[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/21 (2006.01)
G06F 17/30 (2006.01)

[12] 发明专利申请公开说明书

[21] 申请号 200510098131.9

[43] 公开日 2006 年 3 月 29 日

[11] 公开号 CN 1752963A

[22] 申请日 2005.9.7
[21] 申请号 200510098131.9
[30] 优先权
[32] 2004. 9.21 [33] JP [31] 2004 - 273511
[71] 申请人 株式会社东芝
地址 日本东京都
[72] 发明人 铃木优 石谷康人

[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 康建忠

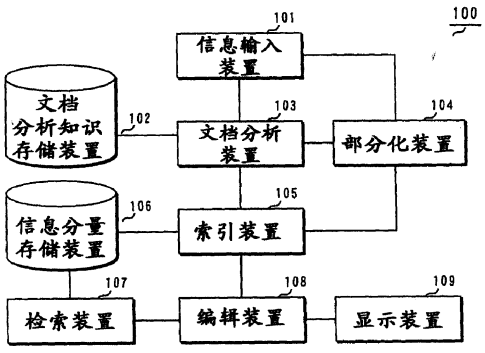
权利要求书 3 页 说明书 32 页 附图 32 页

[54] 发明名称

文档信息处理设备、文档信息处理方法及处理程序

[57] 摘要

提供用于处理文档信息的设备和方法。 根据一个实施例，文档信息处理设备包括用于使用分析文档分析知识来进行对从文档信息输入装置输入的文档信息的文档分析的文档分析装置；用于将该从文档信息输入装置输入的文档信息分割成信息分量以作为编辑单元的部分化装置；用于基于文档分析的结果，为该信息分量生成索引信息并将该索引信息分配给该信息分量的索引装置；和用于关联地存储该信息分量和分配给该信息分量的索引信息的信息分量存储装置。 该设备还可以包括用于检索该信息分量的信息分量检索装置。



1. 一种文档信息处理设备，包括：

文档信息输入装置，用于输入文档信息；

文档分析装置，用于通过使用分析该文档信息的分析知识来进行对该文档信息的文档分析；

部分化装置，用于将该文档信息分割成信息分量以作为编辑单元；

索引装置，用于基于文档分析的结果，为该信息分量生成索引信息并将该索引信息分配给该信息分量；和

信息分量存储装置，用于关联地存储该信息分量和分配给该信息分量的索引信息。

2. 一种文档信息处理设备，包括：

文档信息输入装置，用于输入文档信息；

文档分析装置，用于通过使用分析该文档信息的分析知识来进行对该文档信息的文档分析；

部分化装置，用于将该文档信息分割成信息分量以作为编辑单元；

信息分量选择装置，用于允许用户选择该信息分量；

索引装置，用于基于用户选择的结果，为该信息分量生成索引信息并将该索引信息分配给该信息分量；和

信息分量存储装置，用于关联地存储该信息分量和分配给该信息分量的索引信息。

3. 如权利要求 1 或 2 所述的文档信息处理设备，进一步包括信息分量检索装置，用于从该信息分量存储装置中检索信息分量。

4. 如权利要求 1 或 2 所述的文档信息处理设备，其中该文档分析装置对从下述组中选择的至少一个进行文档分析，该组包括（1）该文档信息的文档结构，（2）包含在该文档信息中的部分的功能性作用，

和（3）包含在该文档信息中的任何单词、从句和句子的语义属性。

5. 如权利要求 1 或 2 所述的文档信息处理设备，其中该文档分析装置通过使用语义分析知识进行对该文档信息的语义分析。

6. 如权利要求 1 或 2 所述的文档信息处理设备，其中该部分化装置基于该文档分析结果，将该文档信息分割成信息分量。

7. 如权利要求 1 或 2 所述的文档信息处理设备，进一步包括:

编辑模板存储装置，用于存储用于编辑该信息分量的编辑模板；和

编辑装置，用于基于该编辑模板、文档分析结果和部分化装置的分割结果中至少一个，对该信息分量进行编辑，以生成新的文档信息。

8. 如权利要求 7 所述的文档信息处理设备，进一步包括编辑模板生成装置，用于基于该文档分析结果和编辑装置编辑的内容来生成编辑模板。

9. 如权利要求 8 所述的文档信息处理设备，进一步包括控制装置，用于将编辑模板生成装置生成的编辑模板存储在编辑模板存储装置中。

10. 如权利要求 1 或 2 所述的文档信息处理设备，进一步包括文档分析知识存储装置，用于存储文档分析结果的结果。

11. 一种文档信息处理方法，包括以下步骤:

输入文档信息；

通过使用分析该文档信息的分析知识来进行对该文档信息的文档分析；

将该文档信息分割成信息分量以作为编辑单元；

基于文档分析的结果，为该信息分量生成索引信息并将该索引信息分配给该信息分量；和

关联地存储该信息分量和分配给该信息分量的索引信息，作为信息分量存储装置中的组。

12. 一种文档信息处理方法，包括以下步骤:

输入文档信息；

通过使用分析该文档信息的分析知识来进行对该文档信息的文档分析;

将该文档信息分割成信息分量以作为编辑单元;

允许用户选择该分割的信息分量;

基于用户选择的结果，为该信息分量生成索引信息并将该索引信息分配给该信息分量；和

关联地存储该信息分量和分配给该信息分量的索引信息，作为信息分量存储装置中的组。

13. 一种计算机可读介质，包含用于执行一种文档信息处理方法的指令，该方法包括:

输入文档信息;

通过使用分析该文档信息的分析知识来进行对该文档信息的文档分析;

将该文档信息分割成信息分量以作为编辑单元;

基于文档分析的结果，为该信息分量生成索引信息并将该索引信息分配给该信息分量；和

关联地存储该信息分量和分配给该信息分量的索引信息，作为信息分量存储装置中的组。

14. 一种计算机可读介质，包含用于执行一种文档信息处理方法的指令，该方法包括:

输入文档信息;

通过使用分析该文档信息的分析知识来进行对该文档信息的文档分析;

将该文档信息分割成信息分量以作为编辑单元;

允许用户选择该分割的信息分量;

基于用户选择的结果，为该信息分量生成索引信息并将该索引信息分配给该信息分量；和

关联地存储该信息分量和分配给该信息分量的索引信息，作为信息分量存储装置中的组。

文档信息处理设备、文档信息处理方法及处理程序

技术领域

本发明涉及一种文档信息处理设备、文本处理信息方法和文档信息处理程序，用于检索／编辑因特网内容、电子邮件等的电子信息、或者通过光学字符阅读器（OCR）或类似技术从打印介质例如纸中提取的电子信息。具体地说，涉及一种文档信息处理设备，其支持或自动执行将电子信息转换成多个部分的操作、检索／获取该部分信息的操作、或者编辑该获取部分和产生新内容的操作。

背景技术

随着因特网的日益普及和数字照相机、扫描仪等的性能增强和广泛使用，一般用户已经开始在商务／家庭应用中从个人计算机上浏览多种类和大量的信息条目。因而就增加了将用户判断为有用的浏览信息条目的那些信息条目或者一些信息条目保存为片断的需求。

作为服从这种需求的一种现有技术，能够直接剪贴（scrap）被浏览的内容的应用软件例如“OneNote (TM)”（由 Microsoft Corporation 制造）或者“kami-copi (TM)”（由 YMIRLINK Inc.制造）已有市售。已经提出了一种用于编辑已经形成组成结构的结构化文档的方法（例如称为专利文档 1），一种用于可编程地模板化在用于医疗应用的成像系统中被浏览的信息条目的排列的方法（例如称为专利文档 2）等等。

专利文档 1：美国专利申请公开 2004/0010755

专利文档 2：美国专利 5,961,610

然而，根据这些现有技术，不能对一个片断的每个部分给出语义或句法信息（例如用以初始化剪贴的信息格式（称为“源信息”），该源信

息中的分量的功能性作用，或者包含在该分量中的个体元素的语义属性）。因此不能增加该剪贴操作的高效性或者由该剪贴操作产生的内容（下文中指“剪贴页（scrap pages）”）的重复使用。更具体地，在根据为某种目的收集的剪贴页而不需要大量劳动就从相同格式的源信息中获取相同功能的片断的情况，或者在剪贴的信息条目已经被安排成某种格式的剪贴页的情况下，存在不能满足其后产生相同格式的剪贴页的需要的问题。

发明内容

本发明的目的是提供一种能够准确获得必要信息的文档信息处理设备。

与本发明一致地，提供一种文档信息处理设备，包括：用于输入文档信息的文档信息输入装置；通过使用用于分析该文档信息的分析知识来进行对该文档信息的文档分析的文档分析装置；用于将文档信息分成作为编辑单元的信息分量的部分化装置；基于文档分析的结果为信息分量生成索引信息和为信息分量分配索引信息的索引装置；和用于相关联地存储信息分量和分配给信息分量的索引信息的信息分量存储装置。

与本发明一致地，还提供一种文档信息处理设备，包括：用于输入文档信息的文档信息输入装置；通过使用用于分析该文档信息的分析知识来进行对该文档信息的文档分析的文档分析装置；用于将文档信息分成作为编辑单元的信息分量的部分化装置；用于允许用户选择信息分量的信息分量选择装置；基于用户选择的结果为信息分量生成索引信息和为信息分量分配索引信息的索引装置；和用于相关联地存储信息分量和分配给信息分量的索引信息的信息分量存储装置。

与本发明一致地，进一步提供一种文档信息处理方法，包括：输入文档信息；通过使用用于分析该文档信息的分析知识来进行对该文档信息的文档分析；将输入的文档信息分成作为编辑单元的信息分量；基于文档分析的结果为信息分量生成索引信息和为信息分量分配索引信息；和相关联地存储信息分量和分配给信息分量的索引信息，作为

在信息分量存储装置的组（set）。

与本发明一致地，此外还提供一种文档信息处理方法，包括：输入文档信息；通过使用用于分析该文档信息的分析知识来进行对该文档信息的文档分析；将输入的文档信息划分成作为编辑单元的信息分量；允许用户选择划分的信息分量；基于用户选择的结果为信息分量生成索引信息和为信息分量分配索引信息；和相关联地存储信息分量和分配给信息分量的索引信息，作为在信息分量存储装置的组。

与本发明一致地，进一步还提供一种计算机可读介质，包含用于执行处理文档信息的方法的指令，该方法包括：输入文档信息；通过使用用于分析该文档信息的分析知识来进行对该文档信息的文档分析；将输入的文档信息分成作为编辑单元的信息分量；基于文档分析的结果为信息分量生成索引信息和为信息分量分配索引信息；以及相关联地存储信息分量和分配给信息分量的索引信息，作为在信息分量存储装置的组。

与本发明一致地，还提供一种计算机可读介质，包含用于执行处理文档信息的方法的指令，该方法包括：输入文档信息；通过使用用于分析该文档信息的分析知识来进行对该文档信息的文档分析；将输入的文档信息分成作为编辑单元的信息分量；允许用户选择该划分的信息分量；基于用户选择的结果为信息分量生成索引信息和为信息分量分配索引信息；以及相关联地存储信息分量和分配给信息分量的索引信息，作为在信息分量存储装置的组。

