(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102196130 B
(45)授权公告日 2014.02.19

(21)申请号 201110059803.0
(22)申请日 2011.03.11
(30)优先权数据
2010-059160 2010.03.16 JP
(73)专利权人 佳能株式会社
地址 日本东京都大田区下丸子 3-30-2
(72)发明人 相马英智 金津知俊 小坂亮 三沢玲司
(74)专利代理机构 北京怡丰知识产权代理有限公司 11293
代理人 迟军
(51)Int.Cl.
H04N 1/00(2006.01)
G06F 17/30(2006.01)

(56)对比文件
CN 101039369 A,2007.09.19, 全文.
CN 101267492 A,2008.09.17, 全文.
CN 101577778 A,2009.11.11, 全文.
US 2005/0111053 A1,2005.05.26, 全文.

审查员 冯薇

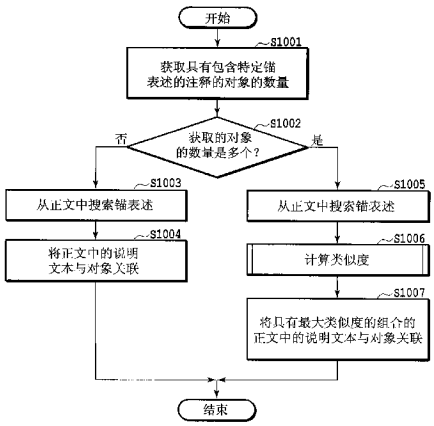
权利要求书2页 说明书18页 附图22页

(54)发明名称

图像处理装置及图像处理方法

(57)摘要

本发明提供一种图像处理装置及图像处理方法。即使当多个对象的注释使用相同锚表述时，本发明也能够将适当的正文中的说明文本作为元数据与对象关联。



CN 102196130 B

1. 一种图像处理装置，所述图像处理装置包括：

区域划分单元，其被配置为将页的图像划分为多个区域；

属性信息添加单元，其被配置为向划分的所述多个区域添加与区域对应的属性；

字符识别单元，其被配置为对由所述属性信息添加单元分别添加了注释属性和正文属性的注释区域和正文区域进行字符识别处理；以及

元数据处理单元，其被配置为将元数据与附有所述注释区域的对象区域进行关联；

其中，所述元数据处理单元包括：

第一提取单元，其被配置为从对所述注释区域的所述字符识别处理的结果中，提取由预定字符串构成的锚表述以及由所述锚表述以外的字符串构成的注释表述；

确定单元，其被配置为确定是否存在附有包含相同锚表述的注释区域的多个对象区域；

第二提取单元，其被配置为从对所述正文区域的所述字符识别处理的结果中，提取包含所述锚表述的说明文本；

第一关联单元，其被配置为在所述确定单元确定存在附有包含所述相同锚表述的注释区域的一个对象区域的情况下，将所述对象区域与由所述第二提取单元提取的所述说明文本获得的元数据进行关联；

类似度计算单元，其被配置为在所述确定单元确定存在附有包含所述相同锚表述的注释区域的多个对象区域的情况下，分别计算包含所述相同锚表述的各个注释区域的注释表述、与由所述第二提取单元提取的包含所述相同锚表述的所述说明文本之间的类似度；以及

第二关联单元，其被配置为基于由所述类似度计算单元计算出的所述类似度，来确定针对所述多个对象区域中的各个对象区域的最佳说明文本，并将由所确定的最佳说明文本获得的元数据与所述各个对象区域进行关联。

2. 根据权利要求 1 所述的图像处理装置，其中，所述对象区域是由所述属性信息添加单元添加了照片、图或者表的任意属性的区域。

3. 根据权利要求 1 所述的图像处理装置，其中，所述第二关联单元将具有所述类似度计算单元计算出的最大类似度的说明文本确定作为针对所述对象区域的最佳说明文本，并将由所述最佳说明文本获得的元数据与所述对象区域进行关联。

4. 根据权利要求 1 所述的图像处理装置，其中，

所述区域划分单元将多页的各个图像划分为多个区域，

所述元数据处理单元还包括第二确定单元，所述第二确定单元被配置为确定由所述第二提取单元提取的所述说明文本和所述对象区域两者是否存在于同一页的图像内，以及

在所述第二确定单元确定由所述第二提取单元提取的所述说明文本和所述对象区域两者存在于同一页的图像内的情况下，所述元数据处理单元跳过由所述类似度计算单元进行的所述类似度的计算，将由从该同一页的图像中提取的说明文本获得的元数据与存在于该同一页的图像内的所述对象区域进行关联。

5. 根据权利要求 1 所述的图像处理装置，其中，所述第二关联单元在所述最佳说明文本的类似度达到预定阈值的情况下，将由所述最佳说明文本获得的元数据与所述对象区域进行关联。

6. 根据权利要求 1 所述的图像处理装置，其中，所述元数据处理单元在所述确定单元确定存在附有包含所述相同锚表述的注释区域的多个对象区域的情况下，提供示出存在所述多个对象区域的警告显示。

7. 根据权利要求 1 所述的图像处理装置，其中，所述图像处理装置还包括第三关联单元，所述第三关联单元被配置为将由所述第一提取单元提取的所述注释表述作为元数据与附有包含所述注释表述的注释区域的对象区域进行关联。

8. 根据权利要求 1 所述的图像处理装置，其中，所述图像处理装置还包括生成单元，所述生成单元被配置为使用关于由所述区域划分单元对页的图像划分的所述多个区域的信息以及与所述对象区域关联的元数据，来生成具有预定格式的电子文档。

9. 一种图像处理方法，所述图像处理方法包括 ：

区域划分步骤，用于将多页的图像划分为多个区域 ；

属性信息添加步骤，用于向划分的所述多个区域中的各个添加与区域对应的属性 ；

字符识别步骤，用于对由所述属性信息添加步骤分别添加了注释属性和正文属性的注释区域和正文区域进行字符识别处理 ；以及

元数据处理步骤，用于将元数据与附有所述注释区域的对象区域进行关联 ；

其中，所述元数据处理步骤包括 ：

第一提取步骤，用于从对所述注释区域的所述字符识别处理的结果中，提取由预定字符串构成的锚表述以及由所述锚表述以外的字符串构成的注释表述 ；

确定步骤，用于确定是否存在附有包含相同锚表述的注释区域的多个对象区域 ；

第二提取步骤，用于从对所述正文区域的所述字符识别处理的结果中，提取包含所述锚表述的说明文本 ；

第一关联步骤，用于在所述确定步骤确定存在附有包含所述相同锚表述的注释区域的一个对象区域的情况下，将由所述第二提取步骤提取的所述说明文本获得的元数据与所述对象区域进行关联 ；

类似度计算步骤，用于在所述确定步骤确定存在附有包含所述相同锚表述的注释区域的多个对象区域的情况下，分别计算包含所述相同锚表述的各个注释区域的注释表述、与由所述第二提取步骤提取的包含所述相同锚表述的所述说明文本之间的类似度 ；以及

第二关联步骤，用于基于由所述类似度计算步骤计算出的所述类似度，来确定针对所述多个对象区域中的各个对象区域的最佳说明文本，并将由所确定的最佳说明文本获得的元数据与所述各个对象区域进行关联。

图像处理装置及图像处理方法

技术领域

[0001] 本发明涉及一种生成用于搜索文档中的对象的电子文档数据的图像处理装置及图像处理方法。

背景技术

[0002] 传统地，考虑了提供一种搜索包含在文档中的诸如照片、图（线图）或者表等的对象的方法。（这里所使用的术语“对象”是指包括字符以外的诸如照片、图（线图）或者表等的对象。）

[0003] 例如，存在如下一种方法：在从文档中提取的对象的附近，添加描述对象的字符串（注释 (caption)）并将其作为元数据关联以使得能够搜索对象。

[0004] 当一般文档中的注释包含诸如图编号（例如“照片 1”、“第一图”或“表 1”）等用于识别对象的表述（以下称为“锚表述 (anchor expression)”）时，在正文中也使用锚表述来说明对象的更为详细的描述。如上所述的锚表述也被用作用于识别文档中的对象的手段。根据日本特开平 11-025113(1999) 号公报中公开的发明，提取包含锚表述的正文中的说明部分（以下称为“正文中的说明文本 (explanatory text)”）并将其作为对象的元数据关联。当与图的对象邻近的注释包含锚表述“图 1”并且正文包含例如“图 1 是 AAA”的说明时，将锚表述“图 1”作为图的对象的标识信息进行关联。同时，还将正文中的说明文本“图 1 是 AAA”作为元数据进行关联，由此提供利用元数据对图的对象的搜索。

[0005] 近年来，例如一些字处理器具有编辑功能（例如自动生成锚表述的功能以及将文档中存在的对象与正文中的说明文本关联的功能）。可以将通过这些功能给出的信息（元数据）存储在电子文档中，由此实现对文档的有效编辑。

[0006] 近年来的扫描器具有诸如自动文档给送器的功能，因此能够容易地读取多页纸。因此，这种扫描器还能够同时读取多种文档。另一方面，当这种扫描器必须读取混合的不同文档时，存在如下可能性，即可能产生具有包含相同锚表述的注释的多个对象。例如，可能存在如下情况：同时读取的多个文档中的一个文档具有注释为“表 1 是 YYY”的表对象，而其中的另一个文档具有注释为“表 1 示出 ZZZ”的表对象。如果在这种情况下简单地进行上述关联处理，则相同的锚表述“表 1”与两个表对象关联，从而无法提供与锚表述“表 1”适当地对应的正文中的说明文本。

[0007] 由于上述情形，期望这样一种方法：即使当必须读取多种文档并且多个注释使用同一锚表述时，也能够将注释或正文中的说明文本作为元数据与对象适当地关联。

发明内容

[0008] 根据本发明的图像处理装置包括：区域划分单元，其被配置为将多页的图像分别划分为多个区域；属性信息添加单元，其被配置为向划分的所述多个区域中的各个添加与区域对应的属性；字符识别单元，其被配置为对由所述属性信息添加单元分别添加了注释属性和正文属性的注释区域和正文区域进行字符识别处理；以及元数据处理单元，其被配

置为将元数据与附有所述注释区域的对象区域进行关联；其中，所述元数据处理单元包括：第一提取单元，其被配置为从对所述注释区域的所述字符识别处理的结果中，提取由预定字符串构成的锚表述以及由所述锚表述以外的字符串构成的注释表述；确定单元，其被配置为确定是否存在附有包含相同锚表述的注释区域的多个对象区域；第二提取单元，其被配置为从对所述正文区域的所述字符识别处理的结果中，提取包含所述锚表述的说明文本；第一关联单元，其被配置为在所述确定单元确定存在附有包含所述相同锚表述的注释区域的一个对象区域的情况下，将所述对象区域与由所述第二提取单元提取的所述说明文本获得的元数据进行关联；类似度计算单元，其被配置为在所述确定单元确定存在附有包含所述相同锚表述的注释区域的多个对象区域的情况下，分别计算包含所述相同锚表述的各个注释区域的注释表述、与由所述第二提取单元提取的包含所述相同锚表述的所述说明文本之间的类似度；以及第二关联单元，其被配置为基于由所述类似度计算单元计算出的所述类似度，来确定针对所述多个对象区域中的各个对象区域的最佳说明文本，并将由所确定的最佳说明文本获得的元数据与所述各个对象区域进行关联。

[0009] 根据本发明，即使存在包含相同锚表述的多个对象，也能够将对象与适当的元数据关联。因此，甚至在包含多个文档的混合的图像数据中，也能够准备向各个对象添加了适当的元数据的电子文档。

[0010] 从以下参照附图对示例性实施例的描述中，本发明的其它特征将变得清楚。

附图说明

[0011] 图 1 是示出根据本发明的图像处理系统的配置的框图；

[0012] 图 2 示出 MFP 100 的配置；

[0013] 图 3 是示出数据处理单元 218 的内部配置的框图；

[0014] 图 4 是示出元数据处理单元 304 的内部配置的框图；

[0015] 图 5A 至图 5C 示出数据处理单元 218 的处理详情，其中图 5A 示出如何将图像数据划分为区域，图 5B 是示出区域划分单元 301、属性信息添加单元 302 以及字符识别单元 303 的处理结果的示例的表，图 5C 示出格式转换单元 305 使用的对应表的示例；

[0016] 图 6A 和图 6B 示出在本发明中生成的电子文档，其中图 6A 示出利用 SVG 格式描述的电子文档的示例，图 6B 以表的形式示出电子文档中的元数据；

[0017] 图 7 是示出根据本发明的图像处理装置的处理的概要的流程图；

[0018] 图 8A 和图 8B 是示出元数据提取和添加处理的流程的流程图；

[0019] 图 9A 和图 9B 示出根据本发明的图像处理装置的处理，其中图 9A 示出输入到数据处理单元 218 的图像数据以及对其的区域划分的示例，图 9B 是示出在图像数据的情况下区域划分单元 301、属性信息添加单元 302 以及字符识别单元 303 的处理结果的示例的表；

[0020] 图 10 是示出根据实施例 1 的、用于将正文中的说明文本与对象区域关联的处理的流程的流程图；

[0021] 图 11 是示出根据实施例 1 的类似度计算处理的流程的流程图；

[0022] 图 12A 和图 12B 示出通过类似度计算处理获得的中间结果，其中图 12A 示出通过到步骤 1103 的处理获得的结果的示例，图 12B 示出在步骤 1105 和步骤 1106 中的处理的示例；

[0023] 图 13 以表的形式示出通过对图 9A 的图像数据 901 至 907 进行根据实施 1 的处理获得的元数据信息；

[0024] 图 14 是示出图 14A 和图 14B 之间的关系的图；

[0025] 图 14A 和图 14B 是示出根据实施例 2 的、用于将正文中的说明文本与对象区域进行关联的处理的流程的流程图；

[0026] 图 15 是示出根据实施例 2 的类似度计算处理的流程的流程图；

[0027] 图 16 是示出根据实施例 3 的、用于将正文中的说明文本与对象区域进行关联的处理的流程的流程图；

[0028] 图 17 示出在操作单元 203 上显示的用户界面 (UI) 画面的示例；

[0029] 图 18 示出在操作单元 203 上显示的警告画面的示例；以及

[0030] 图 19 是示出根据实施例 4 的、用于将正文中的说明文本与对象关联的处理的流程的流程图。

具体实施方式

[0031] [实施例 1]

[0032] 下文将参照附图描述用于实现本发明的实施例。

[0033] 图 1 是示出根据本实施例的图像处理系统的配置的框图。

[0034] 在图 1 中，在办公室 A 中构建的 LAN 102 连接到作为用于实现多种功能（例如复印功能、打印功能、发送功能）的图像处理装置的多功能外围设备 (MFP) 100。LAN 102 还经由代理服务器 103 连接到外部网络 104。客户端 PC 101 经由 LAN 102 接收从 MFP 100 发送的数据并使用 MFP 100 拥有的功能。例如，客户端 PC 101 还可以向 MFP 100 发送打印数据，由此使得通过 MFP 100 来打印基于打印数据的打印物。图 1 的配置是示例。因此，也可以使用其它配置，其中具有与办公室 A 相同的构成部件的多个办公室经由网络 104 连接。网络 104 典型地是由例如互联网、LAN、WAN、电话线路、专用数字线路、ATM、帧中继线路、通信卫星线路、有线电视线路或者数据广播无线线路实现的通信网络。网络 104 可以是任意网络，只要能够通过其发送和接收数据即可。客户端 PC101 和代理服务器 103 的各个终端具有在通用计算机中设置的标准构成部件（包括例如 CPU、RAM、ROM、硬盘、外部存储装置、网络接口、显示器、键盘及鼠标）。

[0035] 图 2 示出 MFP 100 的配置。

[0036] 首先，MFP 100 的配置主要被分为作为图像输入设备的扫描器单元 201、作为图像输出设备的打印机单元 202、由 CPU 205 构成的控制单元 204 以及作为用户接口的操作单元 203。

[0037] 控制单元 204 连接到扫描器单元 201、打印机单元 202 和操作单元 203。控制单元 204 还是连接到 LAN 219 或者作为一般电话线网络的公共线路 (WAN) 220、由此提供图像信息及设备信息的输入和输出的控制器。

[0038] CPU 205 控制包括在控制单元 204 中的各个单元。

[0039] RAM 206 是用于 CPU 205 的操作的系统工作存储器并且还是用于临时存储图像数据的图像存储器。

[0040] ROM 210 是存储诸如系统引导程序等的程序的引导 ROM。

[0041] 存储单元 211 是存储系统控制软件及图像数据的硬盘驱动器。

[0042] 操作单元 I/F 207 是与操作单元 (UI)203 的接口单元，其向操作单元 203 输出要在操作单元 203 上显示的图像数据。操作单元 I/F 207 还具有向 CPU 205 发送由该图像处理装置的用户通过操作单元 203 输入的信息的功能。

[0043] 网络 I/F 208 将该图像处理装置连接到 LAN 219 以提供包 (packet) 类型信息的输入和输出。

[0044] 调制解调器 209 将该图像处理装置连接到 WAN 220 来提供数据解调和调制，由此提供信息的输入和输出。如上所述的设备被布置在系统总线 221 上。

