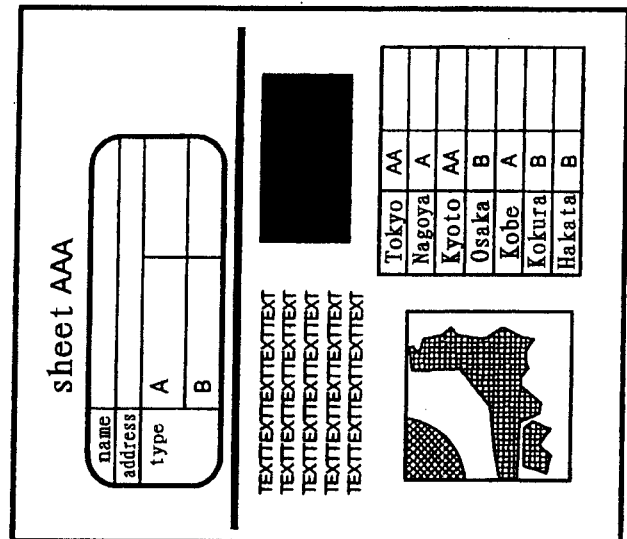


图 5A

图6

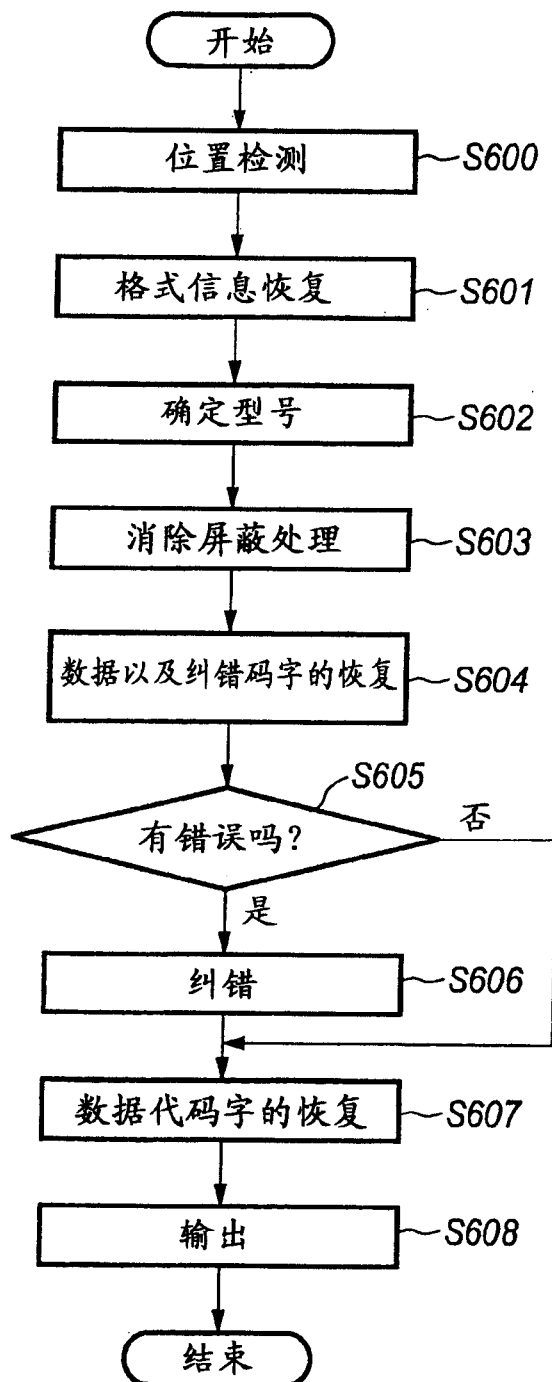