根据本发明的实施例，能够提供一种可以基于文档数据的上下文执行适当的索引的文档信息处理设备。

附图说明

图 1 是根据本发明的第一实施例的示例性文档信息处理设备的框图；

图 2A－2D 是显示了输入到信息输入装置的信息条目的示例的示意图；

图 3A－3C 是显示了输入到信息输入装置的信息条目来源的示例

的示意图；

图 4 是用于解释文档分析装置的处理流程的流程图；

图 5A 和 5B 是分别显示了涉及文档结构分析的知识的示例的示意图；

图 6 是用于解释在输入以 HTML 描述的信息的情况下的文档结构分析处理的流程图；

图 7A－7D 是分别显示了由文档分析装置进行的文档结构分析处理的结果的示例的示意图；

图 8A 是显示了由文档分析装置进行的语义属性分析处理的结果的示例的示意图（在输入图 3A 中的信息的情况下的输出示例）；

图 8B 是显示了由文档分析装置进行的语义属性分析处理的结果的示例的示意图（在输入图 3B 中的信息的情况下的输出示例）；

图 8C 是显示了由文档分析装置进行的语义属性分析处理的结果的示例的示意图（在输入图 3C 中的信息的情况下的输出示例）；

图 8D 是显示了由文档分析装置进行的语义属性分析处理的结果的示例的示意图（在输入图 2D 中的信息的情况下的输出示例）；

图 9 是用于解释由文档分析装置进行的功能性作用分析处理的流程图；

图 10 是显示功能性作用分析知识的示例的示意图；

图 11A 是显示了对图 8A 中的文档数据进行功能性作用分析处理的处理结果的示例的示意图；

图 11B 是显示了对图 8B 中的文档数据进行功能性作用分析处理的处理结果的示例的示意图；

图 11C 是显示了对图 8C 中的文档数据进行功能性作用分析处理的处理结果的示例的示意图；

图 11D 是显示了对图 8D 中的文档数据进行功能性作用分析处理的处理结果的示例的示意图；

图 12 是用于解释部分化装置的处理流程的流程图；

图 13A 是显示在输入图 11A 中文档数据的情况下部分化装置的

处理结果的示例的示意图；

图 13B 是显示在输入图 11B 中文档数据的情况下部分化装置的处理结果的示例的示意图；

图 13C 是显示在输入图 11C 中文档数据的情况下部分化装置的处理结果的示例的示意图；

图 13D 是显示在输入图 11D 中文档数据的情况下部分化装置的处理结果的示例的示意图；

图 14 是用于解释索引装置的处理流程的流程图；

图 15 是显示索引装置的结构的示意图；

图 16 是显示信息分量存储装置的结构的示意图；

图 17A 和 17B 是显示索引策略知识的示例的示意图；

图 18 是用于解释检索装置的处理流程的流程图；

图 19 是显示检索装置的结构的示意图；

图 20 是显示检索策略知识的示例的示意图；

图 21 是显示根据第二实施例的文档信息处理设备的结构的示意图；

图 22 是显示使用编辑装置的编辑工作的屏幕的示例的示意图；

图 23A 和 23B 是显示剪贴簿的数据表示的示例的示意图；

图 24 是用于解释模板生成装置的操作的流程图；

图 25 是显示由模板生成装置从图 23B 转换的模板的示例的示意图；

图 26 是用于解释在编辑装置基于模板来实现编辑处理的情况下的处理流程的流程图；

图 27A 和 27B 是显示一组文档的示意图；

图 28A 和 28B 是显示在图 25 中表示的部分都被替换的情况下的编辑结果的示意图；和

图 29 显示了描述可以实施与本发明一致的系统和方法的示例性硬件结构的示意图。

具体实施方式

下面将参照附图对本发明的实施例进行说明。

（第一实施例）

第一实施例包括一种文档信息处理设备，能够把用户在 PC 上浏览的内容分割和部分化，例如因特网或电子邮件的内容，或者通过使用扫描仪和 OCR 转换成电子文本的纸介质内容，并且允许用户按照需要检索和编辑该部分化信息。

图 1 是显示根据本发明第一实施例的示例性文档信息处理设备的框图。

参照图 1，文档信息处理设备 100 包括信息输入装置 101、文档分析知识存储装置 102、文档分析装置 103、部分化（componentization）装置 104、索引装置 105、信息分量(component)存储装置 106 和检索装置 107。

信息输入装置 101 读出被用户浏览的信息，作为文档信息处理设备 100 的输入。在第一实施例中，被提取的信息可以是因特网、电子邮件、以及印刷在纸张等上的信息通过由扫描仪装载并由现有的 OCR（光学字符阅读器）技术来转换的方式来获取的电子信息的内容。更具体地，信息输入装置 101 与用户浏览这些信息条目所使用的应用软件通信，从而提取该信息。作为信息提取器的应用软件可以是专门为本实施例建立的程序或者其他现有的应用软件。在现有应用软件的情况下，该信息通过在现有应用软件产品之间的通信技术来提取。

文档分析知识存储装置 102 存储用于分析输入到信息输入装置 101 的文档信息的文档分析知识。举例来说，用于该文档信息的语义分析的语义分析知识被存储作为文档分析知识。

文档分析装置 103 基于存储在文档分析知识存储装置 102 中的文档分析知识，分析输入到信息输入装置 101 的文档信息。该分析例如可以是语义分析。

部分化装置 104 基于文档分析装置 103 的文档分析结果，将输入到信息输入装置 101 的信息分割和部分化。下面将通过对该信息分割和部分化得到的各项称为“信息分量（component）”。

索引装置 105 基于文档分析装置 103 的文档分析结果，产生和为部分化装置 104 分割的单个信息分量分配索引，并且将得到的信息分量存储在信息分量存储装置 106 中。

信息分量存储装置 106 存储被索引装置 105 分配有索引的信息分量。

检索装置 107 基于该索引检索存储在信息分量存储装置 106 中的信息分量。

编辑装置 108 通过利用由检索装置 107 检索得到的至少一个信息分量来编辑新内容。该由编辑装置 108 编辑的内容被发送到索引装置 105，并且被作为新信息分量分配索引和存储在信息分量存储装置 106 中。

基于编辑装置 108 的编辑屏幕在显示装置 109 例如 CRT（阴极射线管）显示器或液晶显示器（LCD）上显示。

现在，将使用样本信息对文档信息处理设备 100 的操作进行说明。

图 2A－2D 是显示了输入到信息输入装置 101 的信息条目的示例的示意图。

图 2A－2D 中的所有示例都是 TSB 公司的产品“GBG21”上的信息条目。

图 2A 显示了 TSB 公司的产品的新闻稿的网页内容（以 HTML（超文本标记语言）格式编写的数据），图 2B 显示了在因特网上的新站点中出现的产品介绍报告的网页内容（HTML），图 2C 显示了来自一个商店的电子邮件的直接邮件（具有邮件头的文本），图 2D 显示了目录（通过扫描仪加载的、打印在纸介质上的目录数据）。

图 2A 和 2B 中所示的电子信息条目被从因特网的网页浏览器输入到信息输入装置 101。图 2C 所示的电子信息被从电子邮件应用输入到信息输入装置 101。图 2D 所示的电子信息被从图像扫描数据的浏览器输入到信息输入装置 101。

在与本发明一致的实施例中，文档信息处理设备 100 被实施为应用软件，其中网页浏览器和电子邮件应用软件的功能被作为软件部分

结合，该信息输入装置 101 可以通过该软件部分的应用编程接口（API）接收信息条目的输入。在与本发明一致的另一实施例中，文档信息处理设备 100 实施为与外部软件（例如网页浏览器、电子邮件应用软件等）协同操作的应用软件，信息输入装置 101 通过基于该外部软件和应用软件之间的通信技术的通信来接收信息的输入。

图 2A 和 2B 例示了通过网页浏览器浏览信息条目的情况，并且实际输入到信息输入装置 101 的信息条目来源的示例分别在图 3A 和 3B 中示出。同样地，图 2C 例示了通过电子邮件应用软件浏览信息的情况，而实际输入到信息输入装置 101 的信息来源的示例在图 3C 中示出。图 2D 例示了通过图像扫描数据浏览器浏览信息的情况，并且该信息是以图像数据格式例如标记图像文件格式（TIFF）的二进制数据输入到信息输入装置 101。

信息输入装置 101 将该信息的输入源的类型或标识符作为属性信息附加到该输入信息，并且将所得的信息发送给文档分析装置 103。该“作为属性信息附加的、该信息的输入源的类型或标识符”是用于识别网页浏览器或电子邮件应用软件或者具有能够与信息输入装置 101 通信以接收该信息输入的功能的软件部分的属性信息。

这里，通过示例假定网页浏览器或其软件部分的标识符是“INTERNET”。并且，电子邮件应用软件或其软件部分的标识符假定为“MAIL”。另外，图像扫描数据或其软件部分的标识符假定为“SCAN”。

文档分析装置 103 对输入信息的文档结构、包含在输入信息中的部分的功能性作用(functional role)、单词的语义属性、包含在输入信息中的从句或句子进行文档分析。该文档分析装置 103 的处理将结合图 4 进行说明。

然后，将参照图 4 的流程图对文档分析装置 103 的处理流程进行说明。

参照图 4，文档分析装置 103 根据从信息输入装置 101 输入的属性信息改变对文档结构的分析处理（步骤 S401、步骤 S404 或步骤

S406）。

文档分析装置103判断从信息输入装置101输入的属性信息是否是“SCAN”（步骤S401）。

在步骤S401的判断是“是”的情况，该输入信息是图像扫描数据。因此，文档分析装置103首先执行OCR处理以将图像扫描数据转换成文本（步骤S402），然后将该文本提交到文档结构分析处理（a）（步骤S403）。

利用已知的技术（例如JP-A-2003-288334）能够对该图像扫描数据进行OCR处理和进行文档结构分析处理（a），这里省略了对它们的详细说明。

另一方面，在步骤S401的判断是“否”的情况，文档分析装置103判断从信息输入装置101输入的属性信息是否是“INTERNET”（步骤S404）。

在步骤S404的判断是“是”的情况，该输入信息是用HTML描述的。因此，文档分析装置103执行文档结构分析处理（b），其中考虑HTML的结构（S405）。文档结构分析处理（b）的细节将在以后说明。

另一方面，在步骤S404的判断是“否”的情况，文档分析装置103判断从信息输入装置101输入的属性信息是否是“MAIL”（步骤S406）。

在步骤S406的判断是“是”的情况，认为该输入信息具有电子邮件头(header)。因此文档分析装置103执行文档结构分析处理（c），其中考虑电子邮件头（步骤S407）。文档结构分析处理（c）的细节将在以后说明。

在步骤S406的判断是“否”的情况，就是说，从信息输入装置101输入的属性信息不是标识符“SCAN”、“INTERNET”和“MAIL”中的任一个（步骤S401、S404和S406的判断是“否”），则文档分析装置103执行文档结构分析处理（d），假定该输入信息是用纯文本描述的。

虽然在本示例中仅有标识符“SCAN”、“INTERNET”和“MAIL”被假定为属性信息的情况，但是对于其他标识符也可以执行类似的处

理。

在步骤 S403 的文档结构分析处理（a）、步骤 S405 的文档结构分析处理（b）、步骤 S407 的文档结构分析处理（c）或步骤 S408 的文档结构分析处理（d）之后，文档分析装置 103 执行语义属性分析处理（步骤 S409），进一步执行功能性作用分析处理（步骤 S410），最后分配从信息输入装置 101 发送的属性信息（步骤 S411），从而输出语义分析结果。

虽然图 4 中的处理是按照文档结构分析处理（步骤 S403、S405、S407 或 S408）、语义属性分析处理（步骤 S409）和功能性作用分析处理（步骤 S410）的顺序进行的，这些处理的顺序不需要限制到本发明的任何实施例。并且，如果需要的话，可以选择执行这些处理中的至少一个。

下面将对文档分析装置 103 进行的文档结构分析处理（b）-（d）的处理内容进行说明。

为了进行文档结构分析处理（b）-（d）的分析，文档分析装置 103 参考存储在文档分析知识存储装置 102 中的文档分析知识中关于文档结构分析的知识条目（item）。

有关文档结构分析的知识条目的示例在图 5A 和 5B 中示出。

图 5A 例示了用于分析 HTML 文档结构的知识。

图 5B 例示了用于分析电子邮件或纯文本的文档结构的知识。该用于分析电子邮件和纯文本的文档结构的知识并不需要总是相同的。

在本实施例中，文档结构分析处理（b）（或（c））和（d）之间的区别通过参考相互不同的文档分析知识条目来体现。也就是说，文档结构分析处理（b）-（d）根据图 6 所示的公共处理流程分别参考图 5A 和 5B 中的知识条目。