[0045] 图像总线 I/F 212 是将系统总线 221 连接到用于以高速传送图像数据的图像总线 222 以转换数据结构的总线桥。

[0046] 图像总线 222 由例如 PCI 总线以及 IEEE 1394 构成。图像总线 222 上具有如下设备。

[0047] 光栅图像处理器 (RIP)213 实现分析页面描述语言 (PDL) 代码以将代码展开为具有指定分辨率的位图图像的所谓绘制处理。该展开以像素为单位或者以区域为单位来添加属性信息。这被称为图像区域确定处理。通过图像区域确定处理，针对各个像素或者各个区域添加示出对象类型（例如字符（文本）、线、图形或者图像）的属性信息。例如，依据 PDL 代码中的 PDL 描述的对象的类型，RIP 213 输出图像区域信号。然后，将与信号值所示的属性对应的属性信息和对应于对象的像素或区域关联地存储。因此，图像数据附有与其关联的属性信息。

[0048] 设备 I/F 214 经由信号线 223 将作为图像输入设备的扫描器单元 201 连接到控制单元 204。设备 I/F 214 还经由信号线 224 将作为图像输出设备的打印机单元 202 连接到控制单元 204。设备 I/F 214 提供图像数据的同步 / 异步转换。

[0049] 扫描器图像处理单元 215 对输入的图像数据进行校正、处理和编辑。

[0050] 打印机图像处理单元 216 对要输出到打印机单元 202 的打印输出图像数据进行例如依据打印机单元 202 的校正及分辨率转换。

[0051] 图像旋转单元 217 对输入的图像数据进行旋转以使数据直立并输出该数据。

[0052] 参照图 3 详细描述数据处理单元 218。

[0053] < 数据处理单元 >

[0054] 如图 3 所示，数据处理单元 218 由区域划分单元（区域提取单元）301、属性信息添加单元 302、字符识别单元 303、元数据处理单元 304 以及格式转换单元 305 构成。数据处理单元 218 利用各个处理单元 301 至 305 对输入的图像数据 300 进行预定的处理。最后，生成并输出电子文档 310。

[0055] 区域划分单元 301 接收由扫描器单元 201 读取的图像数据或者接收从诸如客户端 PC 的外部装置接收并存储在存储单元中的图像数据。为了以页为单位从输入的图像数据中提取基于诸如字符、照片、图或者表的各个属性划分的区域，使图像数据中的像素经历诸如提取或者分组的处理。可以通过已知的区域划分方法（区域提取方法）进行该处理。以下描述这种方法的示例。首先，将输入图像进行二值化以生成二值图像。然后，使二值图像具有较低分辨率，由此准备筛选图像（缩小图像）。例如，为了准备 1/(M×N) 稀疏图像 (thinning image)，将二值图像划分为分别对应 M×N 个像素。如果 M×N 个像素包含黑色

像素，则将缩小图像中的相应像素设置为黑色像素。如果 M×N 个像素不包含黑色像素，则将缩小图像中的相应像素设置为白色像素。以这种方式，准备稀疏图像。接下来，提取稀疏图像中黑色像素连接的部分（连接黑色像素），由此准备由连接黑色像素外接的矩形。当布置大小与字符图像的大小类似的矩形（一个字符的矩形）时或者当在短边的附近布置高度或者宽度与字符图像大小类似的相似矩形（连接了多个字符的连接黑色像素的矩形）时，这些矩形可以组成构成字符行的字符图像的可能性高。在这种情况下，将这些矩形相互连接，由此获得表示一个字符行的矩形。在表示一个字符行的、短边长度大体相同并且在列方向上以大体相等的间距布置的矩形的集合的情况下，该集合可以表示可能为正文。因此，将这些矩形连接，并提取得到的部分作为正文区域。通过大小大于字符图像的连接黑色像素来提取照片区域、图区域以及表区域。结果，例如提取了由图 5A 的标号 501 至 506 所示的各个区域。如稍后所述，基于例如大小、长宽比或者黑色像素密度或者连接黑色像素中包括的白色像素的轮廓跟踪结果来确定各个区域的属性。

[0056] 属性信息添加单元 302 向区域划分单元 301 划分的各个区域添加属性信息。如图 5A 所示，基于假定区域划分单元 301 划分了图像数据 500 来进行以下描述。

[0057] 在页中的区域 506 中包括大于或者等于预定数量的字符和行并且具有例如段落的形式。因此，区域 506 被添加“正文”属性。为了简单地示出图，包含在图 5A 的区域 506 中的字符用黑色点表示。然而，区域 506 实际包含多个字符图像。

[0058] 对于其余区域 501 至 505，确定区域 501 至 505 是否是包含字符的区域。具体地说，在包含字符的区域的情况下，字符图像的矩形周期性地出现在区域中。因此，确定区域是否是包含大小与字符图像大小类似的矩形的区域。结果，区域 501、区域 504 和区域 505 被确定为包含字符的区域并被添加有“字符包含区域”属性。

[0059] 另一方面，对于上述区域以外的区域，首先确定区域的大小。当区域大小非常小时，其属性被确定为“噪声”。当区域大小不是很小而是等于或者大于固定大小时，该区域是某个对象的区域。因此，进一步对该区域进行以下确定以确定要向该区域添加的属性（即该区域对应于哪一个对象）。首先，对具有低像素密度的连接黑色像素中的白色像素进行轮廓跟踪。然后，确定白色像素轮廓的外接矩形是否按顺序布置。当白色像素轮廓的外接矩形按顺序布置时，则该区域被确定为具有“表”属性。当白色像素轮廓的外接矩形未按顺序布置时，该区域被确定为具有“图（线图）”属性。该区域以外的具有高像素密度的区域被确定为与图片或者照片相对应并被确定为具有“照片”属性。在上述部分中，基于“照片”、“图（线图）”、以及“表”三种对象属性来对区域进行分类。然而，本发明不限于此。还可以使用其它确定标准以使得能够基于任意类型的对象属性来对区域进行分类。

[0060] 如果在添加有诸如“照片”、“图”或者“表”的属性的对象区域的附近（或正上方或正下方）存在字符包含区域，则确定该区域是用于描述对象区域的字符的区域并由此添加“注释”属性。将添加有“注释”属性的区域与附有该注释的诸如“照片”、“图”或者“表”的区域关联地存储以使得可以识别后者区域。

[0061] 当区域大于正文部分的字符图像并且位于与正文部分的列设置的位置不同的位置时，该区域被添加“标题”属性。当区域大于正文部分的字符图像并且存在于正文部分的列设置的上部时，该区域被添加“副标题”属性。当区域是用于大小比正文部分的字符图像的大小小的字符图像并且存在于原稿的下端或上端时，该区域被添加“页”（或者“页眉”或

者“页脚”)。在区域被确定为字符包含区域而不被确定为“正文”、“标题”、“副标题”、“注释”或者“页”的情况下,该区域被添加“字符”属性。

[0062] 通过如上所述的添加属性信息的处理,对图像数据 500 进行设置以使得分别对区域 501 添加标题属性,对区域 502 添加表属性,对区域 503 添加照片属性,对区域 504 添加字符属性,对区域 505 添加注释属性(其附加区域 503)以及对区域 506 添加正文属性。对区域 501、504 以及 505 的各个属性给出的下划线表示对其添加了“字符包含区域”。

[0063] 字符识别单元 303 对包含字符图像的区域(添加有“字符”属性、“正文”属性、“标题”属性、“副标题”属性以及“注释”属性的区域)进行已知的字符识别处理(OCR 处理)。还可以对“表”中的字符图像的区域添加“表内字符”属性并且随后可以对其进行字符识别处理。然后,将通过字符识别处理获得的字符代码列存储为字符识别信息并将其与目标区域关联。

[0064] 如上所述,存储单元 211 在其中存储由区域划分单元 301、属性信息添加单元 302 以及字符识别单元 303 中的各个单元提取的例如区域的位置及大小、区域属性信息、页信息、字符识别处理的结果的信息(字符代码列)。图 5B 以表的形式示出图像数据 500 的上述处理的结果。因此,该结果以表形式存储在存储单元 211 中。区域 504 是存在于照片 503 中的字符图像的区域。因此,区域 504 被添加“在照片 503 中”属性。尽管表中关于区域的位置和大小的格(例如坐标 X/Y、宽度 W、高度 H)包括诸如 X1 的标记,但是在实际情况下在这些格中包括数值。

[0065] 元数据处理单元 304 将由属性信息添加单元 302 确定为附有注释的区域的对象区域(例如“照片”、“图”或者“表”)与用作用于搜索对象的信息的元数据进行关联。然后,属性信息添加单元 302 将得到的数据存储在存储单元 211 中。具体地说,属性信息添加单元 302 将作为特定对象区域的注释来描述的字符串、与使用跟该字符串中包括的诸如图号(锚描述)的词相同的词的正文中的说明部分(正文中的说明文本)关联,作为用于搜索对象的元数据。

[0066] 通过对各个对象给出的标识符(以下称为“注释标识符”)进行元数据的关联。该注释标识符可以用于将注释或者正文中的说明文本分别与附有注释的对象区域适当地关联。通过向各个对象添加不同的注释标识符,即使当存在具有包含相同锚表述的注释的多个对象时,也能够将元数据与对象适当地关联。在本实施例中,注释标识符是用于唯一识别附有注释的对象区域的 ID(即从值“1”开始的序列号(正整数))。注释标识符还可以是诸如表示注释或者正文的字符识别信息的存储位置的地址或者指针的位置信息或者诸如 XMLPath 或者 URL 的参照信息。

[0067] 格式转换单元 305 使用由上述各个处理单元获得的各种信息(例如页信息、区域的位置或者大小、属性、字符识别信息、元数据),将图像数据转换为具有预定格式的电子文档。预定格式可以是例如 PDF、SVG、XPS 或者 OfficeOpenXML。通过格式转换生成的电子文档包括使用例如图形的页显示信息(例如待显示图像)以及使用诸如字符的语义描述的内容信息(例如元数据)。

[0068] 格式转换单元 305 主要进行以下两种处理。一种处理是对图像区域进行滤波器处理(例如直方图均衡化、平滑、边缘增强、色量化、二值化),以使得图像数据(例如与添加有“图(线图)”属性的区域相对应的部分的图像)能够以具有预定格式的电子文档的形式进

行存储。通过将图像数据转换为矢量路径描述图形数据（矢量数据）或位图描述图形数据（例如 JPEG 数据），能够将图像数据以具有预定格式的电子文档的形式进行存储。可以通过已知的矢量化技术来进行到矢量数据的转换。该转换还生成当针对对象搜索将搜索结果进行识别或强调时显示的诸如框的图形描述（矢量路径描述）。另一种处理是向生成的矢量数据或位图数据中添加存储在存储单元 211 中的区域信息（位置、大小、属性）、用于区域中的字符识别的信息以及元数据，由此准备具有预定格式的电子文档。

[0069] 由格式转换单元 305 应当对各个区域进行的转换处理方法常常依据区域的属性。例如，矢量转换处理适于诸如字符或者线图等由白黑颜色或者少数颜色构成的图形，但不适于诸如照片等具有灰阶的图像区域。为了提供与各个区域的属性对应的适当转换，还可以预先提供如图 5C 所示的对应表。

[0070] 例如，设置图 5C 所示的对应表，以使得对具有“字符”、“图（线图）”以及“表”属性的区域进行矢量转换处理并对具有“照片”属性的区域进行图像剪切处理，作为各自的转换处理。另外，设置如图 5C 所示的对应表，以使得针对各个属性设置是否从图像数据中删除区域的像素信息。例如，当将具有“字符”属性的区域的数据转换为矢量路径描述数据时，设置区域以进行删除处理。在这种情况下，进行用周围颜色填充与由转换的矢量路径覆盖的部分相对应的像素的这种处理。类似地，当将具有“照片”属性的区域剪切作为矩形图像部分时，用例如周围颜色填充与剪切区域相对应的区域。如上所述的删除处理的目的是使用各个区域已经进行了填充处理的图像数据作为构成“背景”图像数据的部分。在背景图像数据（背景图像）中包含由区域划分处理提取的区域以外的其余部分（例如与图像数据中的基色 (foundation) 相对应的像素）。通过将由未示出的矢量转换处理单元或者图像剪切处理单元获得的图形数据（前景图像）叠加在背景图像上，由此显示结果图像来获得电子文档的数据描述。这因此能够防止背景像素（背景颜色）的信息的缺失并能够构建非冗余的图形数据。

[0071] 作为另选方案，还可以预先准备多个对应表，以使得能够依据电子文档的应用来选择适当的对应表。例如，当使用图 5C 的对应表时，主要将对象转换为具有矢量路径描述，因此提供用于放大或者缩小显示的高图像质量。因此，提供其它对应表，例如根据该对应表针对字符图像的各个字符颜色生成不同的二值图像并对所述二值图像进行无损压缩以及将上述图像以外的图像作为背景图像进行例如 JPEG 压缩。前者适于再利用例如图形编辑器的应用。后者适于期望准备在实现高压缩比率的同时能够容易地读取字符图像的这样一种电子文档的情况。通过依据应用使用不同的对应表，能够准备适于用户应用的适当的电子文档。

[0072] 图 6A 示出由数据处理单元 218 生成的电子文档的示例。图 6A 示出基于图 5A 的图像数据 500 生成的、基于可缩放矢量图形 (SVG) 格式描述的电子文档 600。为了简便起见，图 6B 以表的形式示出添加到电子文档 600 的元数据的信息。

[0073] 在这种情况下，图 6A 的标号 601 至 606 分别表示与图像数据 500 中的区域 501 至 506 相对应的图形描述。标号 601 以及 604 至 606 表示使用字符代码的字符图描述。标号 602 表示用于进行了矢量转换的表的框的矢量路径描述。标号 603 表示用于粘附照片图像的描述。描述 603 包括注释标识符 (caption_id)608（“1”）。

[0074] 标号 607 表示用于元数据的描述。描述 607 描述了照片对象中的字符“新产品”、

注释“图 1AAA”、注释中的锚表述“图 1”以及基于所述锚表述和正文的字符串提取的正文中的说明文本“图 1 是关于 AAA 的 XX”。此外，还描述了与注释标识符 608 相同的标识符 609。代替包含锚表述“图 1”的正文中的说明文本（句子）“图 1 是关于 AAA 的 XX”，还可以提取词“AAA”并可以将其作为元数据添加。作为另选方案，还可以使用正文中的说明文本以及词二者作为元数据。其同样适用于注释的字符串。因此，还可以将从注释中提取的词作为元数据添加。

[0075] ＜元数据处理单元＞

[0076] 图 4 是示出元数据处理单元 304 的内部配置的框图。

[0077] 标号 401 表示元数据添加目标选择单元，其进行从输入的图像数据中选择要进行元数据的提取／添加的区域的处理。在本实施例中，选择具有注释区域的图像数据作为要进行元数据（例如注释、锚表述或者正文中的说明文本）的提取／添加的目标。

[0078] 标号 402 表示锚表述提取单元，其从由元数据添加目标选择单元 401 选择的注释区域中提取锚表述。具体地说，锚表述提取单元 402 分析与选择的注释区域关联的字符识别信息，以在字符识别信息中搜索诸如“图 1”的任意锚表述的存在。当找到锚表述时，提取相应部分作为锚表述并提取上述部分以外的其它部分作为注释表述。该提取使用例如字符代码特性或者词典以排除无用的字符串（例如无意义的符号串）。这因此防止了在字符识别中将出现在文档的文本部分的边界处的划分线或图像错误地识别为字符。此外，能够通过使用诸如图号的多语言字符串模式或者针对字符串模式的错误字符识别模式来提取锚表述，由此实现了锚表述提取的精确性并校正了锚表述的字符。还可以对注释表述进行使用自然语言处理的分析或者错误字符识别的校正。因此，还能够校正并排除具有锚表述的边界或者出现在开头或结尾的符号或字符修饰。

[0079] 标号 403 表示正文内搜索单元，其在与正文区域关联的字符识别信息中搜索包含由锚表述提取单元 402 获得的锚表述的正文中表述（句子），以提取其作为正文中的说明文本。为了实现高速搜索，还可以准备和使用搜索索引。索引的准备以及使用索引的高速搜索可以通过已知技术实现并且不是本发明的要点，因此将不再对其进行描述。作为另选方案，可以同时搜索多个锚表述，由此实现高速。还可以将在搜索中找到的正文中的说明文本与使用多语言字符串模式描述的锚表述相组合。

[0080] 标号 404 表示表述类似度计算单元，其基于锚表述将注释表述与正文中的说明文本进行比较以计算它们之间的类似度。具体地说，表述类似度计算单元 404 使用字符级 (character level) 或者自然语言分析来将注释表述与正文中的说明文本进行比较，由此进行基于词或者意义级的比较。然后，在关注两个表述中的锚表述的位置的同时计算类似度。类似度是当两个表述具有与锚表述较近的字符或者词时值增大的标准。稍后将详细描述类似度。

[0081] 标号 405 表示元数据收集／输出单元，其收集由上述各个单元提取的元数据以将元数据与要添加元数据的图像数据进行关联，由此将结果输出到格式转换单元 305。