图 7

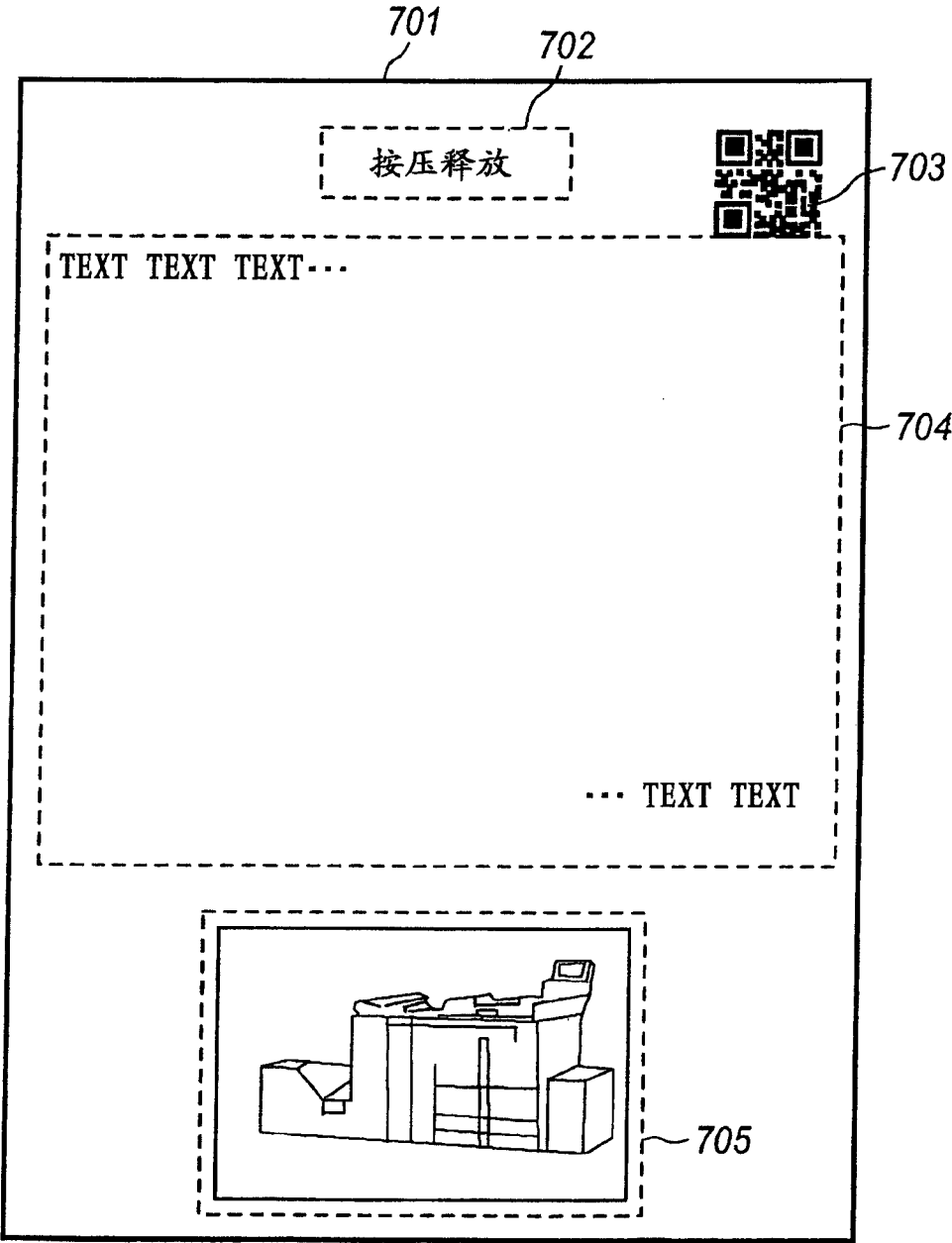
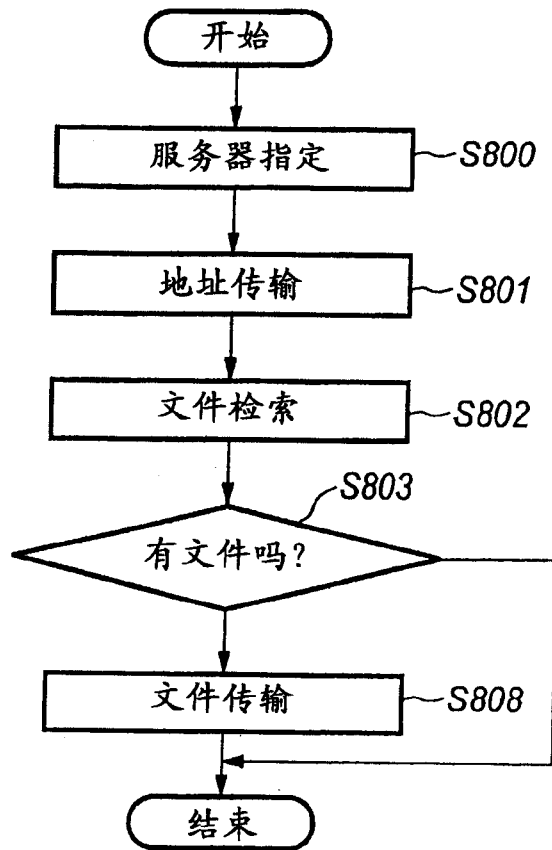