[文档结构分析处理（b）的操作]

首先，将参照图 6 对输入如图 3A 所示以 HTML 描述的信息的情况下文档结构分析处理（b）的操作进行说明。

图 3A 中的信息是以 HTML 描述的，并且分析处理（b）参考图

5A 中的知识。

文档分析装置 103 将图 3A 中的文档信息作为待分析数据加载，并且将该加载的信息赋予变量 D（步骤 S601）。

接着，文档分析装置 103 将表示模式匹配位置（来自文档头的字符的位置包含换行符）的变量 I 清零。

随后，文档分析装置 103 从存储在文档分析知识存储装置 102 中的文档结构分析知识中取出一个分析知识条目（步骤 S603）。这里假定在图 5A 例示的分析知识条目 501 已经被取出。

为了以后执行替换处理，文档分析装置 103 将在步骤 S603 取出的分析知识 501 中作为“文件结构标志”的“<STRUCTURE:TITLE>$1</STRUCTURE:TITLE>”赋予变量 T。

关于存储在变量 D 中的待分析数据，文档分析装置 103 从变量 I 指示的位置中搜索与分析知识 501 的“模式”相匹配的位置（步骤 S605）。

在本实施例中，采用在已知技术中使用的称为“Perl 语言”的正规表示格式作为模式。Perl 语言和该语言的正规表示可以从例如“Learning Perl, 2nd Edition”，Randal L. Schwartz & Tom Christiansen (O'Reilly 1997)中得知，在此引用该参考文献的全文作为参考。

在图 5A 中的分析知识 501 的模式的情况下，待分析数据在字符串“<TITLE>”和“</TITLE>”之间存在至少 0 个字符（*）中的任何字符（.）的情况下匹配。这里，该换行符也包含在任何字符（.）中。并且，在字符串“</TITLE>”在输入信息中出现多次的情况，这里将选择最短的一个匹配字符串。最后，在句子中首次出现的该“<TITLE>-</TITLE>”部分被选择。

文档分析装置 103 判断与该模式匹配的字符串是否被找到作为步骤 S605 搜索的结果（步骤 S606）。

在步骤 606 的判断是“是”的情况，文档分析装置 103 用对应于该模式中的括号的字符串替换“变量 T 中的$n (n = 1, 2, ...)”。在至少有

两个括号对应于变量 T 中的至少两个“n”的情况，使用图 3A 中的文档数据作为示例，第三行中的“<TITLE>PRESS RELEASE</TITLE>”与该模式匹配，并且字符串“PRESS RELEASE”对应于该模式中的括号，从而变量 T 的值变为“<STRUCTURE:TITLE>PRESS RELEASE</STRUCTURE:TITLE>”。表示这时的位置的变量 I 的值是“15”，包含换行符。换句话说，从头数起的第 15 个字符，即“<HTML>[换行符]<HEAD>[换行符]”（“[换行符]”实际上是一个字符）之后紧接的字符，与该模式匹配。

另一方面，在步骤 S606 的判断是“否”的情况，文档分析装置 103 进行到步骤 S611。

然后到步骤 S607，文档分析装置 103 用变量 T 的值“<STRUCTURE:TITLE>PRESS RELEASE</STRUCTURE:TITLE>”替换变量 D 中的字符串“<TITLE>PRESS RELEASE</TITLE>”（步骤 S608）。

文档分析装置 103 将表示该位置的变量 I 的值改变为变量 D 中的替换位置尾部的下一个位置（步骤 S609）。这里，设定 I = 41。换句话说，从头数起的第 41 个字符，即“<HTML>[换行符]<HEAD>[换行符] <STRUCTURE:TITLE>PRESS RELEASE</STRUCTURE:TITLE>”的下一个字符被设定。

步骤 S609 之后，文档分析装置 103 判断被处理的分析知识的“重复标志”的值是否是“1”（步骤 S610）。

如果步骤 S610 是“是”，文档分析装置 103 对于该相同的分析知识再次重复进行步骤 S604 到 S606 的处理，直到与该模式匹配失败。另一方面，如果步骤 S610 是“否”，文档分析装置 103 进行到步骤 S611。

对于所有相应的分析知识条目重复执行步骤 S602－S610 的处理。当对于所有相应的分析知识条目都已经完成该处理时（步骤 S611 为“是”），将变量 D 作为分析结果输出（步骤 S612）。于是，图 6 中的处理流程结束。

图 7A－7D 示出了文档分析装置 103 的文档结构分析处理结果的

示例。

图 7A 例示了在输入图 3A 中的信息的情况下该文档结构处理的示例性结果。因为图 3A 中的输入信息是 HTML，与该文档结构分析结果无关的标记例如“<HTML>”保持在该输出中。如果需要移除该标记，可以通过已知的技术很容易地移除它们。

图 7B 示出了在输入图 3B 中的信息的情况下该文档结构处理的示例性结果。因为在图 3B 中属性信息是“INTERNET”，所以使用图 5A 中的分析知识执行该文档结构分析处理。

图 7C 示出了在输入图 3C 中的信息的情况下该文档结构处理的示例性结果。因为在图 3C 中属性信息是“MAIL”，所以使用图 5B 中的分析知识执行该文档结构分析处理。

因为在图 2D 中属性信息是“SCAN”，所以利用前述的已知技术执行该文档结构分析处理。图 7D 示出了在输入图 2D 中的信息的情况时文档结构处理结果的示例。

然后，可以使用已知技术进行文档分析装置 103 的语义属性分析处理（图 4 中的步骤 S409）。可用的已知技术包含在例如 the research report NL-161-3 (2004) of the 161st Natural Language Processing Research Meeting, the Institute of Information Processing Engineers，这里通过全文引用作为参考。语义属性分析处理的结果取决于存储在文档分析知识存储装置 102 中的、在语义属性分析处理中参考的语义属性分析知识的内容。然而，在本实施例中，假定已经获得图 8A－8D 中所示的处理结果。

接着，将参照图 9 对文档分析装置 103 的功能性作用分析处理（图 4 中的步骤 S410）进行说明。

应用包含在例如以下文档中的技术作为功能性作用分析处理：Masaru SUZUKI et al., “Customer Support Operation with a Knowledge Sharing System KIDS : An Approach based on Information Extraction and Text Structurization”, Proceedings of World Multiconference on Systemics, Cybernetics and Informatics

{SC12001, Vol. 7, pp. 89-94 (2001)}，这里通过引用其全文作为参考。

取决于每个实施例的使用目的，该功能性作用分析处理根据待分析文档的功能性作用有所不同。在本实施例中，对以下功能性作用进行分析：

通告：企业等的新闻稿陈述

报告：叙述事实的新闻或杂志的新闻条目

专栏：陈述某个观点的报告

问候：基于电子邮件等的问候信

解释：术语等的说明注释

图 9 是示出了该功能性作用分析处理的流程的示意图。

参照图 9，文档分析装置 103 加载了待分析数据，进行该文档结构分析处理以及语义属性分析处理，并且将该加载数据赋予变量 D（步骤 S901）。

随后，文档分析装置 103 基于文档结构分析处理的结果对变量 D 的值进行分割。这里将该分割的待分析数据的单个部分称为“单元文档”（步骤 S902）。顺便说一下，该分割成单元文档的得到的单元可以根据每个实施例的使用目的而不同。在第一实施例中，对于该单元使用文档结构分析处理的结果。不过，与本发明原理一致的实施例并不因而局限于此。举例来说，单个句子、单个段落、单个文档、或者类似层次结构的条目都可以设定为单元。可选地，作为修改实施例，在输入是 HTML 的情况，不仅文档结构分析处理的结果而且 HTML 标记本身也可以用于该单元文档分割的定界符。

在分析准备中，为每个功能性作用准备工作变量，并且将它们的值清零（步骤 S903）。

随后，文档分析装置 103 逐一取出该分割的单元文档（步骤 S904）。进而，逐一取出存储在文档分析知识存储装置 102 中的功能性作用分析知识（步骤 S905）。

图 10 示出了功能性作用分析知识。功能性作用分析知识的每一条都用一组三个参数表示：“模式”、“功能性作用”和“权重”。如图 10

所示，每个模式可以很好地对应于多个功能性作用和权重。

然后，文档分析装置 103 检查在步骤 S904 取得的单元文档和在步骤 S905 获得的模式之间的匹配情况（步骤 S906）。在第一实施例中，用于该功能性作用分析知识的模式的说明方法和匹配技术和文档结构分析处理中的相同。

在步骤 S906 中单元文档与模式匹配的情况（步骤 S06 为“是”），文档分析装置 103 向对应的功能性作用的工作变量加入相应的权重（步骤 S907）。在存在多个对应的功能性作用的情况，向所有对应的功能性作用增加各自的权重。

文档分析装置 103 对于功能性作用分析知识的所有条目重复步骤 S905－S907 的处理（步骤 S908）。

随后，在文档分析装置 103 检查完一个单元文档与所有功能性作用分析知识条目的模式之间的比较之后（步骤 S908 为“是”），比较单个的工作变量，并且将对应于最大值工作变量的功能性作用分配给该单元文档（步骤 S909）。这里，在存在多个最大值工作变量的情况，将分配多个功能性作用。在所有工作变量的值均为“0”的情况，将分配作用“不定”作为一个特殊的功能性作用。

进而，当步骤 S903－S909 已经对所有单元文档重复后（步骤 S910），并且对所有单元文档的处理已经结束时（步骤 S910 为“是”），该功能性作用分析处理结束。

例如在功能性作用分析处理中将图 8A 的数据输入到文档分析装置 103 的情况，根据该文档结构分割的第一单元文档变成“<HTML><HEAD>”。因为该单元文档仅由 HTML 标记组成，所以它并不构成用于在本实施例中处理的对象。

下一个单元文档是“PRESS RELEASE”。由于该单元文档与图 10 中所示的功能性作用分析知识的任一个模式都不匹配，所以将功能性作用“不定”分配给它。

进而，假定通过步骤 S903－S910 的环（loop）处理，在步骤 S904 得到了图 8A 中第 7 行开始的单元文档。

对照在步骤 S905 获得的功能性作用分析知识的模式，对单元文档 801 的元素进行连续检查。通过示例在步骤 S904 得到的单元文档 801 与图 10 中指示的知识模式 1001 相匹配（步骤 S906 为“是”），所以该例程进行到步骤 S907，在此将权重“+1”加到作为对应的功能性作用的作用“通告”的工作变量上。因为单元文档 801 与图 10 中所示功能性作用分析知识的任何其他模式不相匹配，所以在步骤 S909 将作用“通告”分配给单元文档 801。

图 11A－11D 所示是对于图 8A－8D 中的各个文档数据的功能性作用分析处理的处理结果的示例。

上面是对本实施例中的文档分析装置 103 的三个处理（文档结构分析处理，语义属性分析处理，和功能性作用分析处理）的处理内容的说明。

接着，将参照图 12 的流程图对图 1 中的部分化装置 104 的处理流程进行说明。

部分化装置 104 首先加载该待分析数据，并将该加载数据赋予变量 D 以备重写（步骤 S1201）。

随后，部分化装置 104 在变量 D 中搜索包围在任何“<FUNCTION:*>”标记内的值（步骤 S1202），并且用“<COMPONENT>”和“</COMPONENT”标记包围该值（步骤 S1203）。例如搜索该标记和插入该标记的处理可以通过已知技术例如现有的 DOM（文档对象模型）或“Xpath”来实现。在步骤 S1202 搜索到多个<FUNCTION:*>标记的情况，对各个标记执行步骤 S1203 的处理。然而，在<FUNCTION:*>标记是连续嵌套模式的情况，只把该连续<FUNCTION:*>标记的最里面的一个的值设定为该处理的对象。