[0082] 标号 406 是元数据处理控制单元，其基于存储在存储单元 211 中的区域信息（位置、大小、属性）411、区域中的字符识别信息 412 及元数据 413，将图像数据 300 分配给适当的处理单元 401 至 405。然后，元数据处理控制单元 406 对元数据处理单元 304 的整体进行控制，以使得元数据收集／输出单元 405 能够将从各个处理单元 401 至 404 输出的数据进

行整合。

[0083] 接下来，下文将参照图 7 的流程图来描述由根据本实施例的图像处理装置 (MFP 100) 进行的处理的概要。基于如下情况进行以下描述，所述情况为针对图 9A 所示的与由两种不同类型的文档构成的总共 7 页相对应的图像数据 (901 至 903 以及 904 至 907)，生成各个对象被添加了适当的元数据的电子文档。

[0084] 各个步骤中的处理（在此，这些处理被描述为由各个处理单元进行的处理）通过使 CPU 205 读取存储在存储单元 211 中的计算机程序来实现。然而，本发明不限于此。例如，各个处理单元还可以全部或部分通过硬件实现，以使得构成数据处理单元 218 的各个单元中的任意一个可以通过诸如电子电路的硬件来实现。

[0085] 首先，在步骤 S701 中，CPU 205 通过扫描器单元 201 读取文档或者读取从例如客户端 PC 101 发送并存储在存储单元 211 中的文档，由此获取多页的图像数据。将获取的图像数据 901 至 907 输入到数据处理单元 218 并将其从元数据处理控制单元 406 发送到区域划分单元 301。

[0086] 在步骤 S702 中，区域划分单元 301 将输入的图像数据 901 至 907 按照其属性以页为单位进行划分。图像数据 901 至 907 被划分为各个区域 908、910、911、912、913、915、917、918、919 以及 920。

[0087] 在步骤 S703 中，属性信息添加单元 302 向在步骤 S702 中划分的各个区域添加属性信息。例如，在第三页的图像数据 903 的情况下，区域 911 被添加“照片”属性，区域 912 被添加“注释”属性。注释 912 还被添加示出附加区域（对象）是区域 911 的信息。

[0088] 在步骤 S704 中，字符识别单元 303 对在步骤 S703 中添加了与字符相关的属性（例如正文、注释、标题、副标题）的区域进行字符识别处理，并将结果作为字符识别信息与区域关联，由此将结果存储在存储单元 211 中。各个区域的字符识别处理结果如在图 9B 所示的表中的字符识别信息的格中所描述的那样。

[0089] 在步骤 S705 中，数据处理单元 218 确定是否对所有页进行了步骤 S702 至 S704 中的各个处理。由于输入了与 7 页相对应的图像数据 901 至 907，因此当如图 9B 中的表所示的所有信息都被存储在存储单元 211 中时，能够确定对所有页的处理完成。当确定对所有页的处理完成时，处理进入步骤 S706。如果发现任何还未进行处理的页，则处理返回步骤 S702 并重复上述处理。

[0090] 在步骤 S706 中，元数据处理单元 304 进行提取和添加元数据的处理。稍后将描述该处理的详情。

[0091] 在步骤 S707 中，格式转换单元 305 使用图 5C 所示的对应表，基于存储在存储单元 211 中的各个信息将图像数据 901 至 907 转换为具有预先指定的预定格式的电子文档。

[0092] （元数据的提取和添加）

[0093] 接下来，下文将参照图 8A 和图 8B 的流程图来描述在上述步骤 S706 中的提取和添加元数据的处理的详情。该处理主要由以下两个处理构成。第一处理是用于提取锚表述及注释表述的第一提取（图 8A）。第二处理是用于提取正文中的说明文本的第二提取（图 8B）。首先，进行用于提取锚表述和注释表述的处理。然后，进行用于提取正文中的说明文本的处理。

[0094] 首先，下文将参照图 8A 的流程图来描述用于提取锚表述及注释表述的处理。

[0095] 在步骤 S801 中，元数据添加目标选择单元 401 参照存储单元 211 中的区域信息，以从添加有“注释”属性的区域中选择还未进行用于提取锚表述及注释表述的处理的一个区域。具体地说，确定是否存在还未处理的注释区域。如果存在还未处理的注释区域，则选择该区域作为处理目标。然后，处理进入步骤 S802。如果不存在具有“注释”属性的区域或者如果已经对所有区域进行了处理，则用于提取锚表述及注释表述的处理完成。在输入图像数据 901 至 907 的情况下，第一例程 (routine) 选择注释区域 912，之后的例程依次选择注释区域 918 和 920。

[0096] 在步骤 S802 中，元数据处理控制单元 406 向选择的附有注释区域的对象区域中添加注释标识符，并在存储单元 211 中确保用于所添加的注释标识符的元数据的存储区域。在选择了注释 912 的情况下，注释标识符“1”被添加到附有注释 912 的照片对象 911 中，并且在存储单元 211 中确保用以存储用于注释标识符“1”的元数据的存储区域。

[0097] 在步骤 S803 中，锚表述提取单元 402 从选择的注释区域的字符识别信息中提取锚表述和注释表述。附加对象的注释可以仅包括锚表述、仅包括注释表述或者包括锚表述和注释表述两者。例如，图的锚表述常常由诸如“图”、“第几图”或者“Fig.”的特定字符串（锚字符串）和编号或符号的组合来表现。因此，预先准备包括登记的锚字符串作为候选的词典。可以将词典中的这些候选与字符识别信息进行比较，由此识别锚表述（锚字符串 + 编号 / 符号）。然后，在注释区域的字符识别信息中，将不构成锚表述的字符串确定作为注释表述。例如，当注释区域的字符识别信息是由字符串“图 1AAA”构成的区域 912 时，“图 1”部分是锚表述而“AAA”部分是注释表述。可能存在注释表述具有非常少的字符数或者注释表述不具有有意义的字符串（例如符号串“-----”）的情况。在这种情况下，存在如下可能性，也即可能将不是字符的标记（例如文档的边界）错误地识别为字符串。因此，不提取该标记作为注释表述。

[0098] 在步骤 S804 中，元数据处理控制单元 406 确定在步骤 S803 中是否从注释区域提取到锚表述和 / 或注释表述。具体地说，确定是否提取到用作添加有注释标识符的对象的元数据的锚表述和注释表述。当提取到这种表述时，处理进入步骤 S805。当未提取到这种表述时，处理返回到步骤 S801。

[0099] 在步骤 S805 中，元数据处理控制单元 406 将从注释区域（锚表述或注释表述或锚表述和注释表述两者）中提取到的元数据，存储在步骤 S802 中确保的用于元数据的存储区域中。

[0100] 通过如上所述的处理，例如提取到的锚表述作为特定对象的元数据通过注释标识符被适当地关联。

[0101] 当用于通过锚表述和注释表述的提取来提取元数据的处理完成时，则开始用于提取正文中的说明文本的处理。

[0102] 下文将参照图 8B 的流程图来描述用于提取正文中的说明文本的处理。

[0103] 在步骤 S806 中，元数据处理控制单元 406 选择还未进行用于提取正文中的说明文本的处理的一个锚表述。具体地说，元数据处理控制单元 406 确定是否存在还未处理的锚表述。如果存在还未处理的锚表述，则选择该锚表述。然后，处理进入步骤 S807。如果不存在锚表述，或者如果对所有的锚表述进行了处理，则用于提取正文中的说明文本的处理完成。在上述示例中，从注释区域 912 中提取“图 1”作为锚表述。因此，首先选择该锚表述。

然后，处理进入步骤 S807。

[0104] 在步骤 S807 中，元数据处理控制单元 406 从存储在存储单元 211 中的正文区域的字符识别信息中提取包含提取的锚表述的部分（句子），以将提取的正文中的说明文本作为元数据与对象关联。在上述示例中，“图 1”被提取作为锚表述。因此，在图像数据 901 的正文区域 908 中包含标号 916 所示的词“图 1”。由于该词与锚表述“图 1”相同，因此正文区域 908 被提取作为正文中的说明文本并作为照片对象 911 的元数据关联。

[0105] 还可以进一步对提取到的作为正文中的说明文本的正文区域进行分析，由此提取元数据。例如，可以使用例如自然语言处理的形态分析来分离词，或者例如可以确定获得的词类 (word class) 或重要的词，由此提取正文区域 908 中的词“照相机”作为元数据。

[0106] 存在提取到正文中的多个说明文本的可能性。还可能存在从提取到的正文中的说明文本进一步获得元数据的可能性。因此，该步骤可以提取多个元数据。当用于提取正文中的说明文本的处理完成时，则处理进入步骤 S808。

[0107] 在步骤 S808 中，元数据处理控制单元 406 确定是否提取到用作添加有注释标识符的对象区域的元数据的正文中的说明文本。如果提取到这种正文中的说明文本，则处理进入步骤 S809。如果未提取到这种正文中的说明文本，则处理返回到步骤 S806，对其它未处理的锚表述重复处理。

[0108] 在步骤 S809 中，元数据处理控制单元 406 将提取到的正文中的说明文本存储（或者添加）在步骤 S802 中确保的用于元数据的存储区域中。在存储之后，处理返回到步骤 S806 以继续对其它未处理的锚表述的处理。

[0109] 通过上述处理，提取到的正文中的说明文本作为特定对象的元数据通过注释标识符被适当地关联。

[0110] 当对所有锚表述的处理完成时，元数据收集 / 输出单元 405 收集得到的元数据并将元数据以格式转换单元 305 能够接收的图像数据的形式添加到图像数据中。然后，将添加有元数据的图像数据发送到格式转换单元 305。

[0111] （正文中的说明文本与对象之间的关联）

[0112] 接下来，下文将参照图 10 的流程图来描述图 8B 的步骤 S807 中的处理的详情。

[0113] 在步骤 S1001 中，元数据处理控制单元 406 获取具有包含在图 8B 的步骤 S806 中选择的特定锚表述的注释的对象的数量。假定例如针对图像数据 901 至 907 选择了锚表述“图 2”的情况。在这种情况下，除了照片对象 919 以外不存在包含“图 2”的对象。因此，获取“1”作为对象的数量。另一方面，当选择了锚表述“图 1”时，照片对象 911 和照片对象 917 作为具有包含“图 1”的注释的对象存在。因此，选择“2”作为对象的数量。

[0114] 在步骤 S1002 中，元数据处理控制单元 406 确定在步骤 S1001 中获取的对象的数量是否是多个。当获取的对象的数量是“1”时（即当不再存在具有包含相同锚表述的注释的其它对象时），处理进入步骤 S1003。当获取的对象的数量是多个时（即当多个对象的注释使用相同的锚表述时），处理进入步骤 S1005。

[0115] 在步骤 S1003 中，正文内搜索单元 403 在正文区域的字符识别信息中搜索特定的锚表述。在上述示例的情况下，包含“图 2”作为锚表述 916 的正文区域 915 被提取作为正文中的说明文本。当在正文区域的字符识别信息中找到包含锚表述的表述时，处理进入步骤 S1004。

[0116] 在步骤 S1004 中，元数据处理控制单元 406 将正文中包含特定锚表述的表述（正文中的说明文本）与对象关联。在上述示例的情况下，将提取到的作为包含锚表述“图 2”的正文中的说明文本的正文区域 915 与照片对象 919 关联。

[0117] 在步骤 S1005 中，正文内搜索单元 403 如步骤 S1003 中那样在正文区域的字符识别信息中搜索特定的锚表述。在上述示例的情况下，找到正文区域 908 和 913 作为包含锚表述“图 1”的正文中的说明文本。当如上所述在正文区域的字符识别信息中找到包含特定的锚表述的正文中的说明文本时，处理进入步骤 S1006。

[0118] 在步骤 S1006 中，元数据处理控制单元 406 针对找到的正文中的说明文本（正文区域）准备用于找到与可能的对象区域的注释之间的最佳对应关系所需的多种组合。然后，对各个组合计算类似度。在上述示例的情况下，例如，在包含相同锚表述“图 1”的注释 912 和 918 与包含该锚表述的正文区域 908 和 913 之间准备两种组合。具体地说，在这种情况下，如果确定了与正文中的一个说明文本相对应的注释，则也确定了其它注释。因此，准备两种组合 :注释 912 和正文区域 908 的组合以及注释 918 和正文区域 908 的组合。作为另选方案，还可以准备相应的其它两种组合 :注释 912 和正文区域 913 的组合以及注释 918 与正文区域 913 的组合。如果存在包含相同锚表述的三个注释（对象）并且找到正文中可以与所述注释相对应的三个说明文本，则至少计算 5 种组合。

[0119] 稍后将描述类似度的计算的详情。在类似度的计算完成之后，处理进入步骤 S1007。

[0120] 在步骤 S1007 中，元数据处理控制单元 406 基于步骤 S1006 中的计算结果认定类似度的值最大的组合是最佳的。然后，元数据处理控制单元 406 将组合中的正文中的说明文本与对象区域关联。

[0121] （用于计算类似度的处理）

[0122] 参照图 11 的流程图，针对上述两种组合（注释 912 和正文区域 908 的第一组合以及注释 918 和正文区域 908 的第二组合）的示例来描述图 10 的步骤 S1006 中的类似度的计算。

[0123] 在步骤 S1101 中，表述类似度计算单元 404 从计算目标的组合的注释中获取注释表述。当第一组合是计算目标时，获取通过从注释 912 中移除锚表述“图 1”而获得的“AAA”作为注释表述。在获取之后，处理进入步骤 S1102。

[0124] 在步骤 S1102 中，表述类似度计算单元 404 通过形态分析对获取的注释表述进行词分离，由此获得关于各个词的词类信息。然后，基于获得的词类信息，选择具有诸如名词的词类的词（以下称为“注释词”）作为比较目标。具体地说，从比较目标中排除不重要的词或锚表述，以使得不选择该词或锚表述。例如，注释表述“AAA”提供一个名词“AAA”。获得注释词“AAA”及词类信息“名词”。结果，选择了“AAA”作为比较目标候选。在该步骤中，可以选择多个注释词。

[0125] 在步骤 S1103 中，表述类似度计算单元 404 将注释词的权重设置为 1。具体地说，在本实施例中，不基于例如距锚表述的距离或者词的词类或词性来计算权重值。然而，当注释表述是具有某一长度的句子时或者当基于例如唯一的表述提取或者词典来计算词的重要性时，也可以使用如上所述通过计算获得的值作为权重。

[0126] 图 12A 以表的形式示出通过如上所述的处理获得的第一组合的结果。在该表中，

注释表述和注释词是“AAA”，注释词属性是“名词”，并选择了比较目标候选。距锚表述的距离信息被设置为信息“-”，其表示不使用该距离信息。将权重设置为“1”。

[0127] 接下来，在步骤S1104中，表述类似度计算单元404在计算目标的组合中获取正文中的说明文本（正文区域）。在第一组合的情况下，获取正文区域908作为包含“图1”作为锚表述909的正文中的说明文本“照相机AAA（图1）用于例如风景的拍摄”。

[0128] 在步骤S1105中，表述类似度计算单元404通过形态分析对获取的正文中的说明文本进行词分离，由此获得各个词的词类信息。然后，基于获得的词类信息，选择具有诸如名词的词类的词（以下称为“说明文本词”）作为比较目标。另外，进行该选择以使得从比较目标中排除不重要的词或锚表述，并且不选择不重要的词或锚表述。在该步骤中，选择了多个说明文本词。

[0129] 在步骤S1106中，表述类似度计算单元404将选择的说明文本词按照在正文中的说明文本中的位置到锚表述更近的顺序（即按照距锚表述的距离（词距离）更短的顺序）布置。将距锚表述的距离简单地设置为从锚表述到说明文本词的词数量。以下，将按照词距离的顺序布置的说明文本词的列称为“说明文本词串”。然后，将包含在该说明文本词串中的说明文本词的数量设置为变量N的值。

[0130] 图12B以表的形式示出对获取的正文中的说明文本“照相机AAA（图1）用于例如风景的拍摄”进行步骤S1105和S1106的处理的结果。三个说明文本词被选择作为比较目标候选，并分别给予值“3”、“2”、“1”作为在说明文本词串中的顺序，以使得可以按照词到锚表述“图1”距离更短的顺序（即按照“AAA”、“照相机”、及“风景”的顺序）提取这些词。在这种情况下，包含在说明文本词串中的说明文本词的数量是3。因此，变量N的值被设置为“3”。

[0131] 在步骤S1107中，表述类似度计算单元404对示出类似度的变量S的值进行初始化。具体地说，表述类似度计算单元404将作为类似度S的预定值设置为“0”。

[0132] 在步骤S1108中，表述类似度计算单元404确定变量N的值是否大于“0”。当变量N大于“0”时，其表示存在作为未处理比较目标的说明文本词。因此，处理进入步骤S1109。当变量N的值是“0”时，其表示已经对作为比较目标的所有说明文本词进行了处理。因此，进行由步骤S1108至步骤S1113构成的例程并完成处理。

[0133] 在步骤S1109中，表述类似度计算单元404从说明文本词串中提取在说明文本词串中的顺序具有与变量N的值相同的值的说明文本词。在上述示例的情况下，首先提取了在说明文本词串中顺序为“3”的说明文本词“AAA”。