图8



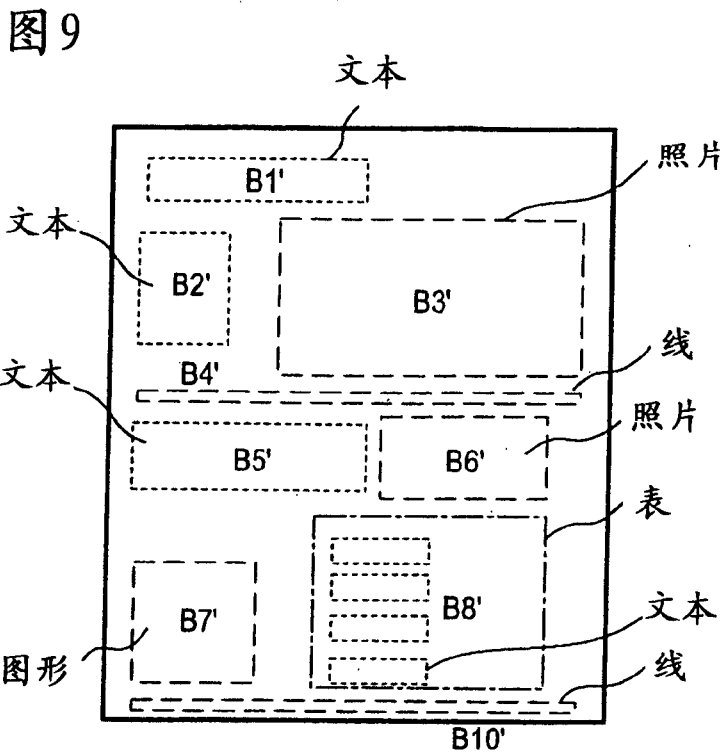


图10

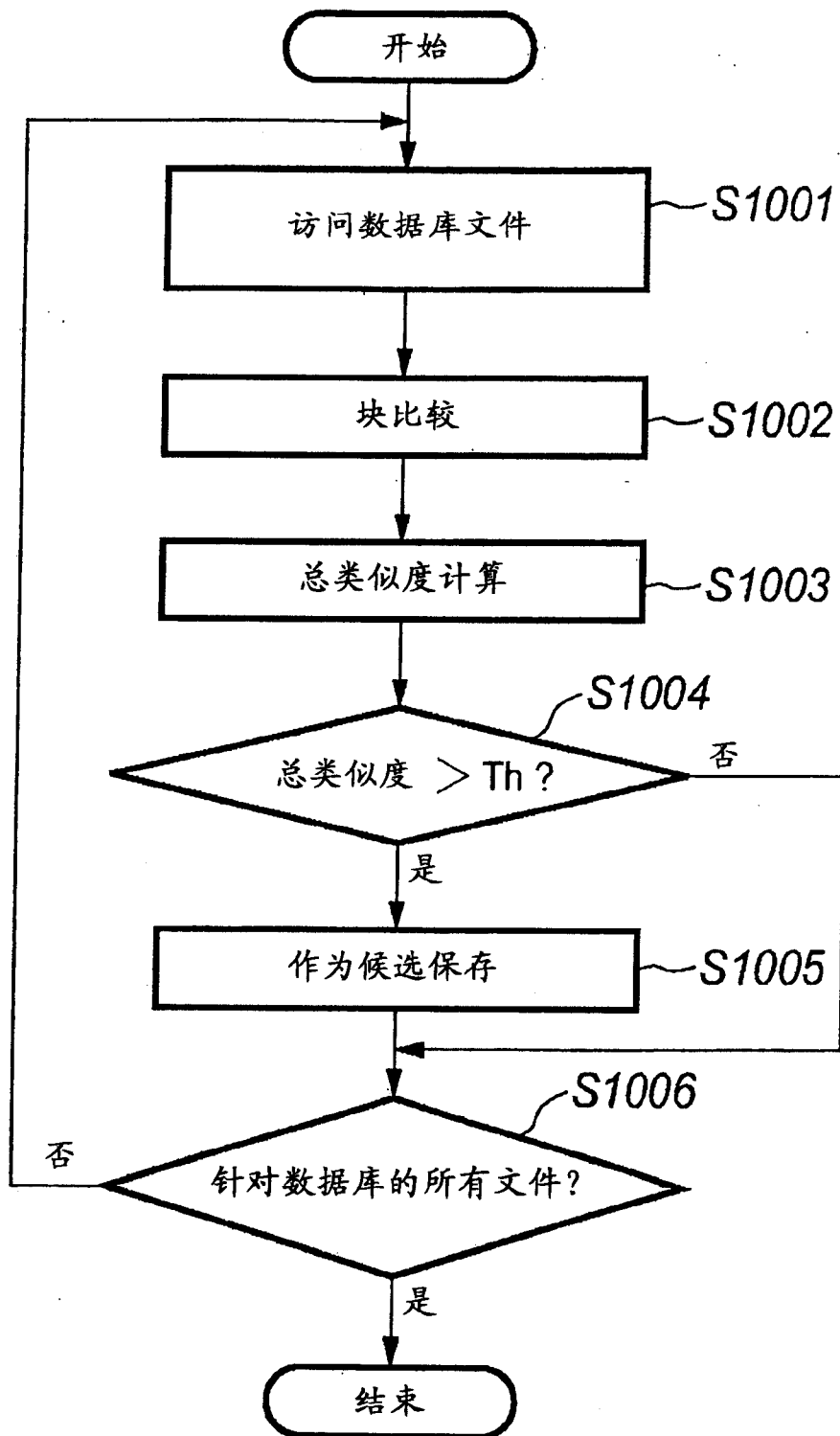