步骤 S1203 之后，部分化装置 104 在变量 D 中搜索包围在“<MEANING:MAIL_ADDRESS>”标记内的值（步骤 S1204），并且用“<COMPONENT>”和“</COMPONENT”标记包围该值（步骤 S1205）。在步骤 S1204 搜索到多个“<MEANING:MAIL_ADDRESS>”标记的情况，对各个标记执行步骤 S1205 的处理。

步骤 S1205 之后，部分化装置 104 在变量 D 中搜索包围在任何“<STRUCTURE:IMG*>”标记内的值（步骤 S1206），并且用“<COMPONENT>”和“</COMPONENT”标记包围该“<STRUCTURE:IMG*>”标记（步骤 S1207）。在步骤 S1206 搜索到多个“<STRUCTURE:IMG*>”标记的情况，对各个标记执行步骤 S1207 的处理。

步骤 S1207 之后，部分化装置 104 输出被在步骤 S1202－S1207 中重写的变量 D，作为分析结果（步骤 S1208）。然后，该部分化处理结束。

下面，将通过示例说明该部分化处理。

例如在输入图 11A 中的文档数据的情况，在步骤 S1202 搜索图 11A 中用参考数字 1101、1102 和 1103 指示的部分，并且将它们分别包围在<COMPONENT>标记内。并且，在步骤 S1204 搜索图 11C 中用参考数字 1105 和 1106 指示的部分，在步骤 S1206 搜索图 11B 中用参考数字 1104 指示的部分。

图 13A－13D 是示出了在输入图 11A－11D 中的各个文档数据的情况下部分化装置 104 的处理结果示例的示意图。

下面，参照图 14 的流程图对图 1 中的索引装置 105 的处理流程进行说明。

索引装置 105 包括索引策略知识存储装置 105a，如图 15 中详细显示。

信息分量存储装置 106 包含文档索引 106a、分量索引 106b 和策略索引 106c，如图 16 中详细显示。

索引装置 105 首先加载该待索引文档数据，并将该加载数据赋予变量 D（步骤 S1401）。

接着，在由部分化装置 104 将该文档数据部分化（步骤 S1402）的情况下，索引装置 105 将变量 D 分割为由分量标记（“<COMPONENT>”和“</COMPONENT>”标记）划界的分量数据。

在步骤 S1402 之后，索引装置 105 将标识符（分量标识符 ID）

分配给各个分量以便以后可以引用该标识符（步骤 S1403）。用于生成该 ID 的方法可以由已知技术实现。该 ID 可以是例如基于随机数的足够位数的数字值、或者字母串。

接着，索引装置 105 索引该文档数据，其中分量 ID 在步骤 S1403 被分配到各个分量，并且将该文档数据和 ID 存储在文档索引 106a 中（步骤 S1404）。该索引技术使用已知的文档数据库技术来实现。

接着，索引装置 105 逐一读出在步骤 S1402 获得的分量数据项（步骤 S1405）。

然后，索引装置 105 在输入到索引装置 105 的原始数据中查找文档结构标记的路径（层级）直到到达在步骤 S1405 提取的分量数据的分量标记。它将该路径转换成向量 v_1（步骤 S1406）。这里，在该分量标记中包含任何文档结构标记的情况，它也应当包含在该向量 v_1 中。

随后，索引装置 105 在输入到索引装置 105 的原始数据中查找功能性作用标记的路径（层级）直到到达在步骤 S1405 提取的分量数据。它将该路径转换成向量 v_2（步骤 S1407）。

在步骤 S1407 之后，索引装置 105 在分量索引 106b 中注册分量数据、分量 ID、向量 v_1 和向量 v_2 四个值（步骤 S1408）。

接着，索引装置 105 获取包含在步骤 S1405 提取的分量数据值中的一组语义属性标记的所有标注，并将该标注转换成向量 v_3（步骤 S1409）。

在步骤 S1409 之后，当向量 v_3 在步骤 S1409 是空向量（其全部由“0”组成）时（步骤 S1410 为“是”），索引装置 105 进行到步骤 S1418（稍后解释），而不执行在策略索引 106c 中注册。当向量 v_3 不是空向量时，索引装置 105 进行到步骤 S1411（步骤 S1410）。稍后将参照图 17A 对该到各个向量 v_1、v_2 和 v_3 的转换（基础）进行说明。

然后，索引装置 105 获取存储在索引策略知识存储装置 105a 中的一个索引策略知识条目（步骤 S1411）。

这里，图 17A 和 17B 中示出了该索引策略知识的示例。该索引

策略知识由索引策略选择向量组成，该索引策略选择向量包含文档结构向量的三个向量：功能性作用向量、语义属性向量和索引策略向量。

图 17A 分别表示该文档结构向量的基础要素：上述的功能性作用向量和语义属性向量。

例如，语义属性向量中仅出现“COMPANY”的状态表示为（1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0）。该索引策略向量与该索引策略选择向量的语义属性向量具有相同的基础（base）。

图 17B 中的数字 901、902 和 903 分别表示索引策略知识的示例。表示为“文档结构”、“功能性作用”和“语义属性”的各个向量是索引策略选择向量的组成向量。图 17B 表示为“策略向量”的向量是索引策略向量。在第一实施例中，假定索引策略知识向量的每个元素都具有“0”或“1”的值。

重新参照图 14 来继续对索引装置 105 的处理的说明。

索引装置 105 计算在步骤 S1411 获取的索引策略知识的每个索引策略选择向量和向量 v_1、v_2 和 v_3 之间的内积（d_1、d_2 和 d_3），并且对该计算值求和以计算该分量数据和索引策略选择向量之间的相似度 S（步骤 S1412）。

索引装置 105 对于所有索引策略知识条目重复执行步骤 S1411 和 S1412 的处理（步骤 S1413）。

在步骤 S1413 之后，当对于所有索引策略知识条目，相似度 S 小于预定阈值 S_lim 时，索引装置 105 继续到步骤 S1418（稍后解释）而不执行在策略索引 106c 中的注册。当对于所有索引策略知识条目，相似度 S 不小于预定阈值 S_lim 时，索引装置 105 继续到步骤 S1415（步骤 S1414）。

在步骤 S1415，索引装置 105 从索引策略知识存储装置 105a 中提取索引策略知识向量 v_s，该索引策略知识向量 v_s 对应于大于阈值 S_lim 并且提供最大相似度 S 的索引策略选择向量（步骤 S1415）。

在步骤 S1415 之后，索引装置 105 将该分量数据的语义属性向量（向量 v_3）的组成与该索引策略知识向量（向量 v_s）之间的积设定

为新向量 v_3（步骤 S1416）。

接着，与其分量 ID 一起，索引装置 105 在策略索引 106c 中将新向量 v_3 的组成注册为具有相应的语义属性的单词的权重，（步骤 S1417）。

索引装置 105 对于包含在所有文档数据（变量 D）的所有分量重复步骤 S1405－S1417 的处理（步骤 S1418）。

例如在将图 13A 的数据作为文档数据输入到索引装置 105 的情况，根据图 14 中的步骤 S1406、S1407 和 S1409，图 13A 中的第一部分 1301 的分量向量变为：

v_1 = (0, 0, 1, 0, 0)

v_2 = (1, 0, 0, 0)

v_3 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)。因为语义属性向量 v_3 没有语义属性标记，所以它是空向量。因而，图 14 中步骤 S1410 的判断变为“是”，并且不在策略索引 106c 中注册向量 v_3。

图 13A 中的下一个部分 1302 的分量向量变为：

v_1 = (1, 0, 0, 0, 0)

v_2 = (0, 1, 0, 0)

v_3 = (1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)

在第一实施例中，即使在该向量中存在多个相同元素的情况下，该向量的各个组成部分也应当取值“0”或“1”。

关于图 13A 中的 1302 部分，与在图 17B 中的参考数字 901、902 和 903 处的索引策略选择向量的相似度分别计算如下。

参考数字 901:

d_1 = 0

d_2 = 1

d_3 = 4

相似度 S = 5

参考数字 902:

d_1 = 0

d_2 = 0

d_3 = 4

相似度 S = 4

参考数字 903:

d_1 = 0

d_2 = 0

d_3 = 1

相似度 S = 1

因此，在参考数字 901 的情况下相似度 S 变为最大。从而，索引装置 105 将新向量（1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0）在策略索引 106c 中注册（register）为具有对应于各个分量的语义属性的单词的权重，其中该新向量（1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0）是通过使向量 v_3 和参考数字 901 处的索引策略向量的单个元素相乘得到的。

更具体地，这里在这种情况下，具有<meaning: COMPANY>标记的“TSB”、具有<meaning: PRODUCT_CLASS>标记的“digital audio player”和“personal computer”、和具有<meaning: PRODUCT_NAME>标记的“GB G21”四项分别具有权重“1”，而具有<meaning: DATE>标记的“April 9”具有权重“0”并因而从策略索引 106c 中排除。

以这种方式，将输入到索引装置 105 的文档数据存储在信息分量存储装置 106 中。

下面，将参照图 18 的流程图对图 1 中的检索装置 107 的处理流程进行说明。

如图 19 详细显示，检索装置 107 包括检索策略知识存储装置 107a。

参照图 18，检索装置 107 接收检索请求的输入（步骤 S1801）。

随后，检索装置 107 判断关于在步骤 S1801 接收的检索请求，语义分析处理和部分化处理是否是未完成的处理（步骤 S1802）。

在步骤 S1802 判断的结果为语义分析处理和部分化处理未完成的

情况（步骤 S1802 为“是”），检索装置 107 通过文档分析装置 103 执行语义分析处理（步骤 S1803），并通过部分化装置 104 执行部分化处理（步骤 S1804）。

接着，检索装置 107 根据分量标记对预先或者在步骤 S1803 和 S1804 进行语义分析处理和部分化处理的检索请求进行分割（步骤 S1805）。

随后，检索装置 107 逐一读出步骤 S1805 分割的分量（步骤 S1806），向量化文档数据中的结构标记路径（步骤 S1807），向量化文档数据中的功能性标记路径（步骤 S1808），以及向量化包含在该分量中的一组语义属性标记的标注（步骤 S1809）。

步骤 S1807－S1809 的向量化处理细节分别与图 14 中的步骤 S1406、S1407 和 S1409 相同。

这里，步骤 S1807 获得的向量用 v_1 表示，步骤 S1808 获得的向量用 v_2 表示，步骤 S1809 获得的向量用 v_3 表示。

从包含在检索装置 107 中的检索策略知识存储装置 107a 中获取一条检索策略知识（步骤 S1810）。计算包含在检索策略知识条目中的文档结构向量、功能性作用向量和语义属性向量之间的内积（d_1、d_2 和 d_3）及包含在该分量中各相应的向量，并且对该计算的值求和以计算在检索策略向量和该分量向量之间的相似度 D_i（步骤 S1811）。该用于计算相似度 D_i 的方法与图 14 中的步骤 S1412 相同。

随后，检索装置 107 对于所有检索策略知识条目查找相似度 D_i，并判断相似度 D_i 的最大值是否小于预定阈值 D_lim（步骤 S1813）。

当相似度 D_i 的最大值小于值 D_lim 时（步骤 S1813 为“是”），将该检索策略向量设为元素全为“0”的空向量（步骤 S1814）。

当相似度 D_i 的最大值不小于值 D_lim 时（步骤 S1813 为“否”），从该提供最大相似度 D_i 的检索策略知识中提取检索策略向量（步骤 S1815）。

随后，检索装置 107 执行检索处理。这里，它输出由下述三个环检索结果结合的一个检索结果。

检索装置 107 在该分量标记值的基础上搜索文档索引，并存储检索文档的检索分数（步骤 S1816）。

接着，关于步骤 S1815 提取的检索策略知识向量，检索装置 107 将与该检索策略知识向量的各个元素相对应的单个意义标记（meaning tags）中包含的单词的权重和作为系数的这些元素相乘，并且搜索该分量索引。进而，检索装置 107 为该单个检索分量的检索分数评分（步骤 S1817）。