[0134] 在步骤S1110中，表述类似度计算单元404基于示出距锚表述的距离的词距离来计算说明文本词的权重。权重具有依据词距离衰减(attenuate)的值并且具有最大值“1”和最小值“0”。在本示例中，权重是词距离的倒数。在说明文本词“AAA”的情况下，其词距离为“2”。因此，计算出“0.5”作为权重。可以通过上述方法以外的各种方法（包括例如权重按线性方式衰减的方法、使用对数的方法或者使用分布函数的方法）来获得权重。然而，也可以使用任意方法，只要该方法提供权重依据距离衰减即可。

[0135] 在步骤S1111中，表述类似度计算单元404将选择的说明文本词与注释词进行比较以计算类似度S。具体地说，如下面公式1所示，当前词与后词具有相同的字符串时，通过将“1”乘以前词和后词的权重来计算类似度S。当前词和后词具有不同的字符串时，通过将

“0”乘以前词和后词的权重来计算类似度 S。

[0136] 类似度 S = (1 或 0)×(注释词的权重)×(说明文本词的权重)…(公式 1)

[0137] 以这种方式,对作为比较目标的说明文本词计算类似度 S。当存在多个注释词时,将注释词与选择的说明文本词逐一比较。然后,将计算出的类似度 S 中具有最大值的类似度 S 确定为 Smax。

[0138] 在本实施例中,通过简单的计算获得类似度S。然而,还可以通过考虑例如同义词、不一致的表示或者诸如平假名、片假名或汉字的表示的这种比较来获得类似度S。还可以通过使用同义词的比较或使用用于基于上下文确定多义性的比较语言的比较来获得类似度S。还可以通过考虑例如字符串的包含率或同一字符混入的比率来获得类似度 S。

[0139] 接下来,在步骤 S1112 中,表述类似度计算单元 404 将在步骤 S1111 中计算出的类似度 S(或 Smax) 的值与先前的类似度 S 的值相加,由此更新类似度 S 的值。在例如第一例程的情况下,将初始值设为“0”。因此,将在步骤 S1111 中计算出的类似度 S 的值直接用作更新的类似度 S。在第二及之后的例程的情况下,将通过先前例程获得的类似度 S 的值与最近确定的类似度 S(或 Smax) 的值相加。

[0140] 在步骤 S1113 中,表述类似度计算单元 404 从变量 N 的值中减去 1(一)。然后,处理返回到步骤 S1108。

[0141] 在图 12B 的示例中,有三个说明文本词作为比较对象候选。因此,将上述例程重复 3 次。

[0142] 在第一例程中,将变量N的值设为“3”。因此,在步骤S1108中的确定之后进行步骤S1109。在步骤 S1109 和 S1110 中,如上所述,在“AAA”的提取之后,计算“0.5”作为词的权重。然后,在步骤 S1111 中,进行计算说明文本词“AAA”与注释词“AAA”之间的类似度 S 的处理。在这种情况下,前词与后词具有相同的字符串。因此,计算类似度 S 为“1×1×0.5”。

[0143] 由于仅有一个注释词“AAA”,因此在步骤 S1112 中将得到的类似度 S = 0.5 加到初始值“0”中。然后,处理进入步骤 S1113。在步骤 S1113 中,从变量 N 的值中减去 1(一)以得到 N =“2”。然后,处理返回到步骤 S1108。之后,将相同的处理重复两次。还计算说明文本词“照相机”与“风景”之间的类似度 S,然后不断更新类似度 S。然而,由于说明文本词“照相机”和“风景”与仅有的注释词“AAA”不同,因此对于说明文本词“照相机”和“风景”两者来说,在步骤 S1111 中计算的类似度 S 的值均是“0”。因此,最终获得的类似度 S 的值是 0.5+0+0 = 0.5。

[0144] 如上所述,描述了第一组合的处理。还对由注释 918 和正文区域 908 构成的第二组合进行类似的处理。在第二组合的情况下,从注释 918 获得的注释表述是“BBB”。因此,注释表述与说明文本词不同。因此,最终获得的类似度 S 的值为 0。结果,在图 10 的步骤 S1007 中,元数据处理控制单元 406 认定类似度 S 为 0.5 的第一组合是最佳组合。具体地说,将具有根据第一组合的注释 912 的照片对象 911 与正文中的说明文本“照相机 AAA(图 1) 用于例如风景的拍摄”(正文区域 908) 关联。

[0145] 代替图 11 的流程图所示的方法,例如,还可以使用其它方法,诸如使用相同字符被使用的比率或相同字符被连续使用的程度的方法或者仅提取重要的表述由此获得用于比较的特定表述(字符串)的方法。

[0146] 图 13 以表的形式示出通过对图像数据 901 至 907 进行上述处理获得的元数据信

息。

[0147] 例如，如标号 1301 所示，通过注释标识符“1”将正文区域 908 的正文中的说明文本、注释 912 及其锚表述“图 1”与照片对象 911 适当地关联。如标号 1302 所示，通过注释标识符“2”同样将正文区域 913 的正文中的说明文本、注释 918 及其锚表述“图 1”与照片对象 917 适当地关联。如标号 1303 所示，通过注释标识符“3”将正文区域 915 的正文中的说明文本、注释 920 及其锚表述“图 2”与照片对象 919 适当地关联。

[0148] 在实际情况下，使用例如 SVG 格式描述图 13 的表中示出的元数据信息。可以使用与图 6A 所述的方法相同的方法来描述元数据信息。具体地说，在图 6A 中，注释标识符 608 被添加到对象数据 603，并且相同的标识符作为注释标识符 609 被添加到对象的元数据 607。类似地，在图 13 中，与添加到各个对象的注释标识符 (1 至 3) 相同的标识符被添加到与各个对象相对应的元数据，以使得能够识别与各个对象相对应的元数据。

[0149] 如上所述，在根据本实施例的图像处理装置中，能够将文档中的对象（例如照片、图或表）与描述其内容的正文中的说明文本适当地关联。结果，即使当在不同文档的不同对象的注释中共同使用特定的锚表述时，能够针对各个对象准备添加有适当元数据的电子文档数据。此外，能够在大大降低由页面布局或页之间的距离带来的影响的同时提取元数据。因此，即使在页随机布置的图像数据中也能够添加适当的元数据。

[0150] [实施例 2]

[0151] 接下来，下文将参照图 14A、图 14B 和图 15 来描述实施例 2 在图 8B 的步骤 S807 中的处理。实施例 2 是能够提高用于提取包含选择的锚表述的正文中的说明文本、以将提取的文本与对象区域关联的处理的效率的实施例。

[0152] 图 14A 和图 14B 是示出本实施例中的将正文中的说明文本与对象区域关联的处理的流程图。对与根据实施例 1 的图 10 的流程图相同的部分进行简单的描述或者不再描述。因此，主要描述不同点。

[0153] 在步骤 S1401 中，获取具有包含特定锚表述的注释的对象的数量。在步骤 S1402 中，当获取的对象的数量是多个时，处理进入步骤 S1405。当对象的数量是“1”时，后续的处理 (S1403 和 S1404) 与图 10 的步骤 S1003 和 S1004 相同。

[0154] 在步骤 S1405 中，元数据处理控制单元 406 搜索正文区域中的锚表述。在步骤 S1406 中，元数据处理控制单元 406 确定得到的正文中的说明文本是否与诸如照片的对象存在于同一页中。在通过扫描器读取这些页而获得的各个页的图像数据的情况下，很少出现一页的数据与其它文档的图像数据混合。因此，当诸如照片的对象与通过搜索找到的正文中的说明文本存在于同一页中时，认为对象与说明文本之间具有对应关系。因此，进行该确定处理以使得在不计算类似度的情况下就能够进行关联。当判断为诸如照片的对象与正文中的说明文本存在于同一页中时，处理进入步骤 S1407。在步骤 S1407 中，元数据处理控制单元 406 将包含锚表述的正文中的说明文本与同一页中的对象区域关联。之后，处理进入步骤 S1409。当判断为诸如照片的对象不与正文中的说明文本在同一页中时，处理进入步骤 S1408。

[0155] 在步骤 S1408 中，元数据处理控制单元 406 确定是否存在其它通过搜索找到的正文中的说明文本。具体地说，元数据处理控制单元 406 确定是否存在必须计算类似度的、对象区域和正文中的说明文本的组合。如果确定不再存在正文中的说明文本，则处理完成。如

果存在其它正文中的说明文本，则处理进入步骤 S1409。

[0156] 在步骤 S1409 中，元数据处理控制单元 406 确定在步骤 S1401 中获取的对象的数量是否是 3 或者更多。当对象的数量是 2 时，处理进入步骤 S1410。当对象的数量是 3 或者更多时，处理进入步骤 S1413。当对象的数量是 3 或者更多时的步骤 S1413 和步骤 S1414 中的处理与图 10 的步骤 S1006 和步骤 S1007 中的处理相同。

[0157] 在步骤 S1410 中，表述类似度计算单元 404 对一个对象区域与正文中的说明文本的组合进行图 11 的流程图的上述处理，由此计算类似度 S。

[0158] 接下来，在步骤 S1411 中，表述类似度计算单元 404 使用在步骤 S1410 中获得的类似度 S，对其它对象区域与正文中的说明文本的组合进行图 15 的流程图所示的处理。具体地说，进行以下处理。

[0159] 首先，在步骤 S1501 中，表述类似度计算单元 404 获取在步骤 S1410 中获得的类似度 S，以将该类似度 S 作为用作稍后描述的步骤 S1514 中的比较目标的类似度 Scomp 保持在 RAM 206 中。然后，处理进入步骤 S1502。

[0160] 在步骤 S1502 到步骤 S1513 中，进行与图 11 的步骤 S1101 到步骤 S1112 的处理相同的处理。当在步骤 S1513 中进行了用于类似度 S 的第一更新的处理时，处理进入步骤 S1514。

[0161] 在步骤 S1514 中，表述类似度计算单元 404 将在步骤 S1501 中获取并保持的值 Scomp 与在步骤 S1513 中更新的类似度 S 的值进行比较以确定这两个值中的哪一个更大。当更新的类似度 S 的值大于 Scomp 时，处理完成。然后，处理进入步骤 S1412。其原因是确定了获得的类似度 S(Scomp) 小于步骤 S1410 中的类似度 S。当更新的类似度 S 的值小于值 Scomp 时，处理进入步骤 S1515 以进行计算类似度 S 的第二例程。

[0162] 当在第二及随后的例程中在没有确定更新的类似度 S 的值大于 Scomp 的情况下变量 N 的值为 0 时，处理完成。然后，处理进入步骤 S1412。在此时点，确定为在步骤 S1410 中获取的类似度 S(Scomp) 大于在步骤 S1411 中获取的类似度 S。

[0163] 然后，在步骤 S1412 中，元数据处理控制单元 406 将类似度 S 的值较大的组合中的正文中的说明文本与对象区域关联。

[0164] 如上所述，根据本实施例，依据诸如照片的对象和正文中的说明文本是否存在于同一页中以及对象的数量是否是 3 或更多，能够省略部分处理。因此，能够以较高的速度进行处理。

[0165] [实施例 3]

[0166] 在实施例 1 和实施例 2 中，仅基于计算出的类似度的值是较大还是较小，来将诸如照片的对象与正文中的说明文本关联。下文将描述在计算类似度之后、确定计算出的类似度的值是否达到预定阈值的实施例。仅当计算出的类似度的值达到预定阈值时，将对象区域与正文中的说明文本关联。

[0167] 图 16 是示出根据本实施例的、用于将正文中的说明文本与对象区域关联的处理的流程的流程图。图 16 的流程图与实施例 1 的图 10 的流程图以及实施例 2 的图 14A 和图 14B 的流程图相对应。因此，对它们之间的共同部分进行简单地描述或者不再进行描述。因此将主要描述不同点。

[0168] 步骤 S1601 到步骤 S1608 中的处理与图 14A 和图 14B 的步骤 S1401 到步骤 S1408

中的处理相同。步骤 S1609 的处理与图 10 的步骤 S1006 的处理相同。当在步骤 S1609 中计算了组合的类似度时，处理进入步骤 S1610。

[0169] 在步骤 S1610 中，表述类似度计算单元 404 将最大的类似度的值与预先设置的预定阈值进行比较，以确定类似度的值是否等于或者大于阈值。当类似度的值等于或者大于阈值时，处理进入步骤 S1611 以将正文中的说明文本与对象区域关联。当类似度的值未达到阈值时，不将正文中的说明文本与对象区域关联并且处理完成。

[0170] 图 17 示出在 MFP 100 的操作单元 203 上显示的用户界面 (UI) 画面的示例。在 UI 画面 1710 上具有用于指定搜索文档中的对象（例如照片、图或者表）的功能的等级的按钮 1702 和 1703。当选择了按钮 1702 时，通过具有高级对象搜索功能的方法（即根据本发明的方法）来准备电子文档。当选择了按钮 1703 时，通过文件大小的压缩优先的传统方法来准备电子文档。标号 1704 表示允许用户将上述阈值设置为任意值的按钮。为了提高文档中的元数据的提取等级，将按钮 1704 滑动到右侧。这因此减小了阈值，由此提取更多的元数据。另一方面，为了降低文档中的元数据的提取等级，将按钮 1704 滑动到左侧。这因此提高了阈值，由此提取更少的元数据。通过该用户界面，用户能够任意地改变阈值，以将元数据的提取等级改变为期望的等级。标号 1705 表示用于取消选择的内容的取消按钮。标号 1706 表示用于确定设置内容的确认按钮。

[0171] 根据本实施例，将类似度的值小于阈值的情况认定为不可能提取正确的元数据的情况，由此防止添加元数据。这因此能够防止正文中的说明文本错误地与对象区域关联而引起的错误元数据的添加的情形。因此，能够正确地进行后续的对象搜索。

[0172] [实施例 4]

[0173] 接下来，描述实施例 4，在实施例 4 中，当发现存在具有包含特定锚表述的注释的多个对象时，通过警告显示来向用户警告该存在。

[0174] 图 18 示出在本实施例中的 MFP 100 的操作单元 203 上显示的警告画面的示例。图 19 是根据本实施例的、用于将正文中的说明文本与对象区域关联的处理的流程的流程图。对与根据实施例 1 的流程图相同的部分进行简单的描述或者不再描述。因此，将主要描述不同点。

[0175] 在步骤 S1901 中，获取具有包含特定锚表述的注释的对象的数量。当在步骤 S1902 中确定获取的对象的数量是多个时，处理进入步骤 S1905。

[0176] 在步骤 S1905 中，元数据处理控制单元 406 在例如操作单元 203 上显示如图 18 所示的警告画面。该警告画面包括表示在不同的图中检测到相同的图编号的消息，并且还包括用于指定是否继续处理的按钮。

[0177] 当用户在警告画面 1801 上选择“继续”按钮 1802 时，处理进入步骤 S1907 以继续处理。步骤 S1907 至步骤 S1909 的处理与图 10 中的步骤 S1005 至步骤 S1007 的处理相同。另一方面，当用户选择“完成”按钮 1803 时，处理停止以返回到扫描开始之前的状态。

[0178] 根据本实施例，当发现存在具有包含相同锚表述的注释的多个对象时，用户能够有机会考虑是否继续处理。因此，为了添加正确的元数据，用户能够有机会停止处理以尝试第二次扫描。

[0179] 本发明的各方面还能够通过读出并执行记录在存储设备上的用于执行上述实施例的功能的程序的系统或装置的计算机（或诸如 CPU 或 MPU 的设备）、以及由系统或装置的

计算机例如读出并执行记录在存储设备上的用于执行上述实施例的功能的程序来执行步骤的方法来实现。鉴于此，例如经由网络或者从用作存储设备的各种类型的记录介质（例如计算机可读介质）向计算机提供程序。

[0180] 虽然参照示例性实施例描述了本发明，但是应当理解，本发明不限于所公开的示例性实施例。应对所附权利要求的范围给予最宽的解释，以使其覆盖所有这种变型、等同结构和功能。