图 11

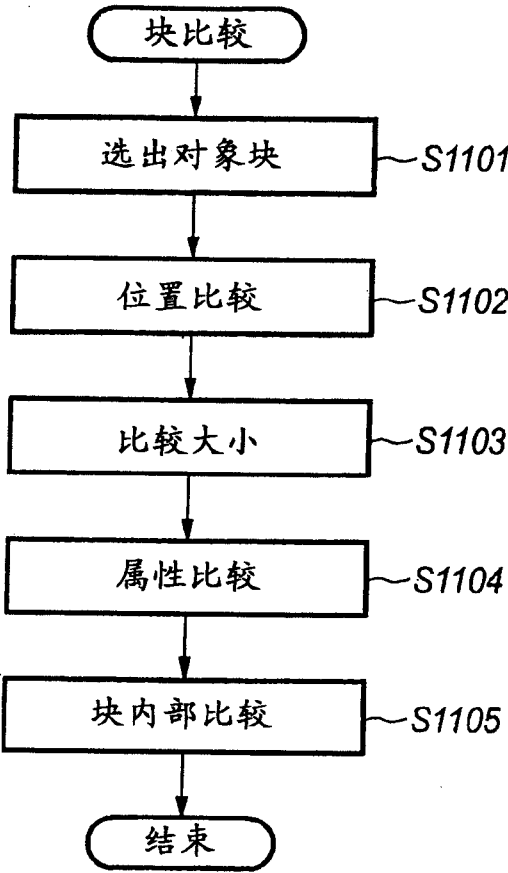


图12

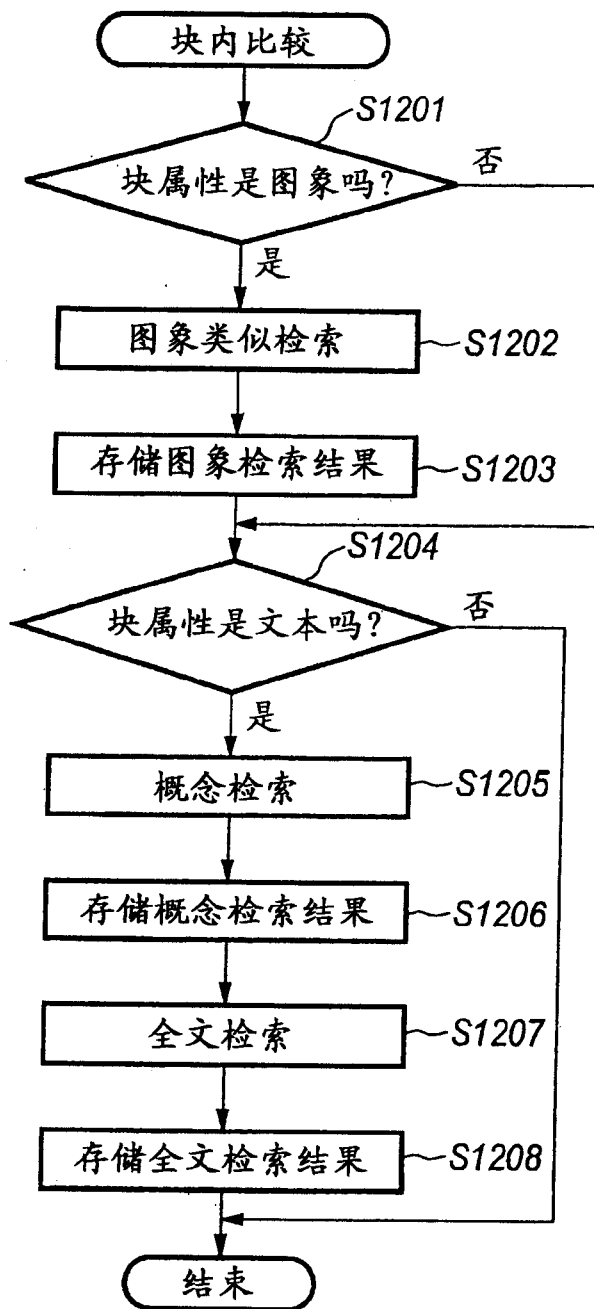