随后，检索装置 107 在该分量标记值的基础上搜索策略索引，并存储单个检索分量的检索分数（步骤 S1818）。顺便说一句，每个检索（评分）处理是已知的技术，这里略去了其详细说明。

然后，检索装置 107 为每个文档或每个分量在步骤 S1816－S1818 存储的分数求和，从而进一步得出结果分数（步骤 S1819）。

紧接着步骤 S1819，检索装置 107 对该部分化检索请求的所有分量执行步骤 S1806－S1819 的处理（步骤 S1820）。

随后，当检索装置 107 已经对整个检索请求执行了该检索处理时，根据在步骤 S1819 求和和存储的分数对该被检索文档或分量进行排序（步骤 S1821），并输出该排序的结果（步骤 S1822）。这里，该文档和分量应当是分别排序和输出的。

现在，重新将图 13D 所示的分量 1303 设定为该检索请求的可行示例，该检索请求作为待注册文档的示例。于是，向量 v_1、v_2 和 v_3 如下：

v_1 = (0, 0, 1, 0, 0)

v_2 = (1, 0, 0, 0)

v_3 = (0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)

这些向量与如图 20 所示的检索策略知识的单个示例的相似度计算如下：

参考数字 2001 的策略向量：

d_1 = 0

d_2 = 0

d_3 = 3

D_i = 3

参考数字 2002 的策略向量:

d_1 = 1

d_2 = 0

d_3 = 3

D_i = 4

参考数字 2003 的策略向量:

d_1 = 0

d_2 = 0

d_3 = 0

D_i = 0

因此，相似度 D_i 变为最大的检索策略知识是参考数字 2002 时的策略向量。

如果 D_i 的最大值小于 4，即参考数字 2002 时的策略向量；则在步骤 S1816 使用（0.5, 0, 0.5, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0）。更具体地，通过将“1”设为具有 PRODUCT_NAME 作为检索请求中的意义标记的单词“GB G21”的权重，将“0.5”设为具有 PRODUCT_CLASS 的单词“portable audio player”的权重，和将“0”设为其他任何单词的权重，从而对该分量索引进行搜索。

虽然在该策略向量中元素 COMPANY 是 0.5，但是在该检索请求中不存在相应的意义标记，所以这里忽略该单词 COMPANY。

关于该检索请求中具有意义标记 COUNT 的“5,000 pieces of music”，该策略向量的对应分量是“0”，所以该单词在步骤 S1816 中也被忽略。

在步骤 S1817，只有通过索引装置 105 在策略索引中注册的单词才成为该检索对象。因而在例如图 13A 的 1302 部分的情况，对如前所述的单词“TSB”、“digital audio player”、“personal computer”和“GB G21”附加重要性。

如上所述，与本发明的原理一致，该索引中的单个单词的权重根据文档结构、功能性作用和文档数据的单个部分包含的语义属性作出适当改变，从而可以提供能够基于文档数据的上下文而执行适当的索引的文档信息处理设备。例如，允许执行高程度控制以有助于在每个上下文中检索重要单词，或者预先去除可能成为无用信息的单词。

而且，还基于检索请求的上下文执行检索，从而可以提供能够精确获得必要信息的文档信息处理设备。例如，当文档数据的该部分（分量）作为检索请求给出时，作为检索关键词的单个单词的权重根据包含该作为检索请求的部分的文档数据的文档结构和功能性作用、和包含在该检索请求中的语义属性而作出适当改变，从而使得基于该检索请求上下文的高度检索控制变得可能。

一般地，本实施例通过由软件控制的计算机来实现。在这种情况下，该软件包括程序和数据，本发明的操作和优点通过物理使用计算机的硬件来实现，并且对可以应用现有技术的部分应用适当的现有技术。进一步，用于实施本发明的硬件和软件的具体方法和结构、通过该软件处理的范围等是可选地可以改变的。因而，在随后说明中，参考虚拟功能框图，其中构成本发明的各个功能用方框说明。顺便说一句，用于通过操作计算机来实现本发明的程序也是本发明的一个方面。

（第二实施例）

现在，将参照附图对本发明的第二实施例进行说明。在第二实施例中，用户可以通过应用模板容易地进行编辑。与第一实施例中相同的结构、操作等将使用相同的参考数字和符号来表示，并且在说明中将被省略。

图 21 是示出根据本发明的第二实施例的文档信息处理设备的结构的示意图。

如图 21 所示，与图 1 相比，文档信息处理设备 100 还提供有模板生成装置 2101 和模板存储装置 2102。

编辑装置 108 通过使用由检索装置 107 检索的至少一个信息分量来编辑新内容。编辑装置 108 向索引装置 105 发送该编辑内容。然后，

索引装置 105 提供索引作为新的信息分量并将该信息分量存储在信息分量存储装置 106 中。

这里，编辑装置 108 通过使用由检索装置 107 检索的信息分量来编辑新内容。然而，编辑装置 108 可以通过使用由不同于检索装置 107 的任何其他装置获得的信息分量来进行编辑，以这种方式使得该输出到文件的信息分量例如通过文件名来调用。并且，编辑装置 108 可以根据模板来处理编辑。其中，模板存储装置 2102 存储编辑装置 108 执行编辑所使用的模板。

存储在模板存储装置 2102 的模板可以由未包含在本发明的文档信息处理设备中的任何其他装置生成，或者它们也可以通过反映用户使用编辑装置 108 执行的编辑处理的内容来生成。

模板生成装置 2101 在基于文档分析装置 103 的文档分析结果和编辑装置 108 的编辑处理内容的基础上，为该编辑处理生成模板并将该生成的模板存储在模板存储装置 2102 中。

首先，将对编辑装置 108 进行说明。

图 22 示出了应用该编辑装置 108 的编辑工作的屏幕的示例。

数字 2203 表示了作为该编辑工作的工作空间的剪贴簿。数字 2201 表示包含在图 2B 中的分量。数字 2202 表示包含在图 2A 中的分量。

分量 2201 和 2202 被安排在剪贴簿 2203 中。

这种编辑工作通过现有技术部分提到的现有技术的软件产品来实现。

图 23A 和 23B 中示出了该剪贴簿的数据表示的示例。

图 23A 示出了不包含分量的状态下的剪贴簿数据。图 23B 示出了剪贴簿 2203 状态下的剪贴簿数据。包含在图 23B 中的单个分量具有在图 14 的流程图中步骤 S1403 提供的特定 ID。因此，即使在通过编辑装置 108 执行编辑工作之后，该单个分量也是可识别的。

下面，参照图 24 的流程图对模板生成装置 2101 的操作进行说明。

首先，模板生成装置 2101 获取（fetch）包含在剪贴簿中的一个

分量（步骤 S2401），并从信息分量存储装置 106 中为该获取的分量提取所述分量 ID（步骤 S2402）。

随后，模板生成装置 2101 以在步骤 S2402 提取的分量 ID 为线索，获取原始包含该分量的文档数据（步骤 S2403）。

模板生成装置 2101 在该文档数据中查找文档结构标记的路径(层级）直到到达该分量数据的分量标记，并将该路径转换为向量 v_1（步骤 S2404）。这里，在该分量标记中包含任何文档结构标记的情况，该文档结构标记也被包含在该向量 v_1 中。同样地，模板生成装置 2101 查找功能性作用标记的路径（层级）直到到达该文档数据的分量数据，并将该路径转换为向量 v_2（步骤 S2405）。

进而，模板生成装置 2101 获取包含在该分量数据值中的语义属性标记的所有标注，并将该标注转换为向量 v_3（步骤 S2406）。

处理步骤 S2404、S2405 和 S2406 分别与图 14 流程中的步骤 S1406、S1407 和 S1409 相似。

紧接在步骤 S2406 之后，模板生成装置 2101 将这三个生成的向量 v_1、v_2 和 v_3 转换成各个字符串，并用该字符串替换该剪贴簿中的该分量信息（步骤 S2407）。

对剪贴簿中的所有分量重复步骤 S2401－S2407 的处理。

当对于剪贴簿中的所有分量都已完成该处理时（步骤 S2408 为“是”），模板生成装置 2101 通过目前已知的 GUI 技术请求用户输入该生成模板的名字（步骤 S2409）。进而，模板生成装置 2101 将该分量部分已经被替换的剪贴簿作为模板存储到模板存储装置 2102 中，其中为其提供步骤 S2409 输入的模板名字。

以这种方式，模板生成装置 2101 生成该模板并将该生成模板存储在模板存储装置 2102 中。

从而图 25 示出了通过模板生成装置 2101 从图 23B 转换为模板的示例。

现在，将参照图 26 对编辑装置 108 在基于模板进行编辑处理的情况的处理流程进行说明。

在这种情况下，用户将要提交给该编辑处理的多个文档输入到编辑装置 108。当该一组文档没有经过语义分析和部分化时，分别通过已经解释过的文档分析装置 103 和部分化装置 104 进行语义分析和部分化。

首先，编辑装置 108 接收该一组文档的输入（步骤 S2601）。这里，将考虑一次输入所有文档的情况，但是该文档也可以逐一给出以便对其进行连续处理。

接着，编辑装置 108 利用提供给该模板的名字作为线索加载由用户预先选择的模板，并且将该模板复制到缓冲器中以便稍后重写该模板（步骤 S2602）。

随后，编辑装置 108 从该模板获取一个分量（步骤 S2603）。

然后，编辑装置 108 从步骤 S2603 获取的模板中提取文档结构向量（v_1）、功能性作用向量（v_2）和语义属性向量（v_3），这些向量是通过模板生成装置 2101 获取并且如前面结合图 24 所解释的对于该模板的每个分量说明的（步骤 S2604－S2606）。

紧接在步骤 S2604 后，编辑装置 108 从步骤 S2601 输入的一组文档中获取一个文档（步骤 S2607），并从该获取的文档中提取一个分量（步骤 S2608）。

随后，以分别与图 24 中步骤 S2404、S2405 和 S2406 相同的过程，编辑装置 108 关于步骤 S2608 提取的分量查找文档结构矢量（v_1'）、功能性作用向量(v_2')和语义属性向量(v_3')（步骤 S2609－S2611）。

接着，对于在步骤 S2604－S2606 提取的向量和在步骤 S2609－S2611 提取的向量，编辑装置 108 计算向量 v_1 和 v_1'之间的内积（s_1）、向量 v_2 和 v_2'之间的内积（s_2）、和向量 v_3 和 v_3'之间的内积（s_3），从而计算在个分量之间的相似度 S_i（= s_1+ s_2+ s_3）。临时存储个计算的相似度（步骤 S2612）。

随后，编辑装置 108 对于包含在布置 S2607 获取的文档中的所有分量重复 S2608－S2612 的处理（步骤 S2613），并且进一步对于步骤 S2601 输入的该文档组中的所有文档重复该处理（步骤 S2614）。

步骤 S2614 之后，编辑装置 108 从步骤 S2612 临时存储的单个相似度 S_i 获取最大值（S_max）（步骤 S2615）。

随后，如果个最大值（S_max）小于预定阈值（S_lim）（步骤 S2616 为“否”），则编辑装置 108 删除在缓冲器中复制的模板的相应的分量部分的值（步骤 S2617）。相反地，如果个最大值（S_max）至少等于预定阈值（S_lim）（步骤 S2616 为“是”），则编辑装置 108 从个文档中的分量中选择最大化该相似度 S_i 的分量（步骤 S2618），并用该选择的分量替换在缓冲器中复制的模板的相应的分量部分的值（步骤 S2619）。

接着，编辑装置 108 对于在步骤 S2602 输入的模板中包含的所有分量重复步骤 S2603－S2619 的处理（步骤 S2620）。

该缓冲器中的模板因为已经按照上述处理流程正确地进行了该替换处理，从而作为编辑结果输出（步骤 S2621）。于是，该处理结束。

我们来考虑例如当指定图 25 中所示的模板和将图 27A 和 27B 中的数据作为一组文档输入的情况。

关于图 25 中用参考数字 2501 表示的该模板的部分，向量如下：

v_1 = (1, 0, 0, 0, 0)

v_2 = (0, 1, 0, 0)

v_3 = (1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)