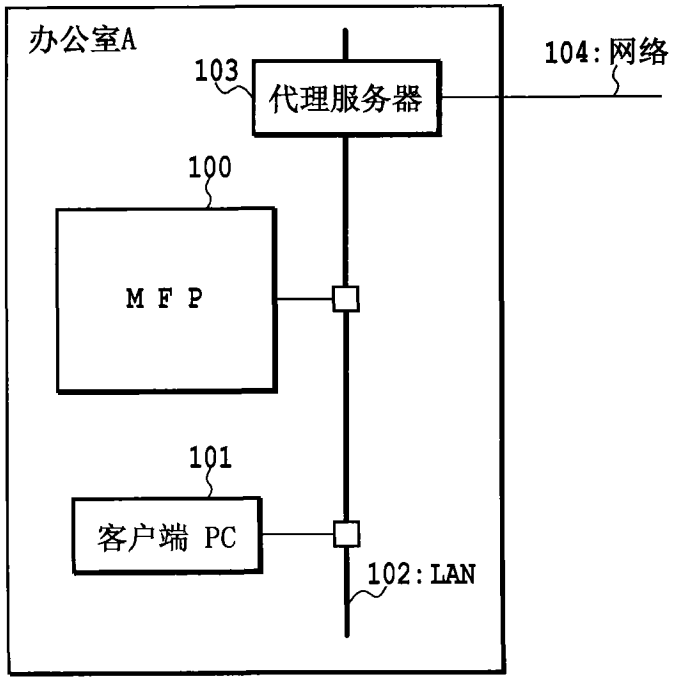


图 1

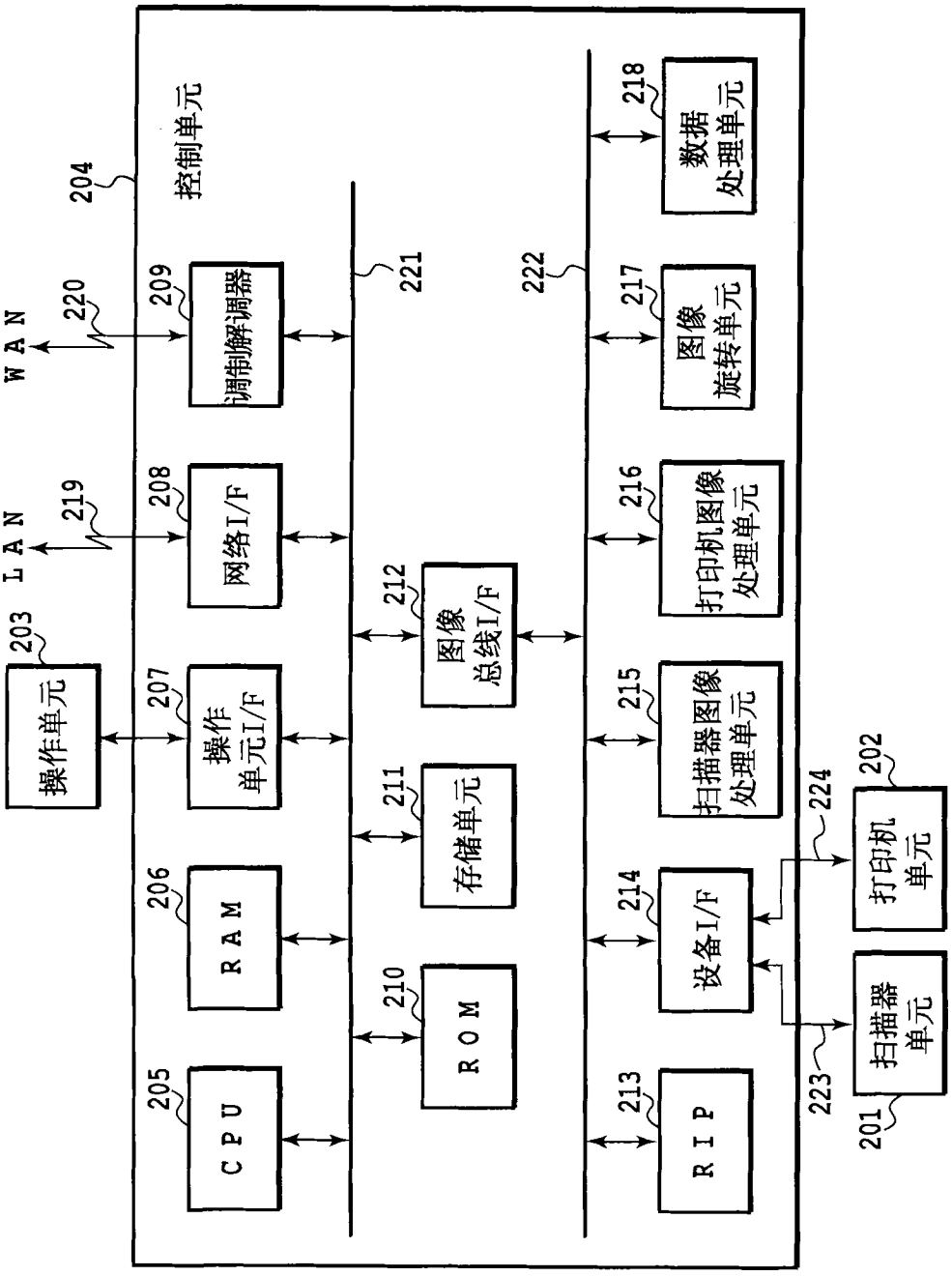

控制单元
204
WAN
220
LAN
219
203
操作单元
209
调制解调器
208
网络I/F
207
操作单元I/F
206
RAM
205
CPU
221
212
图像总线I/F
211
存储单元
210
ROM
222
218
数据处理单元
217
图像旋转单元
216
打印机图像处理单元
215
扫描器图像处理单元
214
设备I/F
213
RIP
224
202
打印机单元
223
201
扫描器单元

图 2

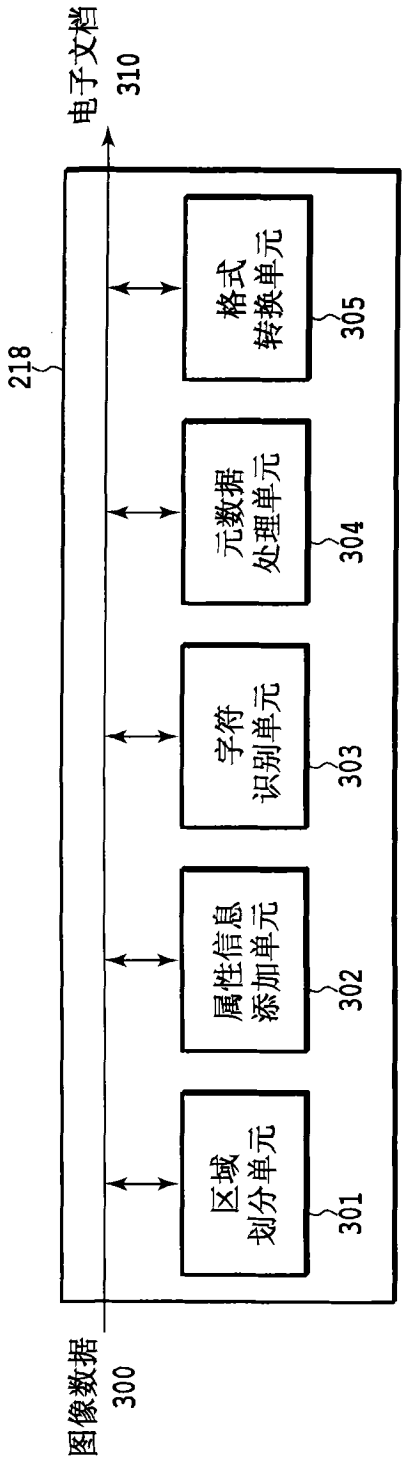

电子文档
310
218
格式
转换单元
305
元数据
处理单元
304
字符
识别单元
303
属性信息
添加单元
302
区域
划分单元
301
图像数据
300

图 3

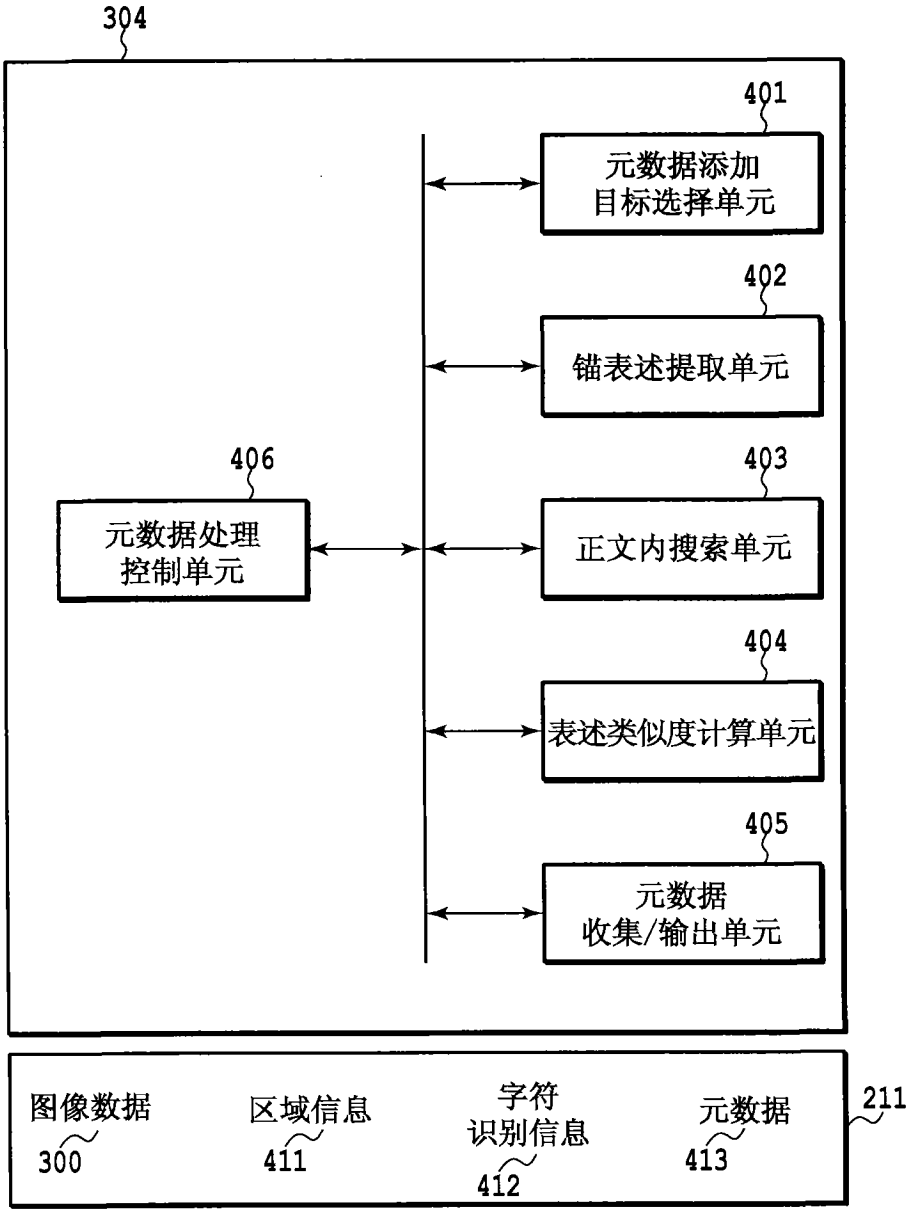

304
401
元数据添加
目标选择单元
402
锚表述提取单元
406
元数据处理
控制单元
403
正文内搜索单元
404
表述类似度计算单元
405
元数据
收集/输出单元
图像数据
300
区域信息
411
字符
识别信息
412
元数据
413
211

图 4

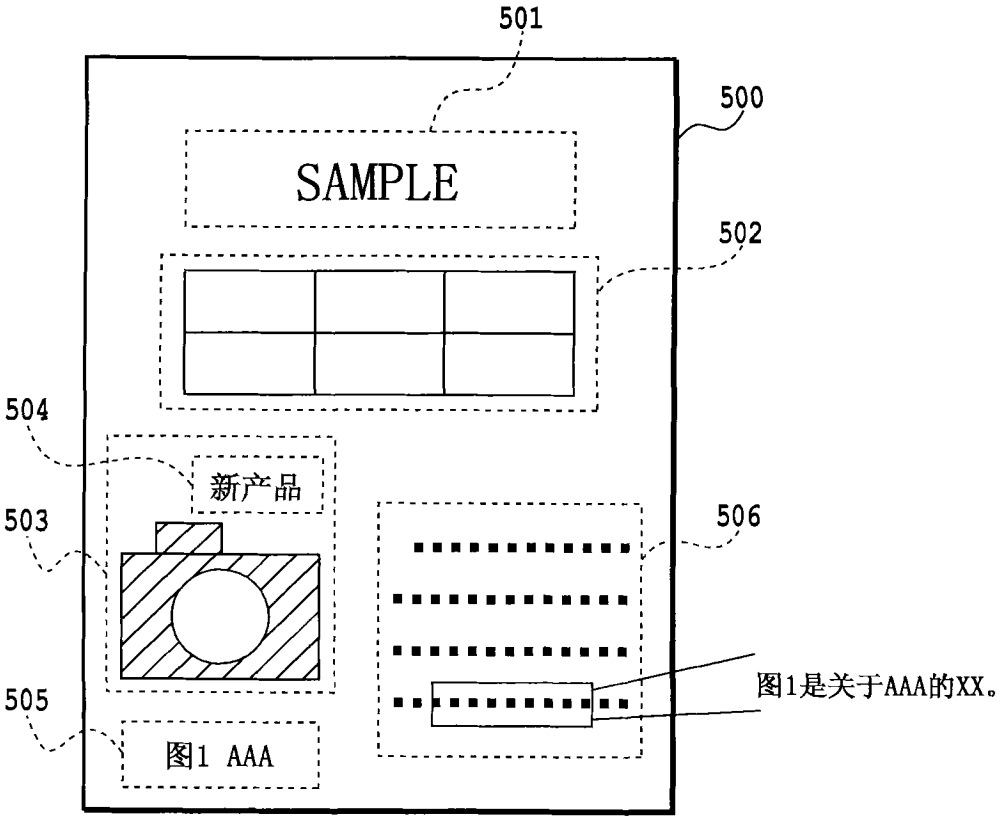


图 5A

区域	属性	附有注释的区域	坐标X	坐标Y	宽度W	高度H	页	字符识别信息
501	标题	-	X1	Y1	W1	H1	1	SAMPLE
502	表	-	X2	Y2	W2	H2	1	-
503	照片	-	X3	Y3	W3	H3	1	-
504	字符（照片503中）	-	X4	Y4	W4	H4	1	新产品
505	注释	503	X5	Y5	W5	H5	1	图1 AAA
506	正文	-	X6	Y6	W6	H6	1	…图1是关于AAA的XX…

图 5B

处理详情 \ 属性	字符	照片	图（线图）	表
转换处理	矢量转换	图像剪切	矢量转换	矢量转换（表框部分）
删除处理	是	是	是	是

图 5C

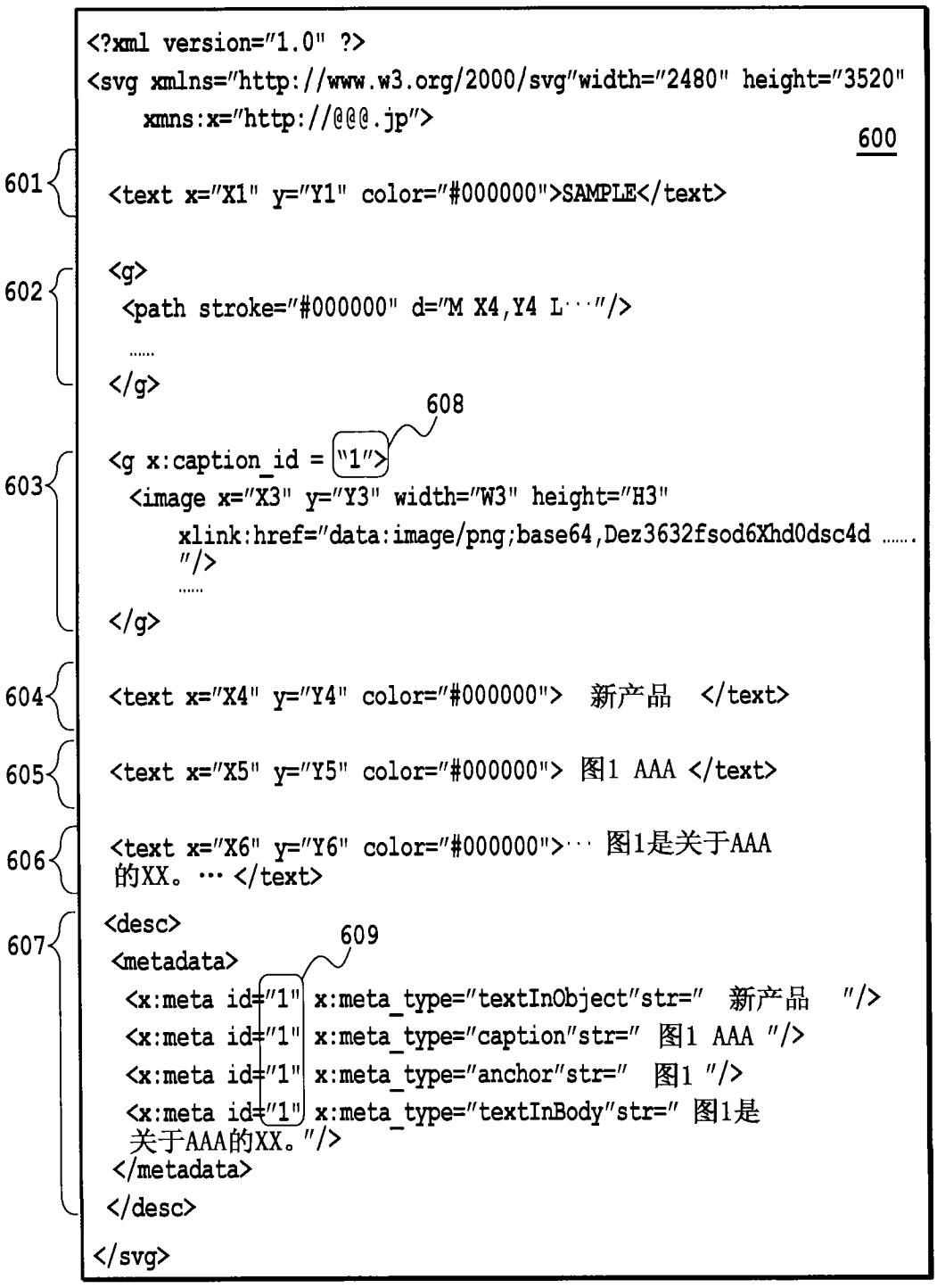

<?xml version="1.0" ?>
<svg xmlns="http://www.w3.org/2000/svg"width="2480" height="3520"
xmns:x="http://@@@.jp">
600
601
<text x="X1" y="Y1" color="#000000">SAMPLE</text>
602
<g>
<path stroke="#000000" d="M X4,Y4 L···"/>
……
</g>
608
603
<g x:caption_id = "1">
<image x="X3" y="Y3" width="W3" height="H3"
xlink:href=" …….
"/>
……
</g>
604
<text x="X4" y="Y4" color="#000000"> 新产品 </text>
605
<text x="X5" y="Y5" color="#000000"> 图1 AAA </text>
606
<text x="X6" y="Y6" color="#000000">··· 图1是关于AAA
的XX。 ··· </text>
607
<desc>
609
<metadata>
<x:meta id="1" x:meta_type="textInObject"str=" 新产品 "/>
<x:meta id="1" x:meta_type="caption"str=" 图1 AAA "/>
<x:meta id="1" x:meta_type="anchor"str=" 图1 "/>
<x:meta id="1" x:meta_type="textInBody"str=" 图1是
关于AAA的XX。"/>
</metadata>
</desc>
</svg>