图13

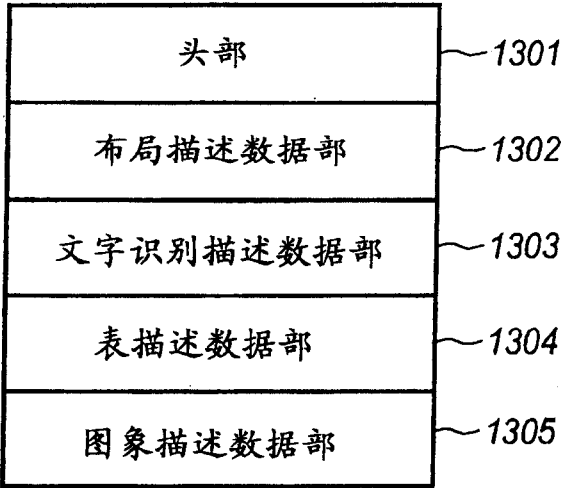


图 14

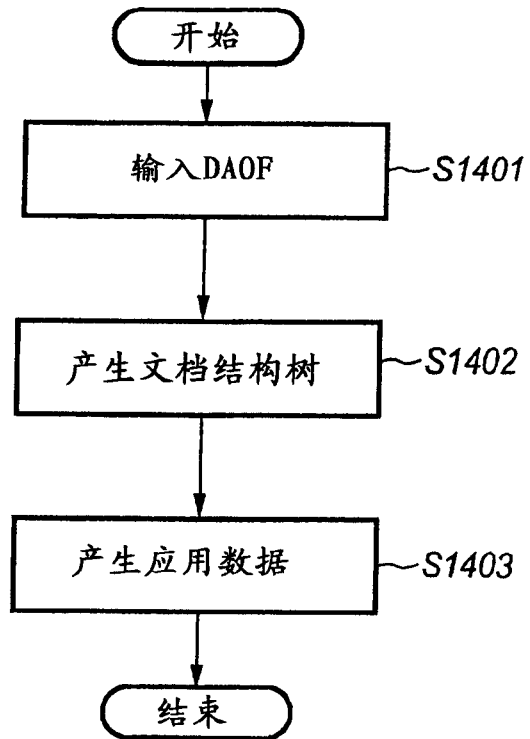