关于图 27A 和 27B 中用参考数字 2701－2706 表示的各个部分，向量如下：

部分 2701:

v_1’ = (0, 0, 1, 0, 0)

v_2’ = (1, 0, 0, 0)

v_3’ = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

部分 2702:

v_1’ = (1, 0, 0, 0, 0)

v_2’ = (0, 1, 0, 0)

v_3' = (1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)

部分 2703:

v_1' = (1, 0, 0, 0, 0)

v_2' = (1, 0, 0, 0)

v_3' = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)

部分 2704:

v_1' = (0, 0, 1, 0, 0)

v_2' = (1, 0, 0, 0)

v_3' = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

部分 2705:

v_1' = (1, 0, 0, 0, 0)

v_2' = (0, 0, 1, 0)

v_3' = (1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)

部分 2706:

v_1' = (0, 0, 0, 0, 1)

v_2' = (0, 0, 0, 0)

v_3' = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

因此，与部分 2501 的相似度分别计算如下:

部分 2701: S_i = 0

部分 2702: S_i = 6

部分 2703: S_i = 1

部分 2704: S_i = 0

部分 2705: S_i = 5

部分 2706: S_i = 0

因而，相似度在部分 2702 最大。如果阈值 S_max 至多等于 5，则用部分 2702 替换图 25 中模板的部分 2501。

这个例子表示，部分 2702 和 2705 作为语义属性向量与部分 2501 等价，但是考虑到功能性作用向量的差别，选择部分 2702 作为更合适的分量。

同样地，对于参考数字 2502 表示的部分的向量：

v_1 = (0, 0, 0, 0, 1)

v_2 = (0, 0, 0, 0)

v_3 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

相似度为：

部分 2701: S_i = 0

部分 2702: S_i = 0

部分 2703: S_i = 0

部分 2704: S_i = 0

部分 2705: S_i = 0

部分 2706: S_i = 1

因而，相似度在部分 2706 最大。如果阈值 S_max 是“0”，则用部分 2706 替换图 25 中模板的部分 2502。

这里假定两个部分2501和2502均被替换，则编辑结果变为图28A所示。图 28B 示出了该编辑结果由浏览器显示的例子。

如上所述，根据本发明，能够提供一种文档信息处理设备，除了具有第一实施例的优点外，还具有能够易于收集被加入到产生的剪贴页的剪贴片断的优点。也就是说，用户能够很方便地再次产生与模板相似的剪贴页。例如根据图 26 的流程，编辑装置 108 能够根据存储在模板存储装置 2102 中的模板自动执行编辑处理。

而且，该剪贴页的模板是根据产生的剪贴页中的剪贴分量的组合而生成的。因而能够提供一种文档信息处理设备，当用户再次产生相似的剪贴页时，用户可以容易地根据该模板产生该剪贴页。

本发明的文档信息处理设备可以通过由计算机例如工作站（WS）或者个人计算机（PC）激活程序实现。

图 29 示出了描述一个示例性计算机的示意图，其中可以实现与本发明一致的系统和方法。该计算机包括执行该程序的中央处理单元 2901，存储被处理的程序和数据的存储器 2902，存储该程序、待搜索数据和 OS（操作系统）的磁盘驱动器 2903，以及用于从光盘读取程

序和数据和向光盘写入的光盘驱动器 2904。

进一步，该计算机包括作为在显示装置等上显示屏幕的界面的图像输出单元 2905，从键盘、鼠标、触摸板等接收输入的输入接收单元 2906，作为向外部装置传送输出或者从外部装置接收输入的接口（例如 USB（通用串行总线）或音频输出终端）的输出／输入单元 2907。此外，该文档信息处理设备包括显示装置 2908 例如 LCD、CRT 或投影机，输入装置 2909 例如键盘或鼠标，外部装置 2910 例如存储卡阅读器或扬声器。

中央处理单元 2901 从磁盘驱动器 2903 中读出程序并存入存储器 2902 中，然后运行该程序，从而实施图 1 所示的各个功能块。在运行程序期间，可以从磁盘驱动器 2903 读取一些或所有待搜索数据并存入存储器 2902 中。

作为基础操作，通过输入装置 2909 接收用户作出的检索请求，根据该检索请求搜索存储在磁盘驱动器 2903 和存储器 2902 中的待搜索数据。并且，在显示装置 2908 上显示检索结果。

显示在显示装置 2908 上的检索结果可以进一步通过声音提供给用户，例如使用作为外部装置 2910 连接的扬声器。可选地，将该检索结果使用作为外部装置 2910 连接的打印机以打印的形式提供。

本发明并不限于这些实施例，而是可以通过在不脱离本发明主旨的范围内修改组成元素而在改进的基础上形成。而且，可以通过适当地组合实施例中公开的多个组成元素而形成各种新技术。例如，可以从实施例中所示的所有组成元素中省略一些组成元素。而且，可以适当地组合不同实施例中的组成元素。

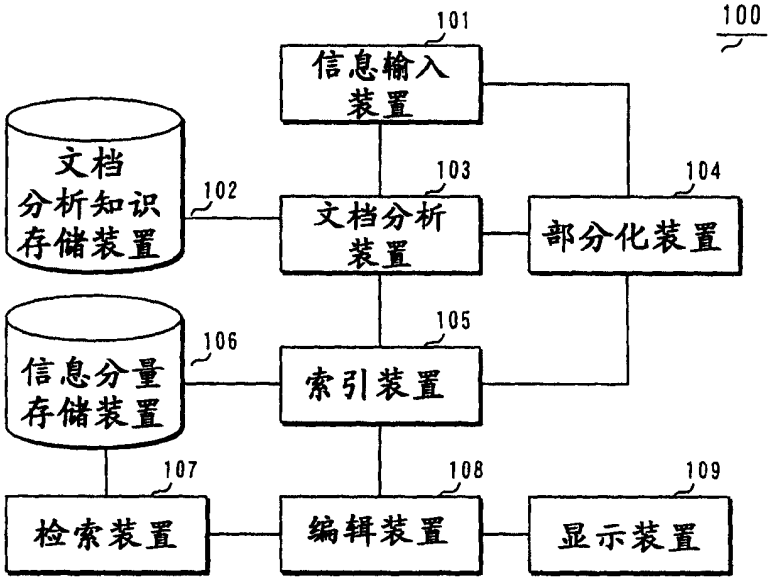

100
101
信息输入装置
文档分析知识存储装置
102
103
文档分析装置
104
部分化装置
信息分量存储装置
106
105
索引装置
107
检索装置
108
编辑装置
109
显示装置

图1

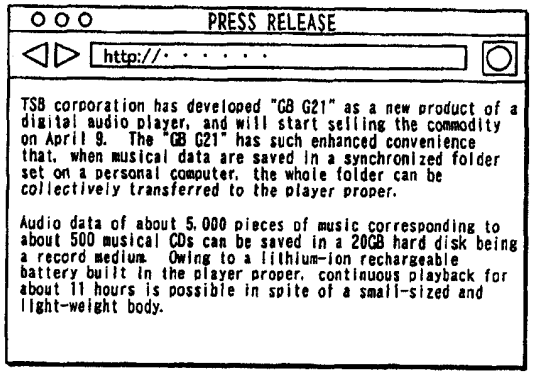


图2A

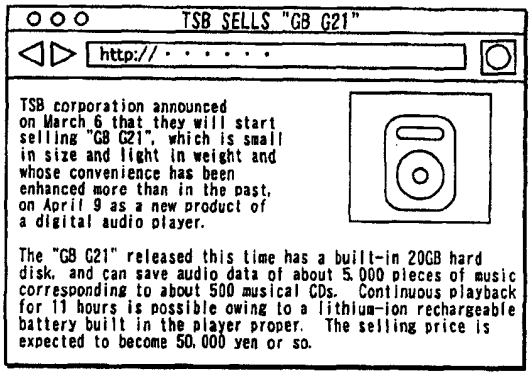


图2B

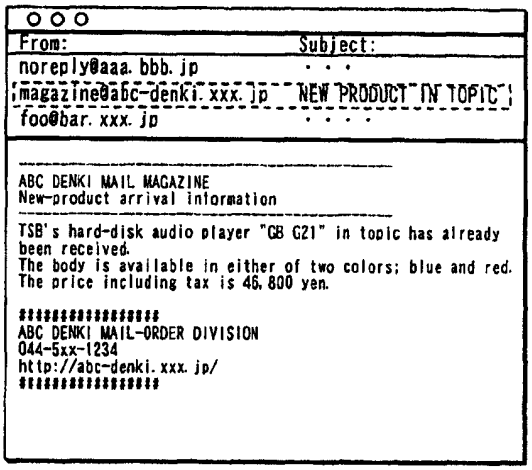


图2C

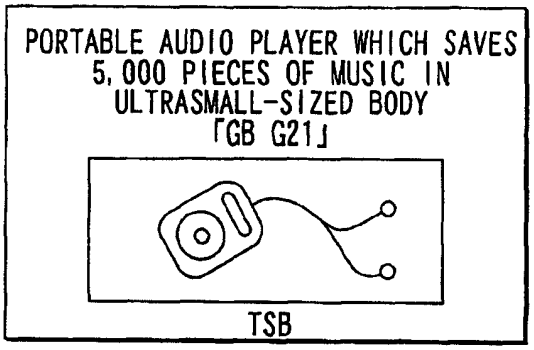


图2D

```
<HTML>
<HEAD>
<TITLE>PRESS RELEASE</TITLE>
</HEAD>
<BODY>
<P>
TSB corporation has developed "GB G21" as a new product of a digital audio
player, and will start selling the commodity on April 9. The "GB G21" has
such enhanced convenience that, when musical data are saved in a synchronized
folder set on a personal computer, the whole folder can be collectively
transferred to the player proper.
</P>
<P>
Audio data of about 5,000 pieces of music corresponding to about 500 musical
CDs can be saved in a 20GB hard disk being a record medium. Owing to a
lithium-ion rechargeable battery built in the player proper, continuous
playback for about 11 hours is possible in spite of a small-sized and light-
weight body.
</P>
</BODY>
</HTML>
```

图 3A

```
<HTML>
<HEAD>
<TITLE>TSB SELLS "GB G21"</TITLE>
</HEAD>
<BODY>
<P>
TSB corporation announced on March 6 that they will start selling "GB G21",
which is small in size and light in weight and whose convenience has been
enhanced more than in the past, on April 9 as a new product of a digital
audio player.
</P>
<IMG src="/img/g21.jpg" alt="TSB G21BODY">
<P>
The "GB G21" released this time has a built-in 20GB hard disk, and can save
audio data of about 5,000 pieces of music corresponding to about 500 musical
CDs. Continuous playback for 11 hours is possible owing to a lithium-ion
rechargeable battery built in the player proper. The selling price is
expected to become 50,000 yen or so.
</P>
</BODY>
</HTML>
```

图 3B

```
From: magazine@abc-denki.xxx.jp
To: customer@abc-denki.xxx.jp
Subject: NEW PRODUCT IN TOPIC

-----------------------------------------
ABC DENKI MAIL MAGAZINE
New-product arrival information
-----------------------------------------
TSB's hard-disk audio player "GB G21" in topic has already been received.
The body is available in either of two colors; blue and red. The price
including tax is 46,800 yen.


##################
ABC DENKI MAIL-ORDER DIVISION
044-5xx-1234
http://abc-denki.xxx.jp/
##################
```