图 6A

注释标识符	添加目标	提取自	页	提取信息类型	提取的信息
1	503	504	1	对象中的字符	新产品
1	503	505	1	注释	图1 AAA
1	503	505	1	锚表述	图1
1	503	506	1	正文中的说明文本	图1是关于AAA的XX。

图 6B

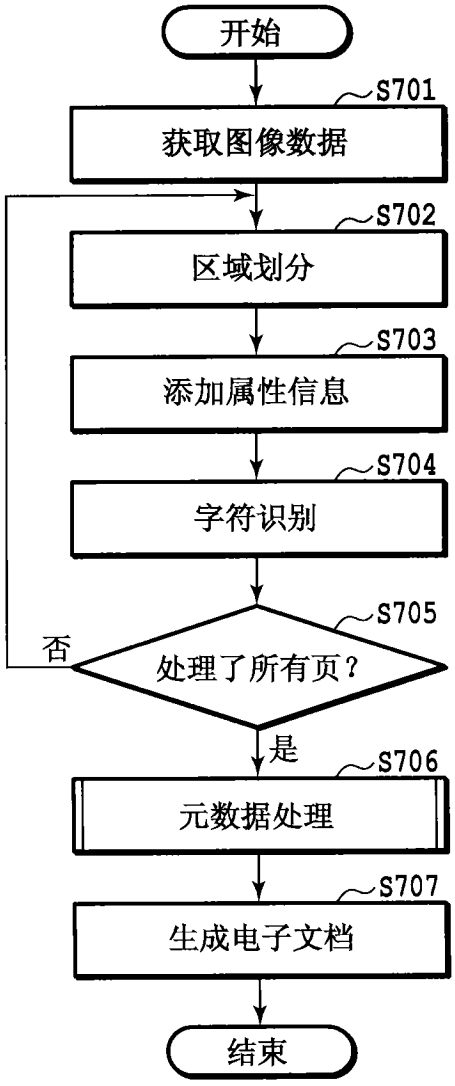


图 7

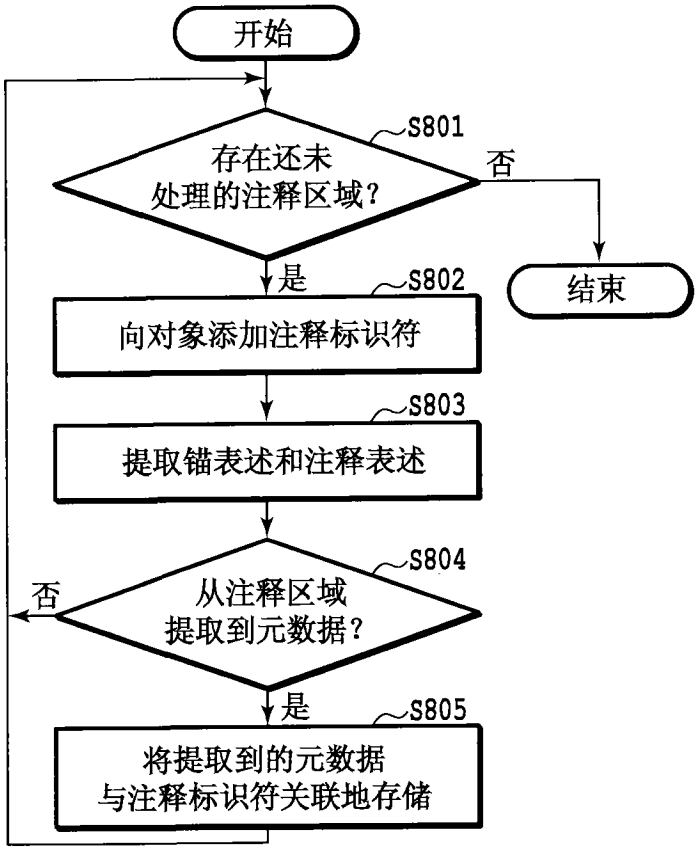


图 8A

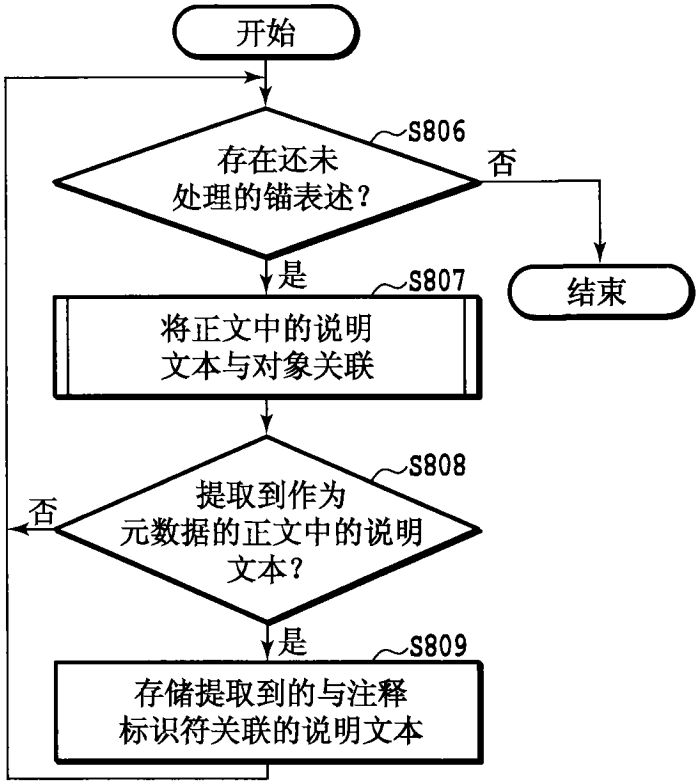

开始
S806
存在还未
处理的锚表述？
否
结束
是
S807
将正文中的说明
文本与对象关联
S808
提取到作为
元数据的正文中的说明
文本？
否
是
S809
存储提取到的与注释
标识符关联的说明文本

图 8B

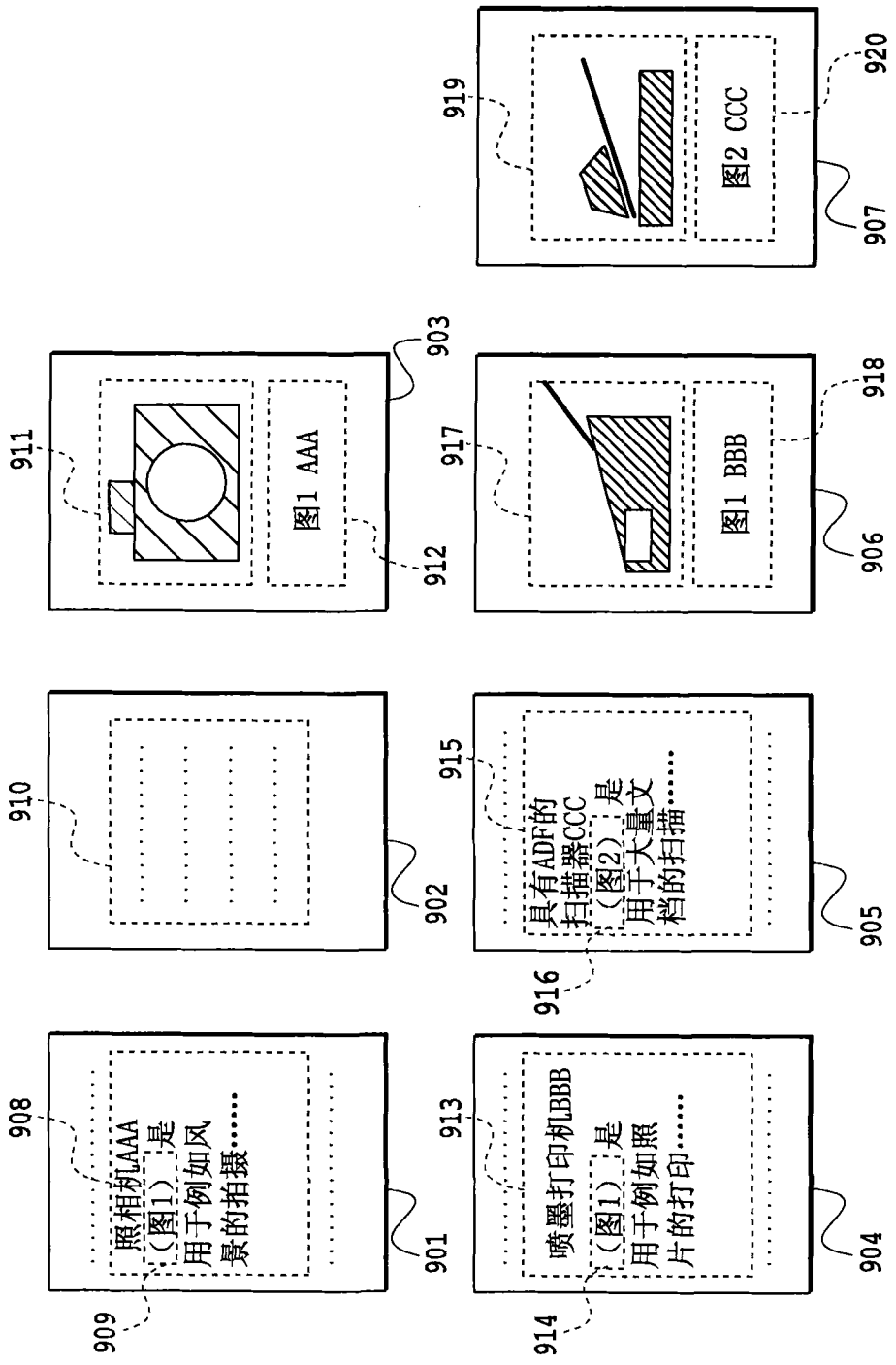

919
920
907
图2 CCC
911
903
912
图1 AAA
917
918
906
图1 BBB
910
902
915
具有ADF的扫描器CCC（图2）是用于大量文档的扫描……
905
916
908
照相机AAA（图1）是用于例如风景的拍摄……
901
909
913
喷墨打印机BBB（图1）是用于例如照片的打印……
904
914

图 9A

区域	属性	附有注释的区域	坐标X	坐标Y	宽度W	高度H	页	字符识别信息
908	正文	-	X08	Y08	W08	H08	1	…照相机AAA（图1）是…
910	正文	-	X10	Y10	W10	H10	2	……
911	照片	-	X11	Y11	W11	H11	3	-
912	注释	911	X12	Y12	W12	H12	3	图1 AAA
913	正文	-	X13	Y13	W13	H13	4	…喷墨打印机BBB（图1）是…
915	正文	-	X15	Y15	W15	H15	5	…具有ADF的扫描器CCC（图2）是用于大量…
917	照片	-	X17	Y17	W17	H17	6	-
918	注释	917	X18	Y18	W18	H18	6	图1 BBB
919	照片	-	X19	Y19	W19	H19	7	-
920	注释	919	X20	Y20	W20	H20	7	图2 CCC

图 9B

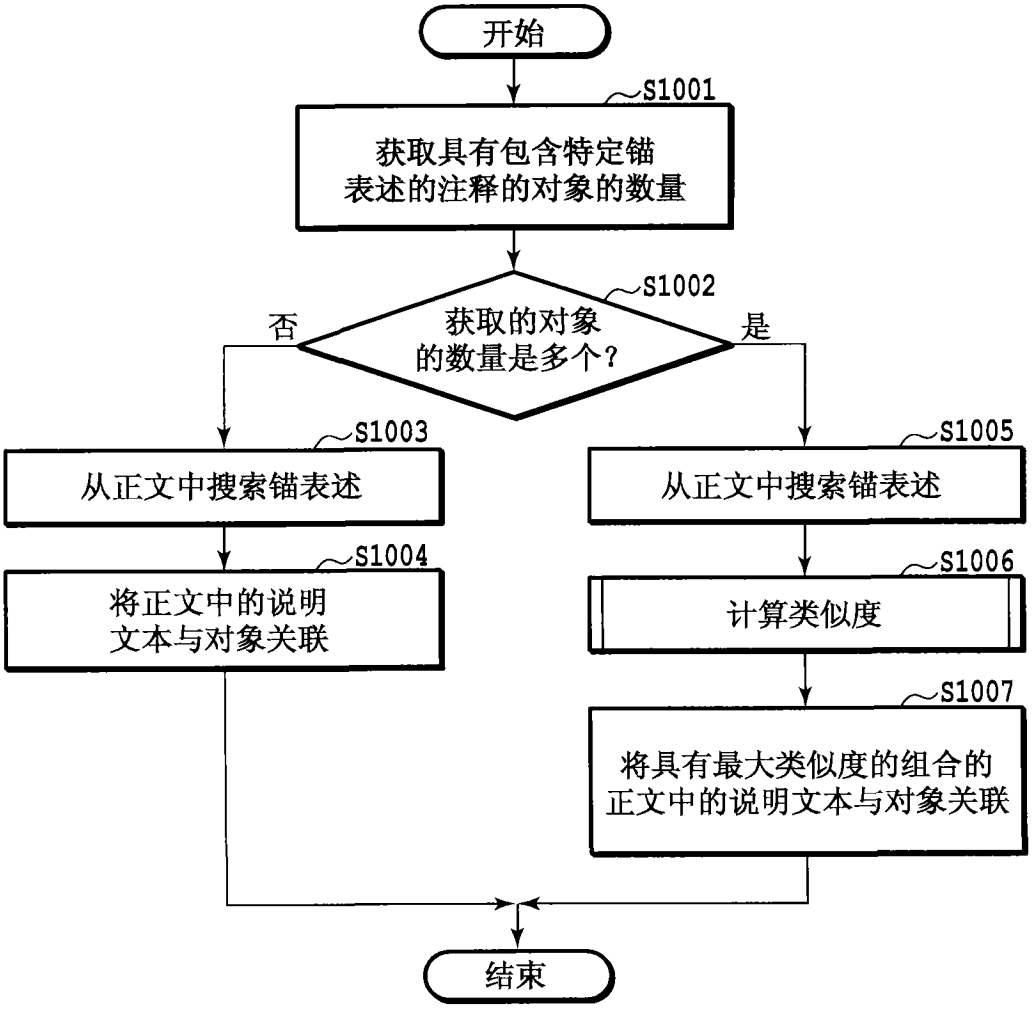

开始
S1001
获取具有包含特定锚表述的注释的对象的数量
S1002
获取的对象的数量是多个？
否
是
S1003
从正文中搜索锚表述
S1004
将正文中的说明文本与对象关联
S1005
从正文中搜索锚表述
S1006
计算类似度
S1007
将具有最大类似度的组合的正文中的说明文本与对象关联
结束