图15

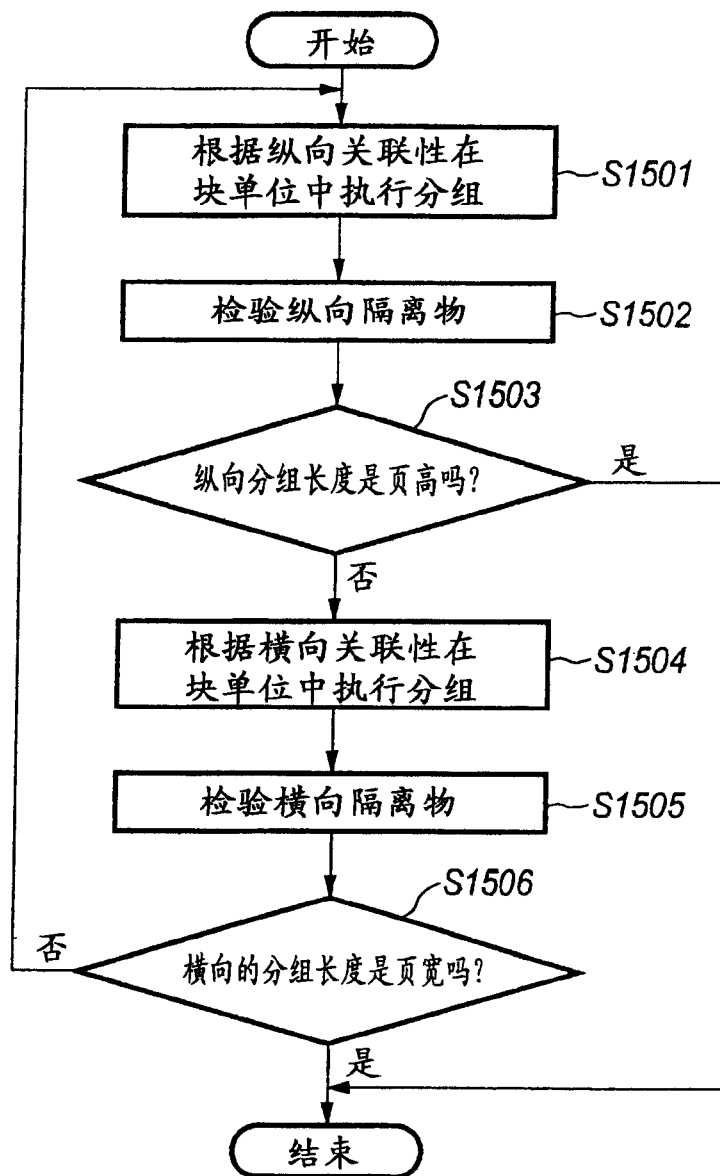


图16A