图 3C

图1中文档分析装置103的处理流程

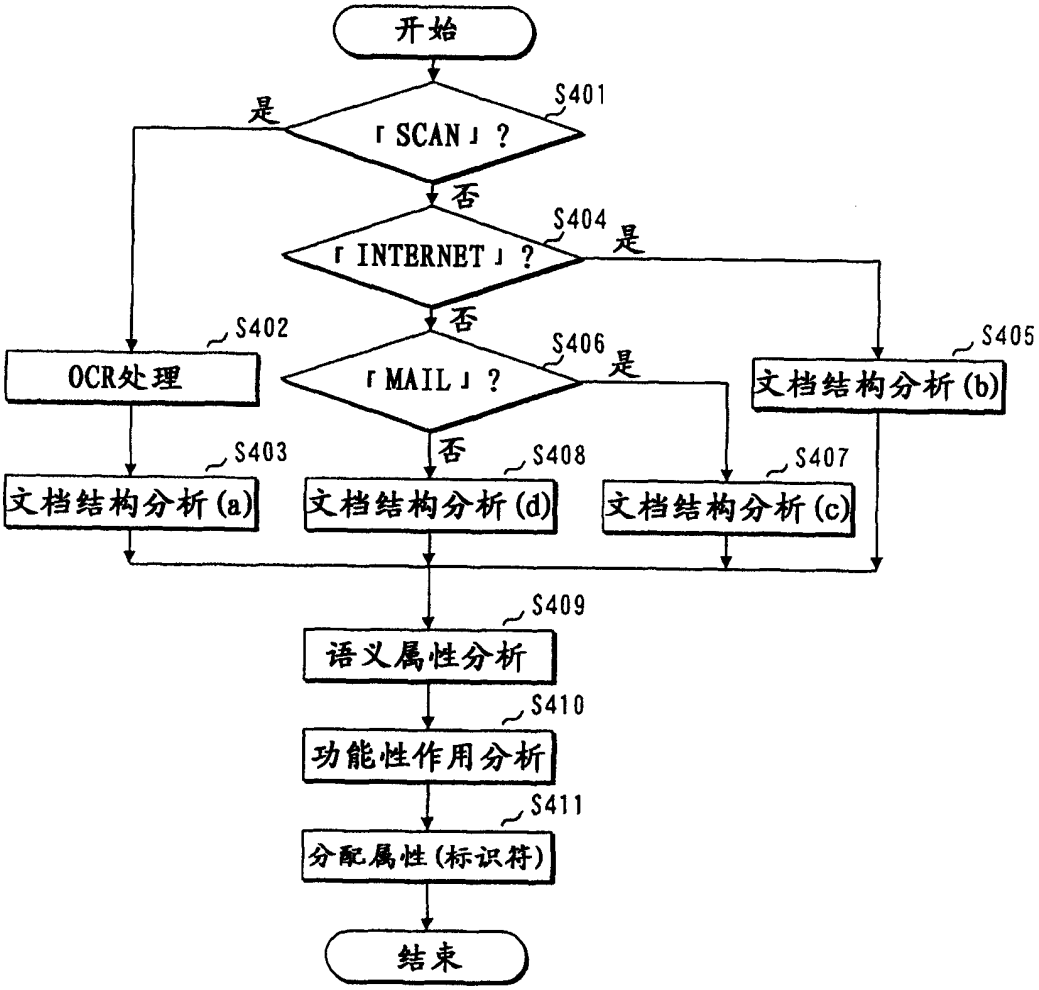


图4

有关文档结构分析的知识条目的示例

用于分析HTML文档结构的知识的示例

分析知识模式	重复标记	文档结构标记
<TITLE>(.*)</TITLE>	0	<STRUCTURE:TITLE>$1</STRUCTURE:TITLE>
<P>(.*)</P>	1	<STRUCTURE:PARAGRAPH>$1</STRUCTURE:PARAGRAPH>
<IMG src="(.*)" alt="(.*)">	1	<STRUCTURE:IMG src="$1" alt="$2"/>
<IMG alt="(.*)" src="(.*)">	1	<STRUCTURE:IMG src="$2" alt="$1"/>
<IMG src="(.*)">	1	<STRUCTURE:IMG src="$1"/>

501

图5A

用于分析电子邮件文档结构的知识的示例

模式	重复标记	文档结构标记
[Ss]ubject:¥s*(.*)¥n	0	<STRUCTURE:TITLE> $1 </STRUCTURE:TITLE>
##+(.*)##+¥s*$	0	<STRUCTURE:SIGNATURE> $1 </STRUCTURE:SIGNATURE>
--+(.*)--+¥s*$	0	<STRUCTURE:SIGNATURE> $1 </STRUCTURE:SIGNATURE>
==+(.*)==+¥s*$	0	<STRUCTURE:SIGNATURE> $1 </STRUCTURE:SIGNATURE>
##+(.*)##+(¥s*¥S)	1	<STRUCTURE:HEADER> $1 </STRUCTURE:HEADER> $2
--+(.*)--+(¥s*¥S)	1	<STRUCTURE:HEADER> $1 </STRUCTURE:HEADER> $2
==+(.*)==+(¥s*¥S)	1	<STRUCTURE:HEADER> $1 </STRUCTURE:HEADER> $2

图5B

图4中文档结构分析的处理流程

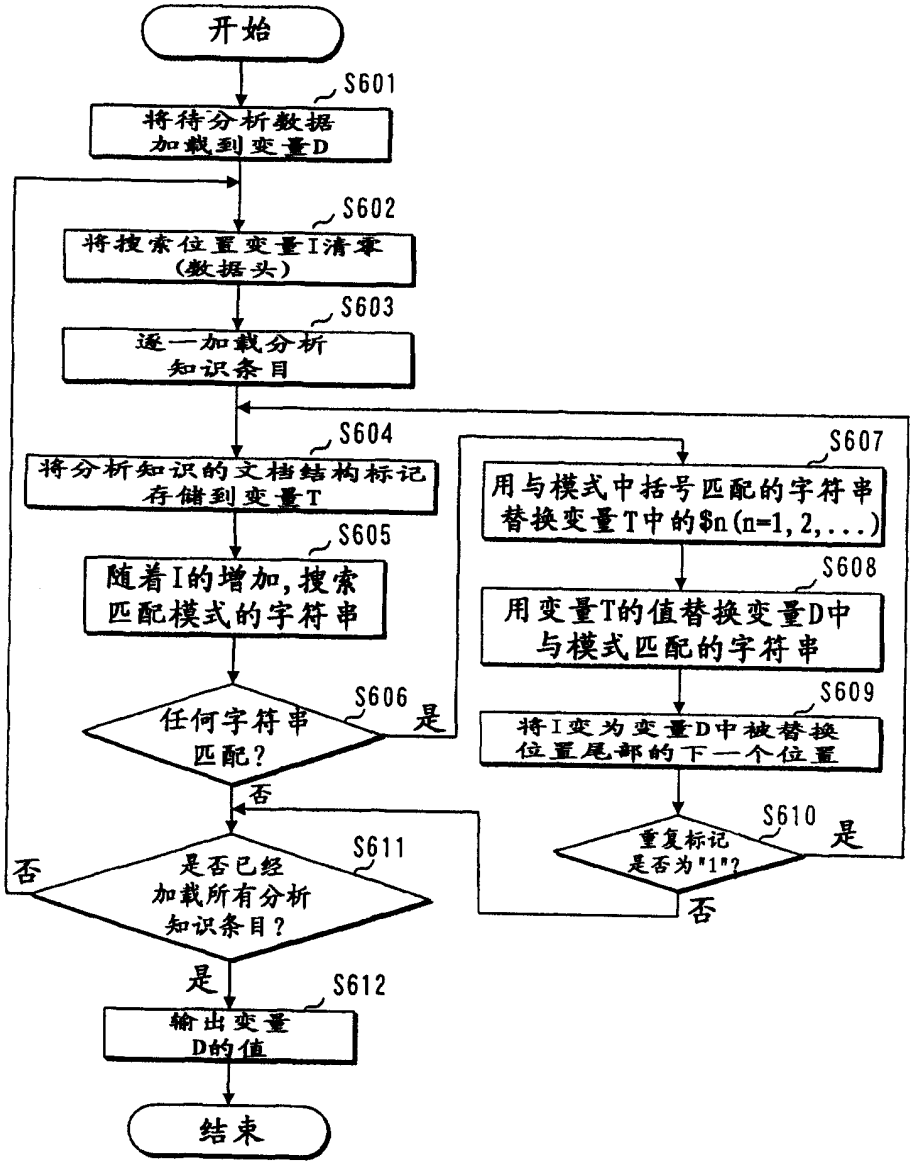


图6

```
<HTML>
<HEAD>
<STRUCTURE: TITLE> PRESS RELEASE </STRUCTURE: TITLE>
</HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
TSB corporation has developed "GB G21" as a new product of a digital
audio player, and will start selling the commodity on April 9.  The "GB
G21" has such enhanced convenience that, when musical data are saved in
a synchronized folder set on a personal computer, the whole folder can
be collectively transferred to the player proper.
</STRUCTURE: PARAGRAPH>
<STRUCTURE: PARAGRAPH>
Audio data of about 5,000 pieces of music corresponding to about 500
musical CDs can be saved in a 20GB hard disk being a record medium.
Owing to a lithium-ion rechargeable battery built in the player proper,
continuous playback for about 11 hours is possible in spite of a small-
sized and light-weight body.
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图 7A

```
<HTML>
<HEAD>
<STRUCTURE: TITLE> TSB SELLS "GB G21" </STRUCTURE: TITLE>
</HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
TSB corporation announced on March 6 that they will start selling "GB
G21", which is small in size and light in weight and whose convenience
has been enhanced more than in the past, on April 9 as a new product of
a digital audio player.
</STRUCTURE: PARAGRAPH>
<STRUCTURE: IMGsrc="/img/g21.jpg"alt="TSB G21 body"/>
<STRUCTURE: PARAGRAPH>
The "GB G21" released this time has a built-in 20GB hard disk, and can
save audio data of about 5,000 pieces of music corresponding to about
500 musical CDs.  Continuous playback for 11 hours is possible owing to
a lithium-ion rechargeable battery built in the player proper.  The
selling price is expected to become 50,000 yen or so.
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图 7B

```
From: magazine@abc-denki.xxx.jp
To: customer@abc-denki.xxx.jp

<STRUCTURE: TITLE> NEW PRODUCT IN TOPIC </ STRUCTURE:
TITLE>

<STRUCTURE: HEADER>
ABC DENKI MAIL MAGAZINE
New-product arrival information
</STRUCTURE: HEADER>
TSB's hard-disk audio player "GB G21" in topic has
already been received.
The body is available in either of two colors; blue and
red.  The price including tax is 46,800 yen.

<STRUCTURE: SIGNATURE>
ABC DENKI MAIL-ORDER DIVISION
044-5xx-1234
http://abc-denki.xxx.jp/
</STRUCTURE: SIGNATURE>
```

图 7C

```
<STRUCTURE: TITLE>
"GB G21"
PORTABLE AUDIO PLAYER WHICH SAVES 5,000 PIECES OF MUSIC
IN ULTRASMALL-SIZED BODY
</STRUCTURE: TITLE>

<STRUCTURE: IMGsrc="file:..../..../temp01.jpg"/>
```

图 7D

```
<HTML>
<HEAD>
<STRUCTURE: TITLE> PRESS RELEASE </STRUCTURE: TITLE>
</HEAD>
<BODY>                                                    801
<STRUCTURE: PARAGRAPH>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation has developed "<MEANING: PRODUCT_NAME>
GB G21 </MEANING: <PRODUCT_NAME>" as a new product of a <MEANING: PRODUCT_CLASS> digital audio
player </MEANING: PRODUCT_CLASS>, and will start selling the commodity on <MEANING: DATE>
April 9 </MEANING: DATE>.  The "GB G21" has such enhanced convenience that, when musical data
are saved in a synchronized folder set on a <MEANING: PRODUCT_CLASS> personal computer </
MEANING: PRODUCT_CLASS>, the whole folder can be collectively transferred to the player
proper.
</STRUCTURE: PARAGRAPH>
<STRUCTURE: PARAGRAPH>
Audio data of about <MEANING: COUNT> 5,000 pieces of music </MEANING: COUNT> corresponding to
about <MEANING: COUNT> 500 musical CDs </MEANING: COUNT> can be saved in a <MEANING: CAPACITY>
20GB </MEANING: CAPACITY> <MEANING: PERIPHERAL> hard disk </MEANING: PERIPHERAL> being a
record medium.  Owing to a <MEANING: PERIPHERAL> lithium-ion rechargeable battery </MEANING:
PERIPHERAL> built in the player proper, continuous playback for about <MEANING: DURATION> 11
hours </MEANING: DURATION> is possible in spite of a small-sized and light-weight body.
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图8A