图 10

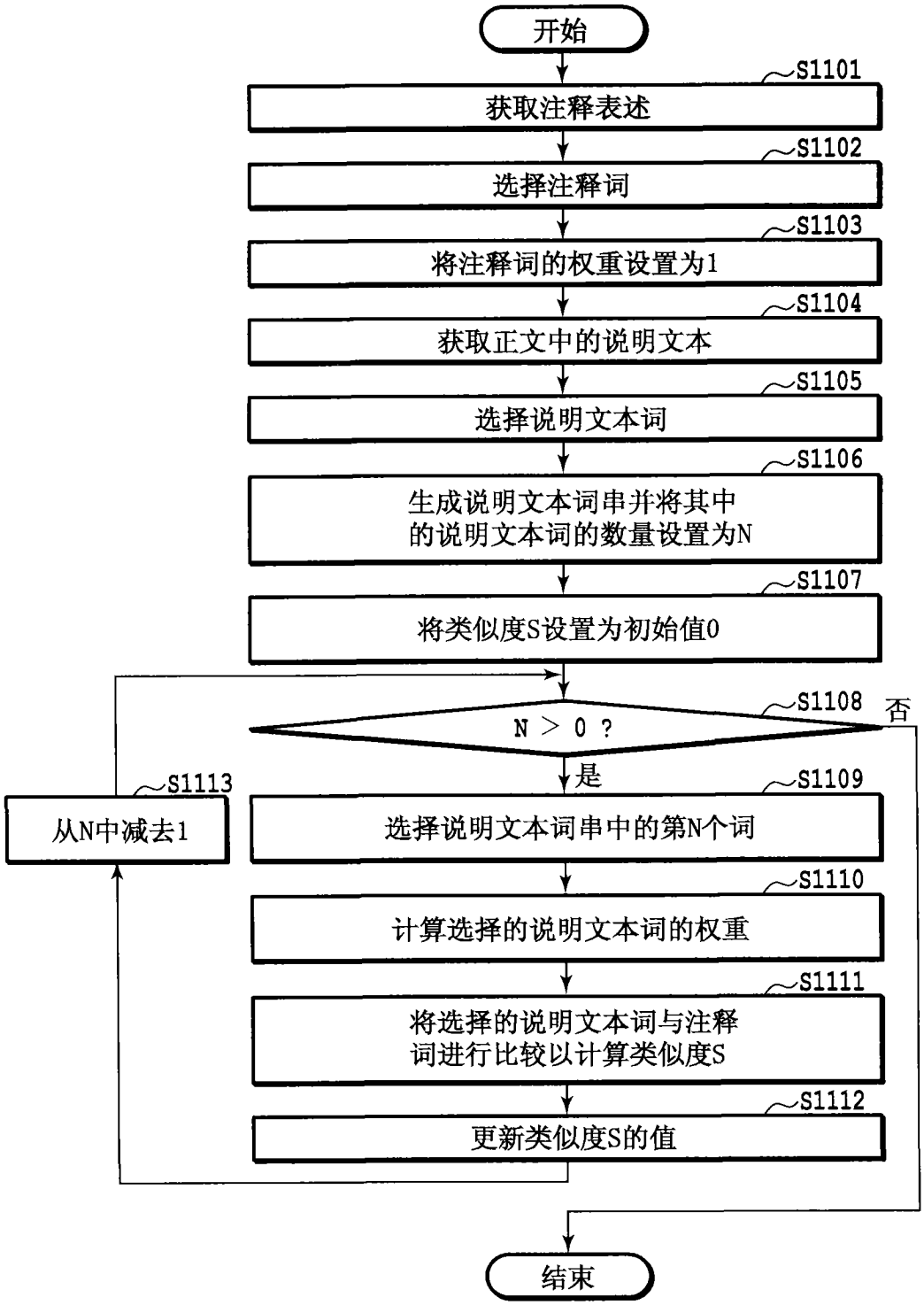

开始
S1101
获取注释表述
S1102
选择注释词
S1103
将注释词的权重设置为1
S1104
获取正文中的说明文本
S1105
选择说明文本词
S1106
生成说明文本词串并将其中
的说明文本词的数量设置为N
S1107
将类似度S设置为初始值0
S1108
N > 0 ?
否
是
S1109
选择说明文本词串中的第N个词
S1110
计算选择的说明文本词的权重
S1111
将选择的说明文本词与注释
词进行比较以计算类似度S
S1112
更新类似度S的值
S1113
从N中减去1
结束

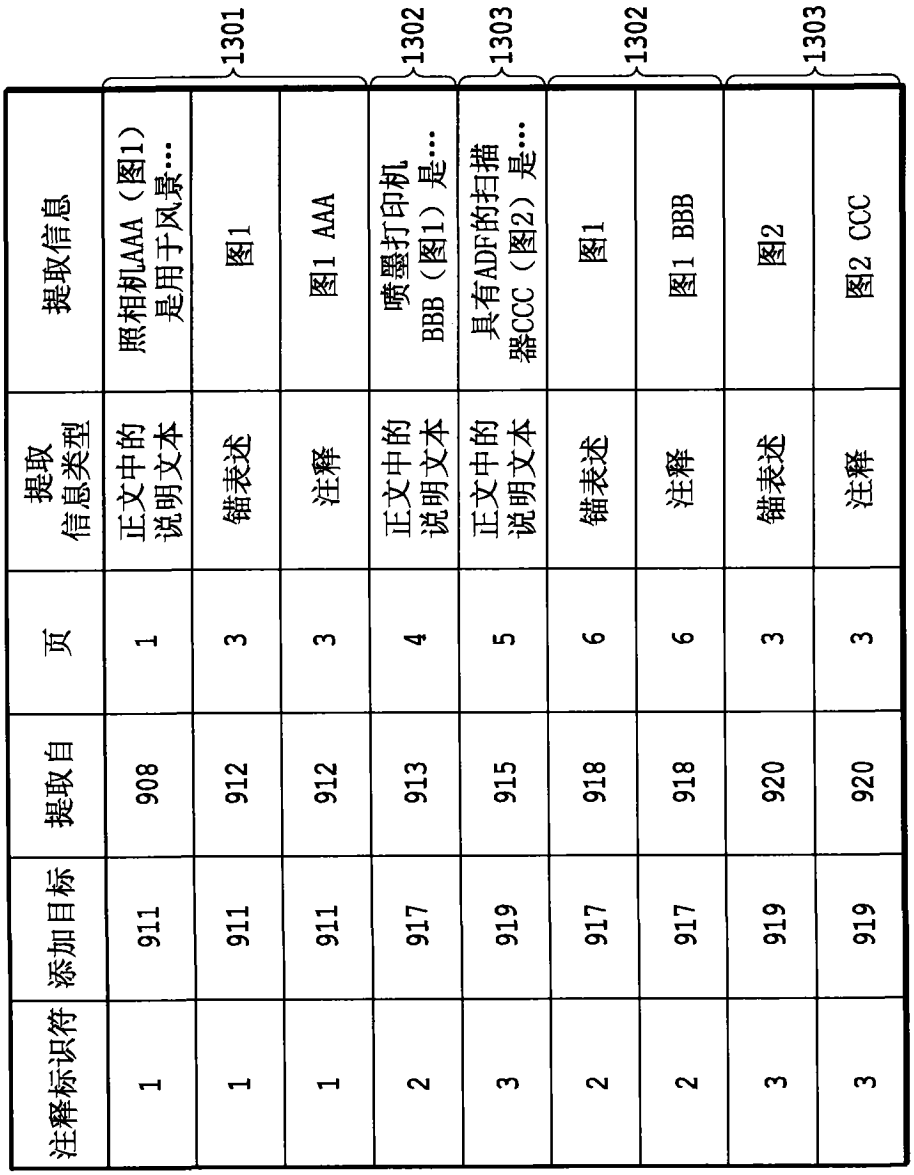
图 11

注释表述	AAA
注释词	AAA
注释词的属性	名词
比较目标候选？	○
词距离	-
权重	1

图 12A

正文中的说明文本	照相机AAA（图1）是用于例如风景的拍摄……。									
说明文本词	照相机	AAA	（	图1	）	是	，	风景	例如	…
说明文本词的属性	名词	名词	符号	（锚表述）	符号	助词	逗号	名词	助词	…
比较目标候选	○	○	x	(x)	x	x	x	○	x	…
说明文本词串的顺序	2	3	-	-	-	-	-	1	-	…
词距离	3	2	1	0	1	2	3	4	5	…
权重	0.33	0.5	-	-	-	-	-	0.25	-	…

图 12B



注释标识符	添加目标	提取自	页	提取信息类型	提取信息	
1	911	908	1	正文中的说明文本	照相机AAA（图1）是用于风景…	1301
1	911	912	3	锚表述	图1	
1	911	912	3	注释	图1 AAA	
2	917	913	4	正文中的说明文本	喷墨打印机BBB（图1）是…	1302
3	919	915	5	正文中的说明文本	具有ADF的扫描器CCC（图2）是…	1303
2	917	918	6	锚表述	图1	1302
2	917	918	6	注释	图1 BBB	
3	919	920	3	锚表述	图2	1303
3	919	920	3	注释	图2 CCC	

图 13

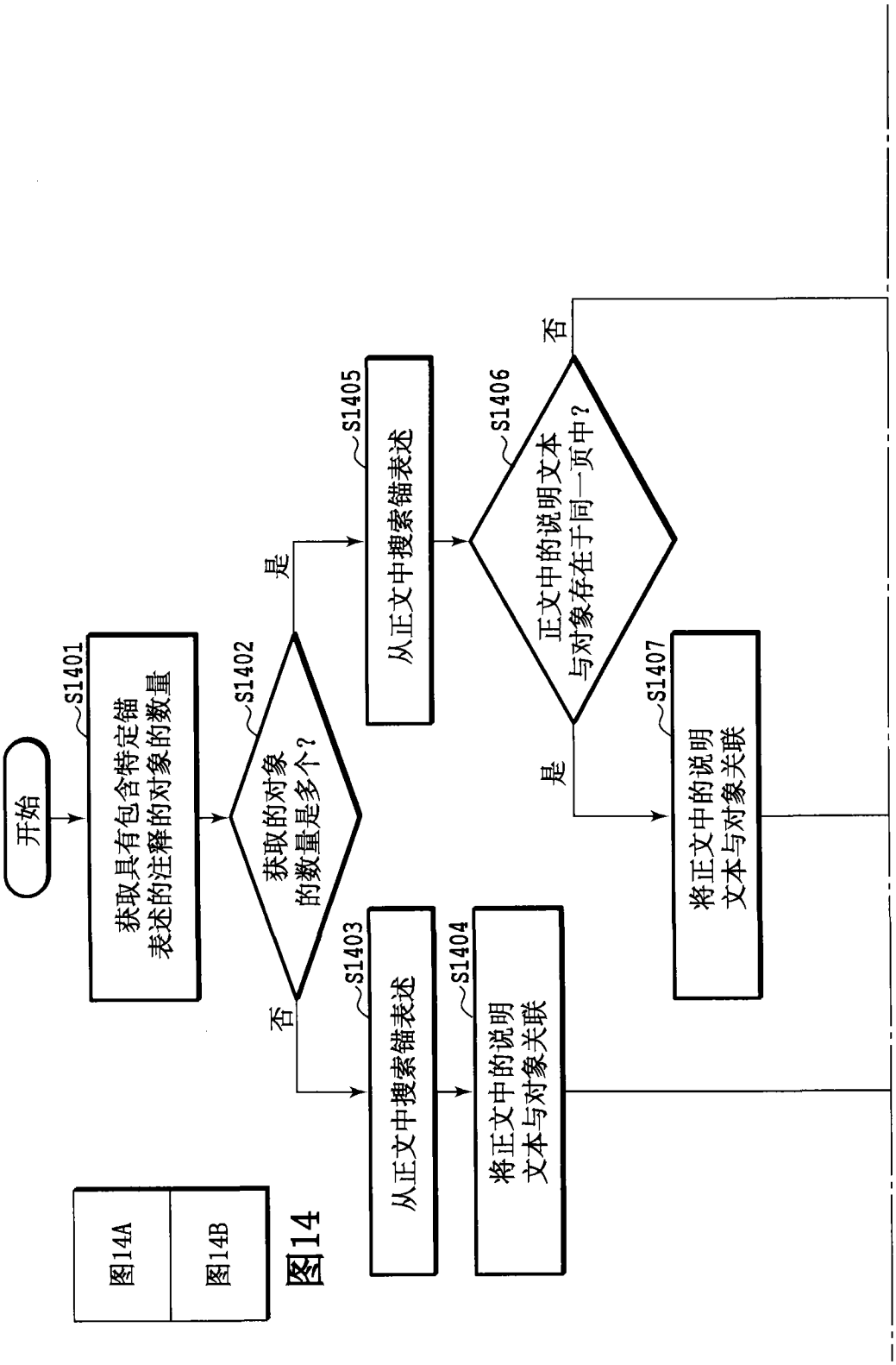

开始
S1401
获取具有包含特定锚表述的对象的数量
S1402
获取的对象的数量是多个？
是
否
S1403
从正文中搜索锚表述
S1404
将正文中的说明文本与对象关联
S1405
从正文中搜索锚表述
S1406
正文中的说明文本与对象存在于同一页中？
是
否
S1407
将正文中的说明文本与对象关联
图14A
图14B

图14

图 14A

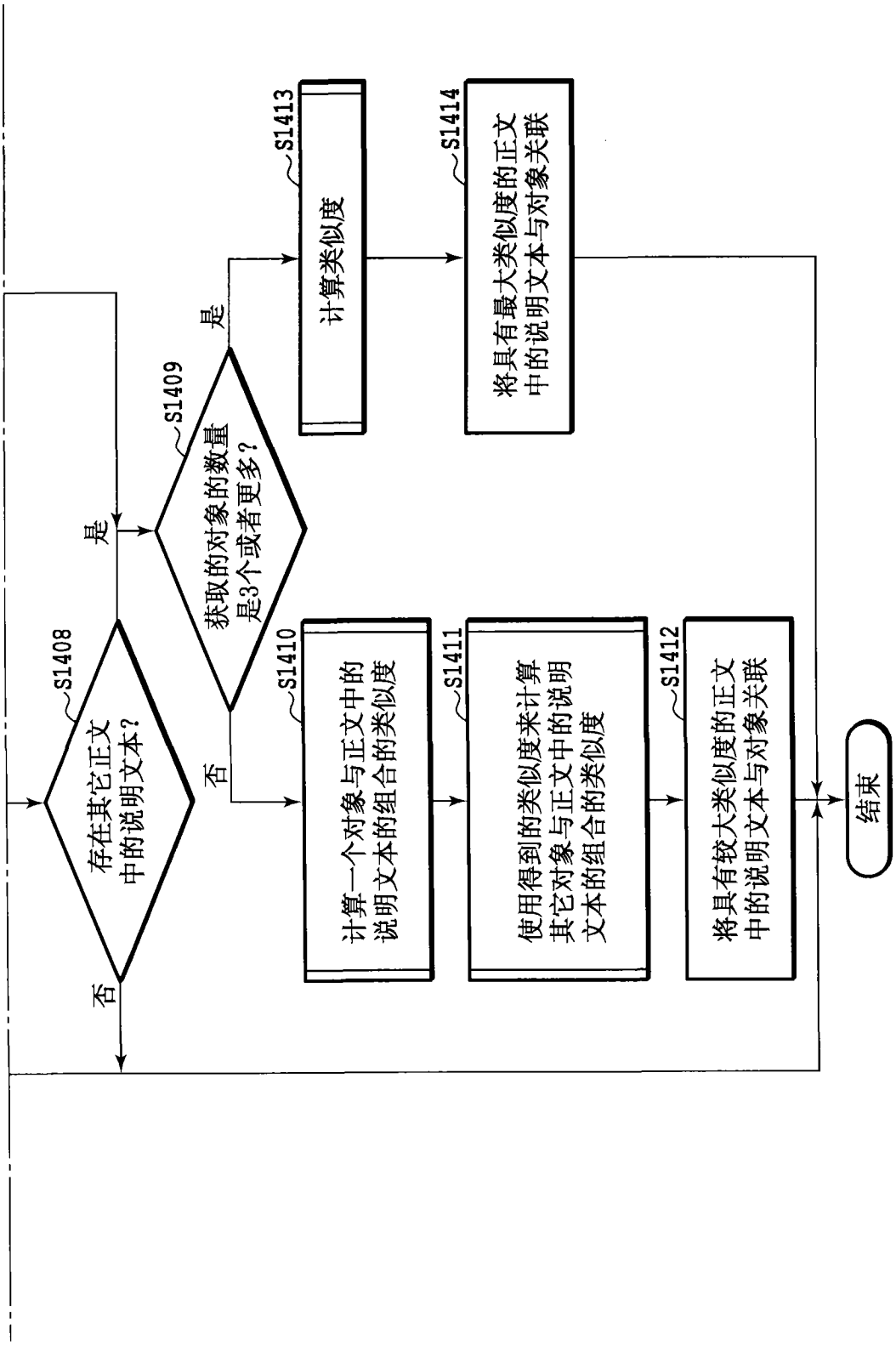

S1408
存在其它正文中的说明文本?
是
否
S1409
获取的对象的数量是3个或者更多?
是
否
S1410
计算一个对象与正文中的说明文本的组合的类似度
S1411
使用得到的类似度来计算其它对象与正文中的说明文本的组合的类似度
S1412
将具有较大类似度的正文中的说明文本与对象关联
S1413
计算类似度
S1414
将具有最大类似度的正文中的说明文本与对象关联
结束