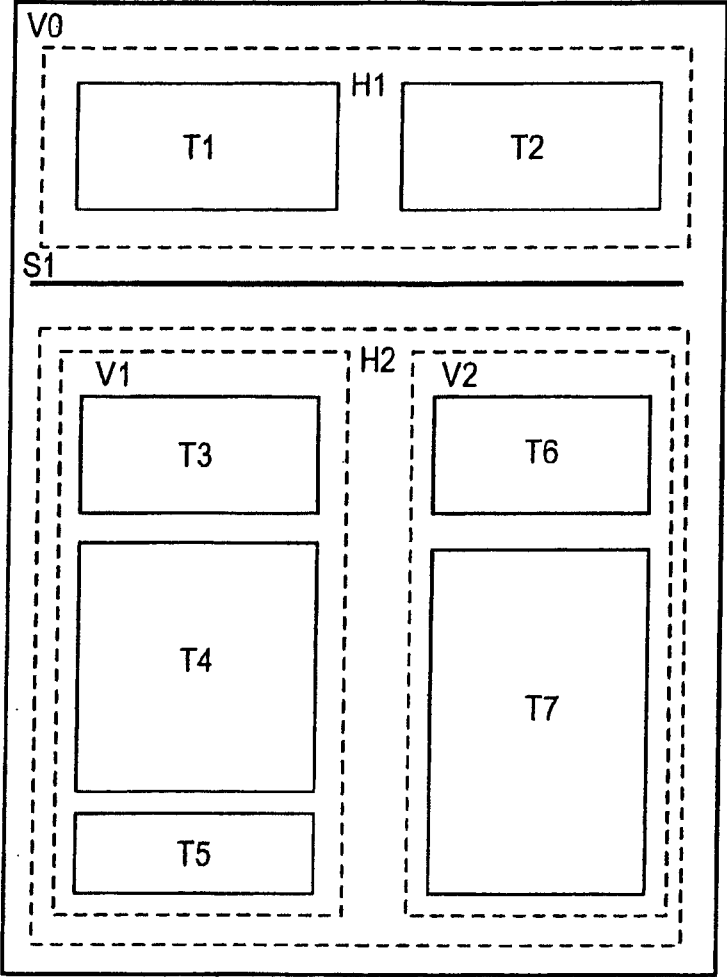


图16B

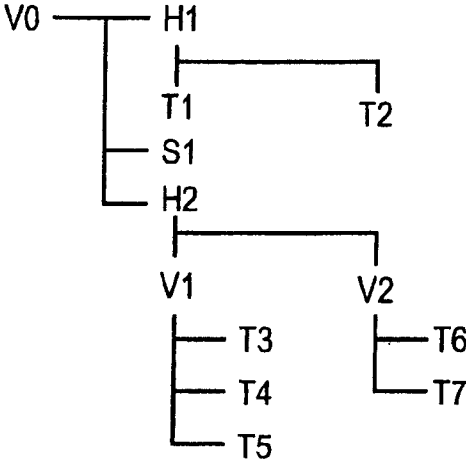


图 17