图 14B

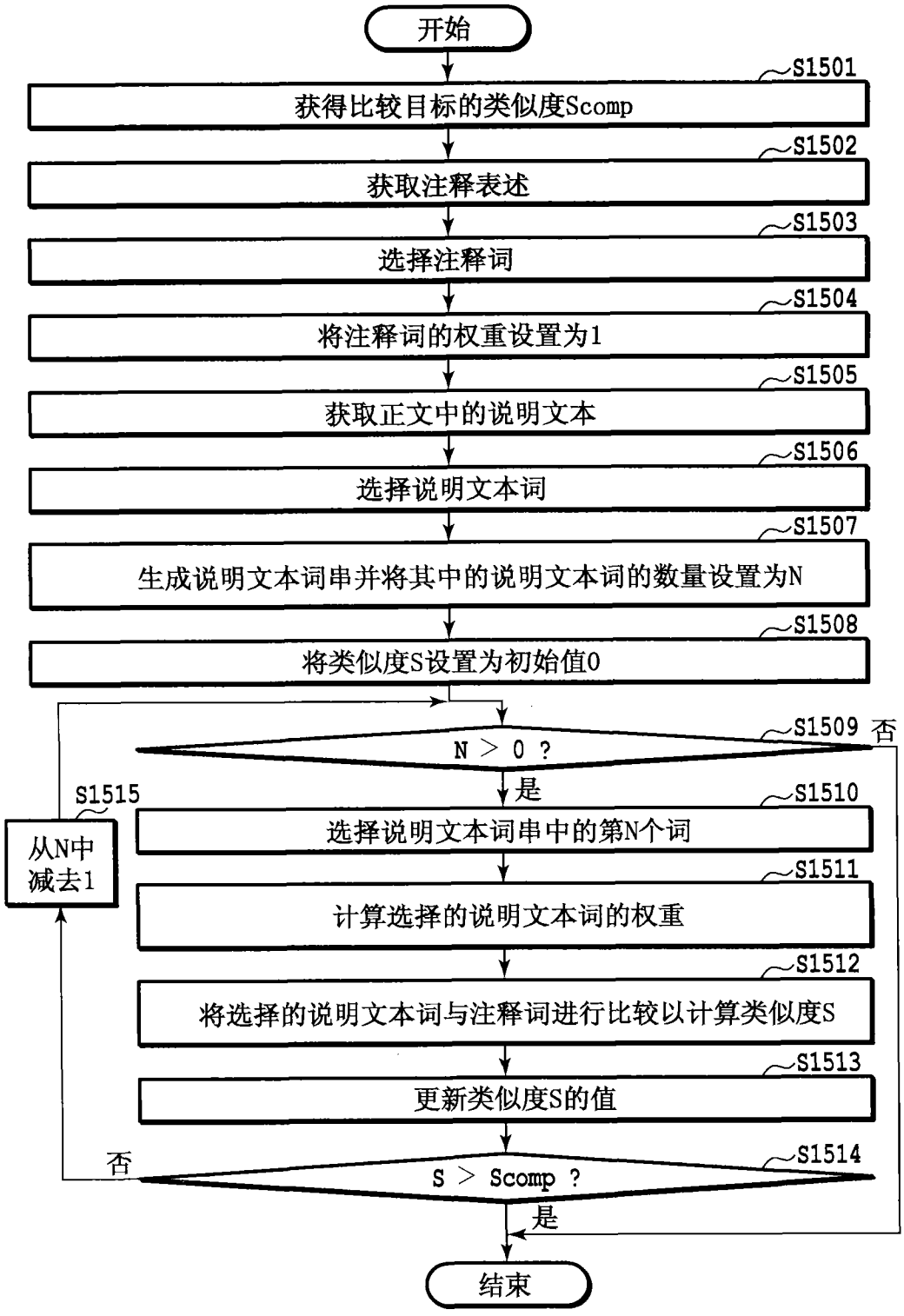

开始
S1501
获得比较目标的类似度Scomp
S1502
获取注释表述
S1503
选择注释词
S1504
将注释词的权重设置为1
S1505
获取正文中的说明文本
S1506
选择说明文本词
S1507
生成说明文本词串并将其中的说明文本词的数量设置为N
S1508
将类似度S设置为初始值0
S1509
N > 0 ?
否
是
S1510
选择说明文本词串中的第N个词
S1511
计算选择的说明文本词的权重
S1512
将选择的说明文本词与注释词进行比较以计算类似度S
S1513
更新类似度S的值
S1514
S > Scomp ?
否
是
S1515
从N中减去1
结束

图 15

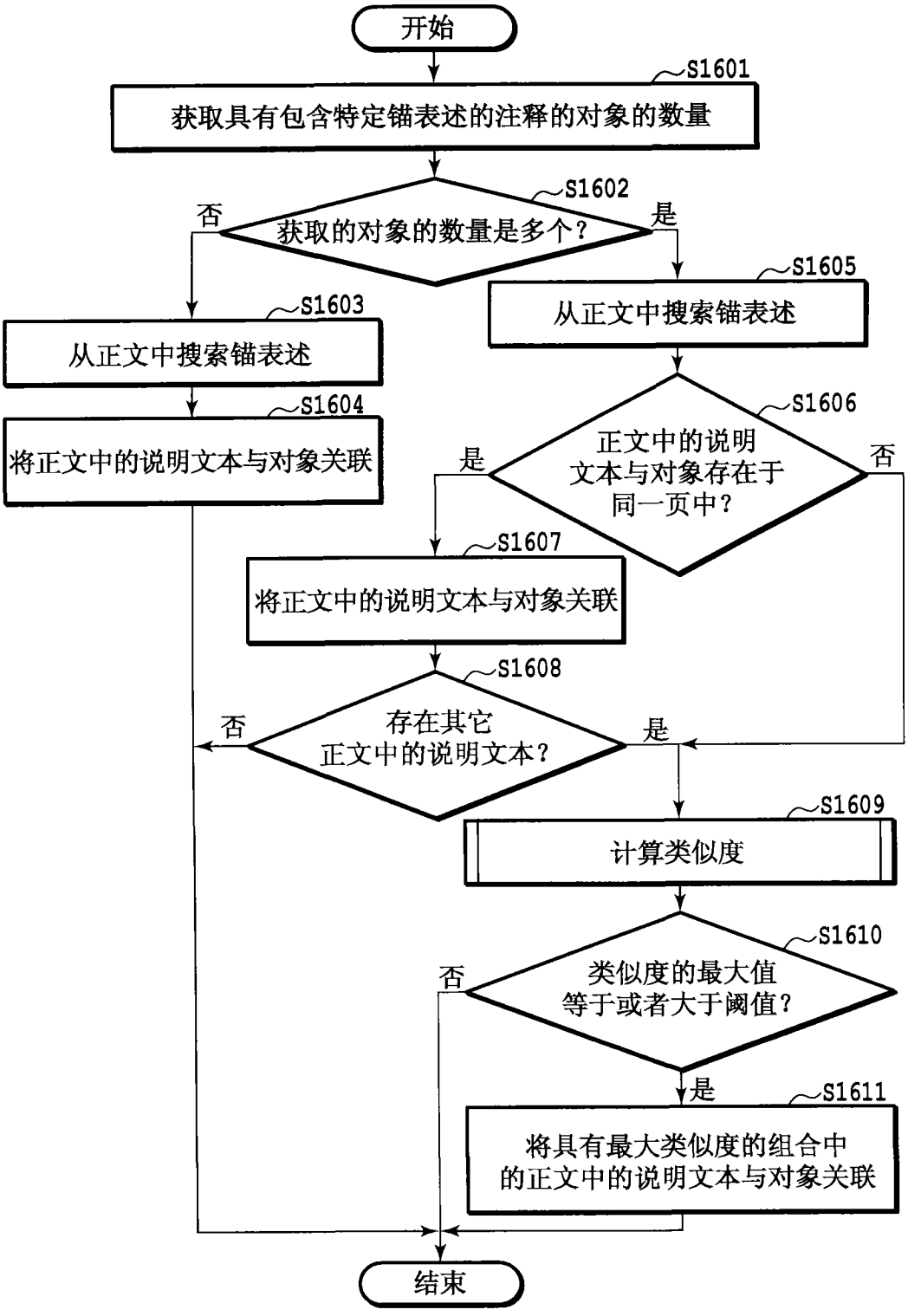

开始
S1601
获取具有包含特定锚表述的注释的对象的数量
S1602
获取的对象的数量是多个？
否
是
S1603
从正文中搜索锚表述
S1604
将正文中的说明文本与对象关联
S1605
从正文中搜索锚表述
S1606
正文中的说明文本与对象存在于同一页中？
是
否
S1607
将正文中的说明文本与对象关联
S1608
存在其它正文中的说明文本？
否
是
S1609
计算类似度
S1610
类似度的最大值等于或者大于阈值？
否
是
S1611
将具有最大类似度的组合中的正文中的说明文本与对象关联
结束

图 16

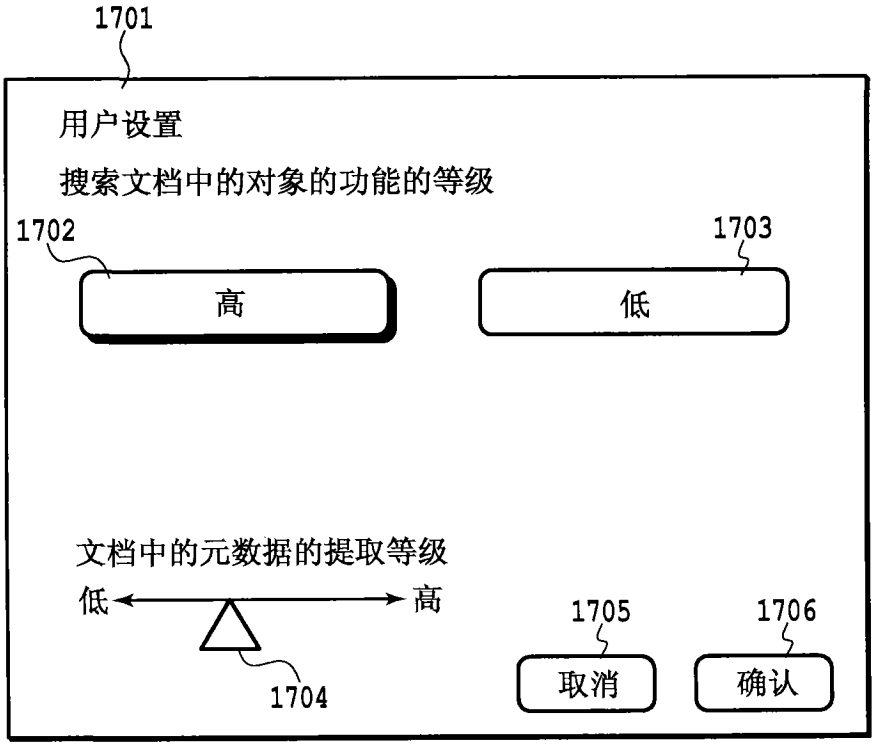

1701
用户设置
搜索文档中的对象的功能的等级
1702
高
1703
低
文档中的元数据的提取等级
低
高
1704
1705
取消
1706
确认

图 17

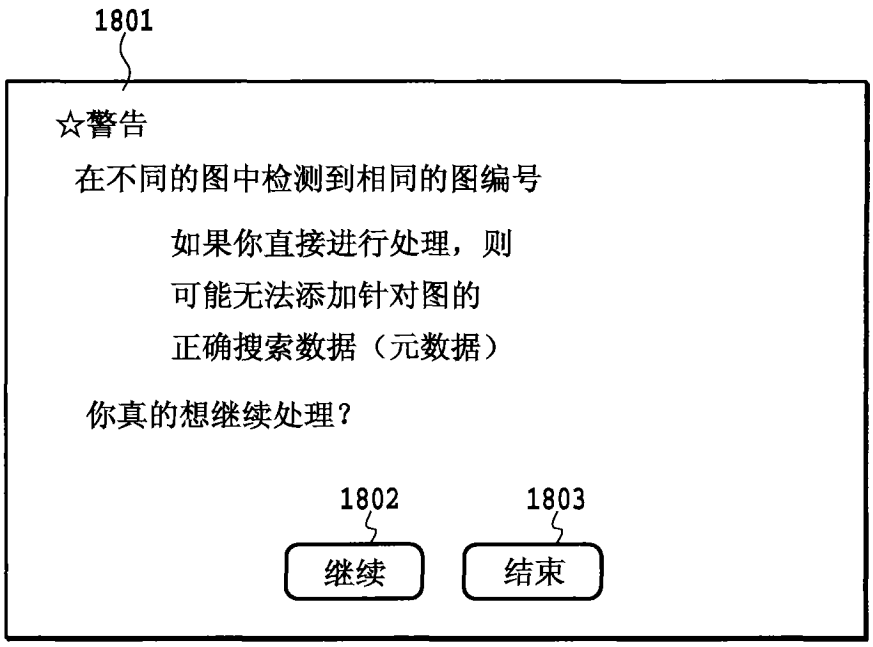

1801
☆警告
在不同的图中检测到相同的图编号
如果你直接进行处理，则
可能无法添加针对图的
正确搜索数据（元数据）
你真的想继续处理？
1802
继续
1803
结束

图 18

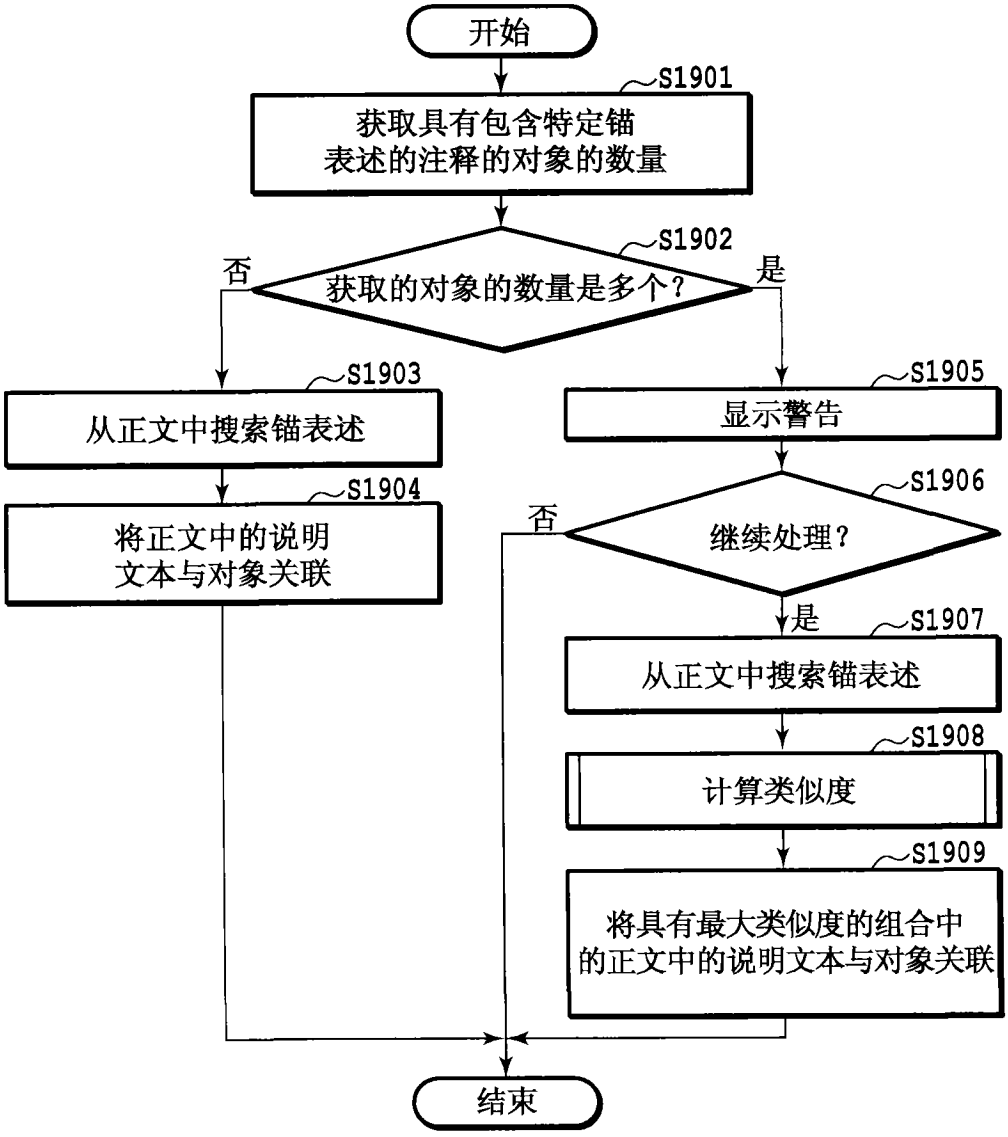

开始
S1901
获取具有包含特定锚
表述的注释的对象的数量
S1902
否
获取的对象的数量是多个？
是
S1903
从正文中搜索锚表述
S1904
将正文中的说明
文本与对象关联
S1905
显示警告
S1906
否
继续处理？
是
S1907
从正文中搜索锚表述
S1908
计算类似度
S1909
将具有最大类似度的组合中
的正文中的说明文本与对象关联
结束

图 19