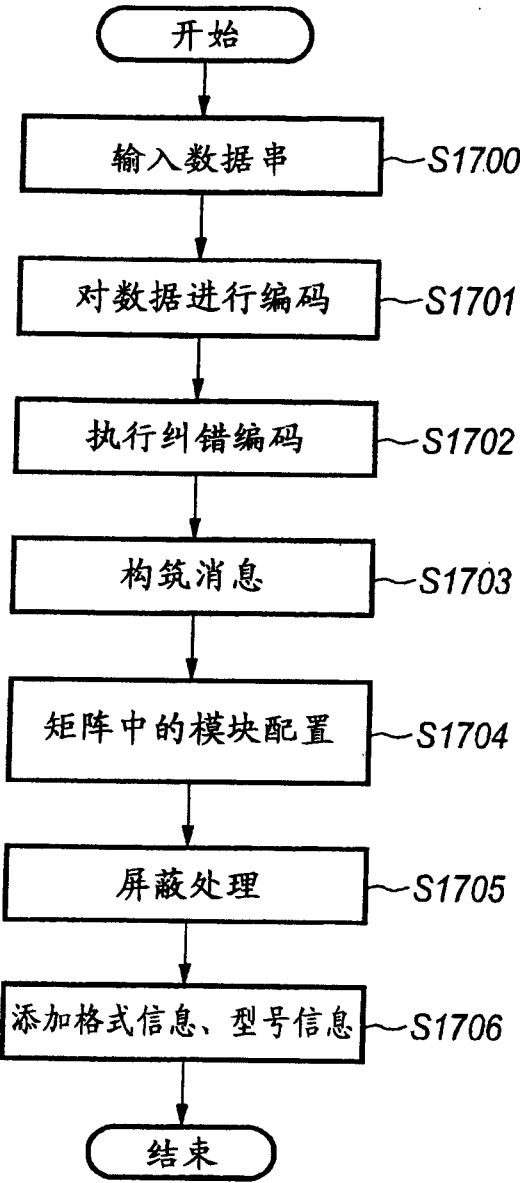


图18

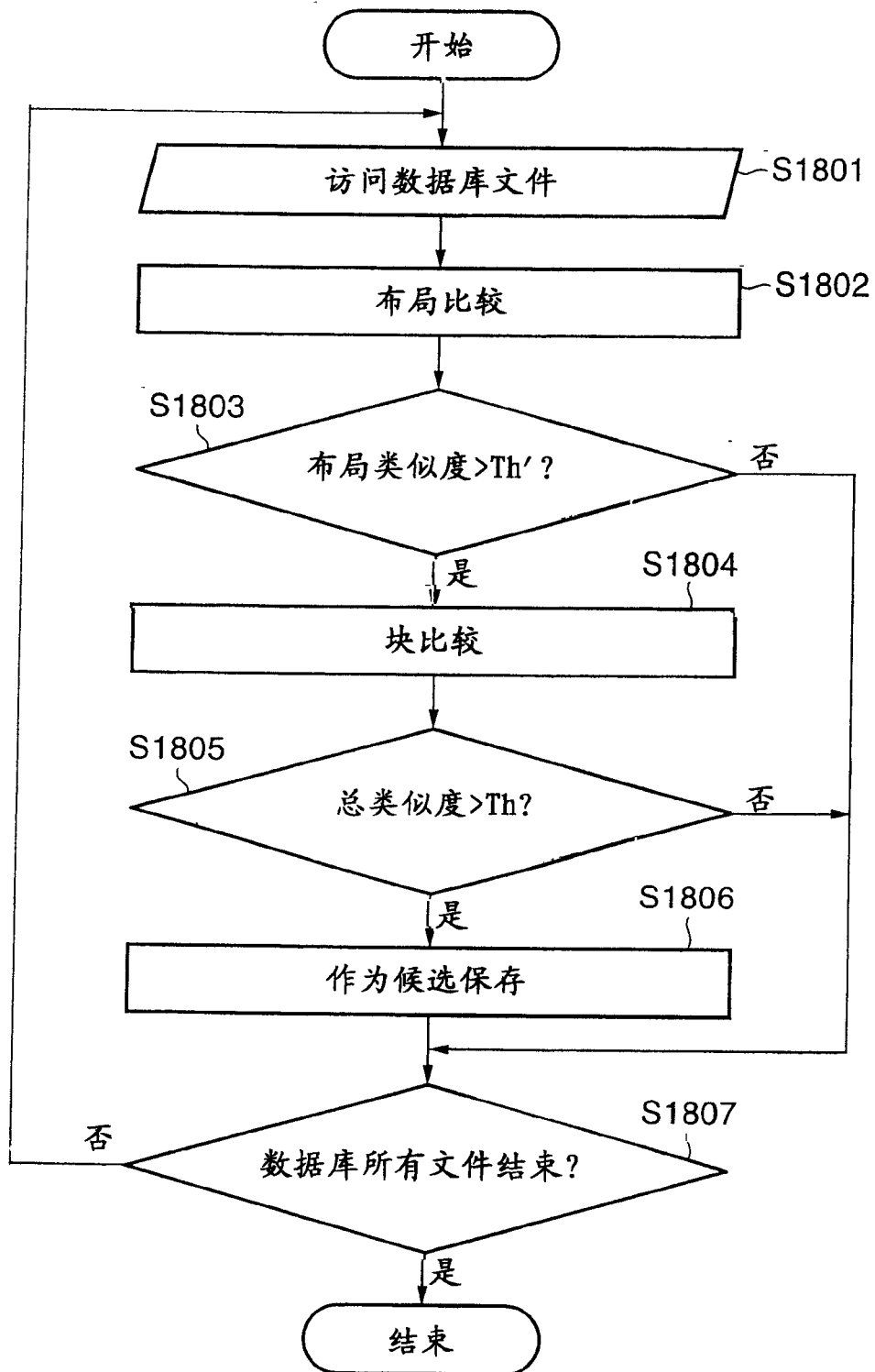


图 19