```
<HTML>
<HEAD>
<STRUCTURE: TITLE> <MEANING:COMPANY> TSB </MEANING: COMPANY> SELLS "<MEANING: PRODUCT_NAME> GB
G21 </MEANING: PRODUCT_NAME>" </STRUCTURE: TITLE>
</HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation announced on <MEANING: DATE> March 6 </
MEANING: DATE> that they will start selling "<MEANING: PRODUCT_NAME> GB G21 </MEANING:
PRODUCT_NAME>", which is small in size and light in weight and whose convenience has been
enhanced more than in the past, on <MEANING: DATE> April 9 </MEANING: DATE> as a new product
of a <MEANING: PRODUCT_CLASS> digital audio player </MEANING: PRODUCT_CLASS>.
</STRUCTURE: PARAGRAPH>
<STRUCTURE: IMGsrc="/img/g21.jpg"alt="TSB G21 body"/>
<STRUCTURE: PARAGRAPH>
The "<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>" released this time has a built-
in <MEANING: CAPACITY> 20GB </MEANING: CAPACITY> <MEANING: PERIPHERAL> hard disk </MEANING:
PERIPHERAL>, and can save audio data of about <MEANING: COUNT> 5,000 pieces of music </
MEANING: COUNT> corresponding to about <MEANING: COUNT> 500 musical CDs </MEANING: COUNT>.
Continuous playback for <MEANING: DURATION> 11 hours </MEANING: DURATION> is possible owing to
a <MEANING: PERIPHERAL> lithium-ion rechargeable battery </MEANING: PERIPHERAL> built in the
player proper.  The selling price is expected to become <MEANING: PRICE> 50,000 yen </MEANING:
PRICE> or so.
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图8B

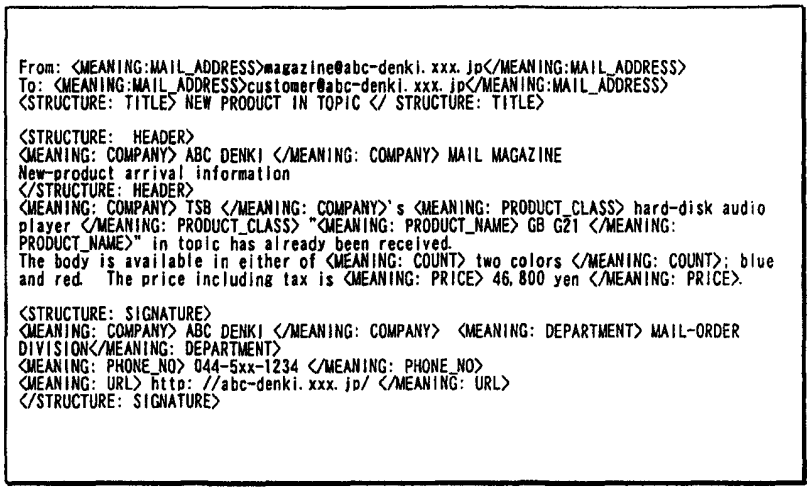

From: <MEANING:MAIL_ADDRESS>magazine@abc-denki.xxx.jp</MEANING:MAIL_ADDRESS>
To: <MEANING:MAIL_ADDRESS>customer@abc-denki.xxx.jp</MEANING:MAIL_ADDRESS>
<STRUCTURE: TITLE> NEW PRODUCT IN TOPIC </ STRUCTURE: TITLE>
<STRUCTURE: HEADER>
<MEANING: COMPANY> ABC DENKI </MEANING: COMPANY> MAIL MAGAZINE
New-product arrival information
</STRUCTURE: HEADER>
<MEANING: COMPANY> TSB </MEANING: COMPANY>'s <MEANING: PRODUCT_CLASS> hard-disk audio player </MEANING: PRODUCT_CLASS> "<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>" in topic has already been received.
The body is available in either of <MEANING: COUNT> two colors </MEANING: COUNT>; blue and red. The price including tax is <MEANING: PRICE> 46,800 yen </MEANING: PRICE>.
<STRUCTURE: SIGNATURE>
<MEANING: COMPANY> ABC DENKI </MEANING: COMPANY> <MEANING: DEPARTMENT> MAIL-ORDER DIVISION</MEANING: DEPARTMENT>
<MEANING: PHONE_NO> 044-5xx-1234 </MEANING: PHONE_NO>
<MEANING: URL> http://abc-denki.xxx.jp/ </MEANING: URL>
</STRUCTURE: SIGNATURE>

图8C

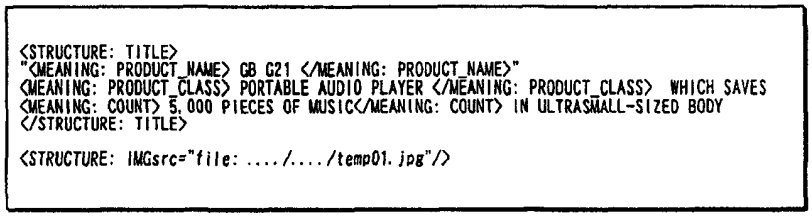

<STRUCTURE: TITLE>
"<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>"
<MEANING: PRODUCT_CLASS> PORTABLE AUDIO PLAYER </MEANING: PRODUCT_CLASS> WHICH SAVES
<MEANING: COUNT> 5,000 PIECES OF MUSIC</MEANING: COUNT> IN ULTRASMALL-SIZED BODY
</STRUCTURE: TITLE>
<STRUCTURE: IMGsrc="file: ..../..../temp01.jpg"/>

图8D

图4中S410功能性作用分析的流程

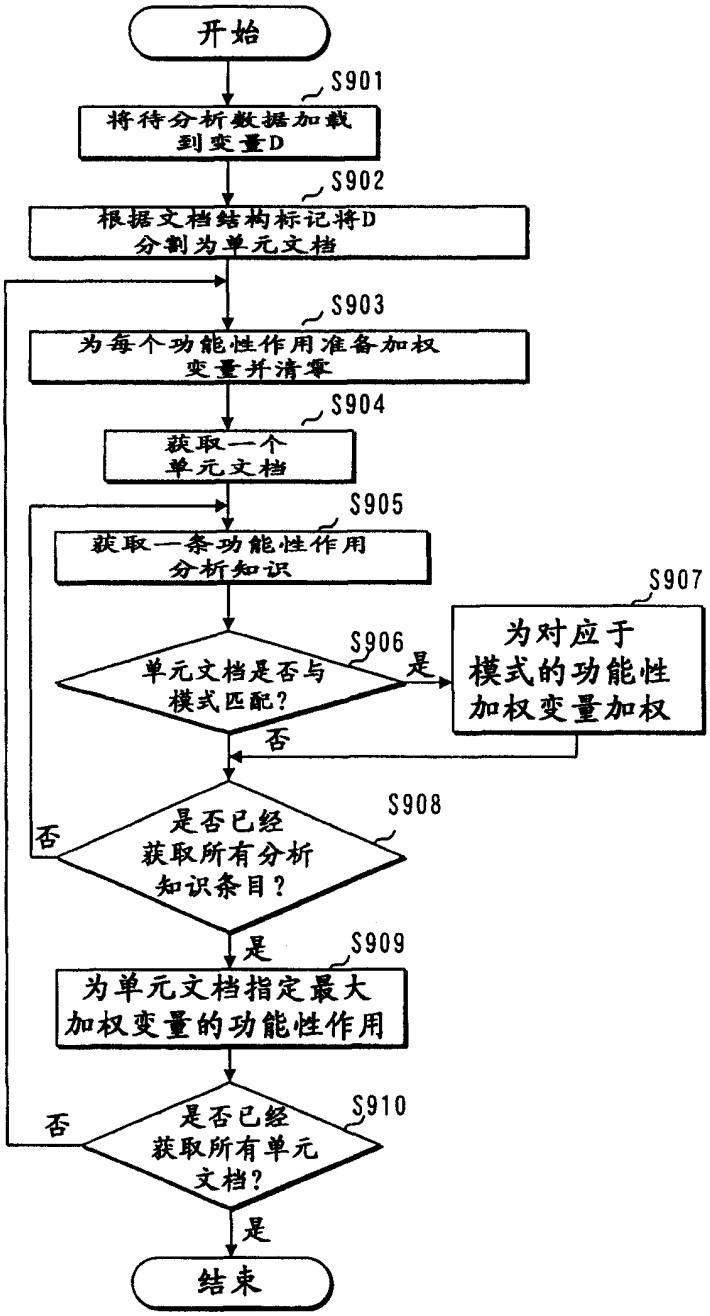


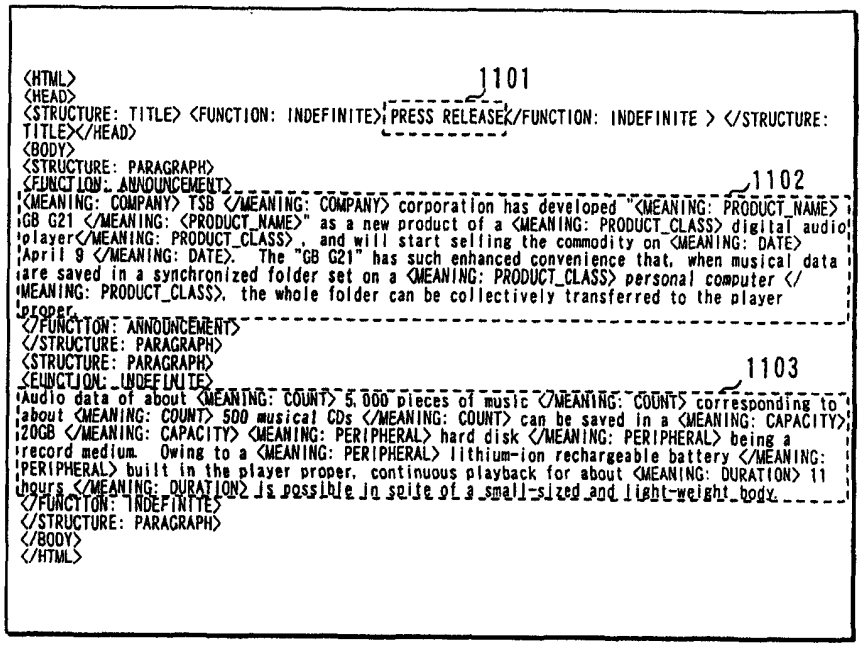
图9

功能性作用分析知识

模式	功能性作用	权值
<MEANING: COMPANY> .* </MEANING: COMPANY> [^\.]* will	ANNOUNCEMENT	+ 1
OUR COMPANY	ANNOUNCEMENT	+ 3
ANNOUNCED	ACCOUNT COLUMN	+ 2 + 1
ANNOUNCES	ACCOUNT	+ 2
APPEALED	ACCOUNT COLUMN	+ 1 + 1
IS EXPECTED	ACCOUNT COLUMN	+ 1 + 1
WILL BECOME	COLUMN	+ 3
MAY POSSIBLY	COLUMN	+ 1
IN DANGER OF	COLUMN	+ 1
guess*?	COLUMN	+ 1
TERM	EXPLANATION	+ 1
SIGNIFIES	EXPLANATION	+ 2

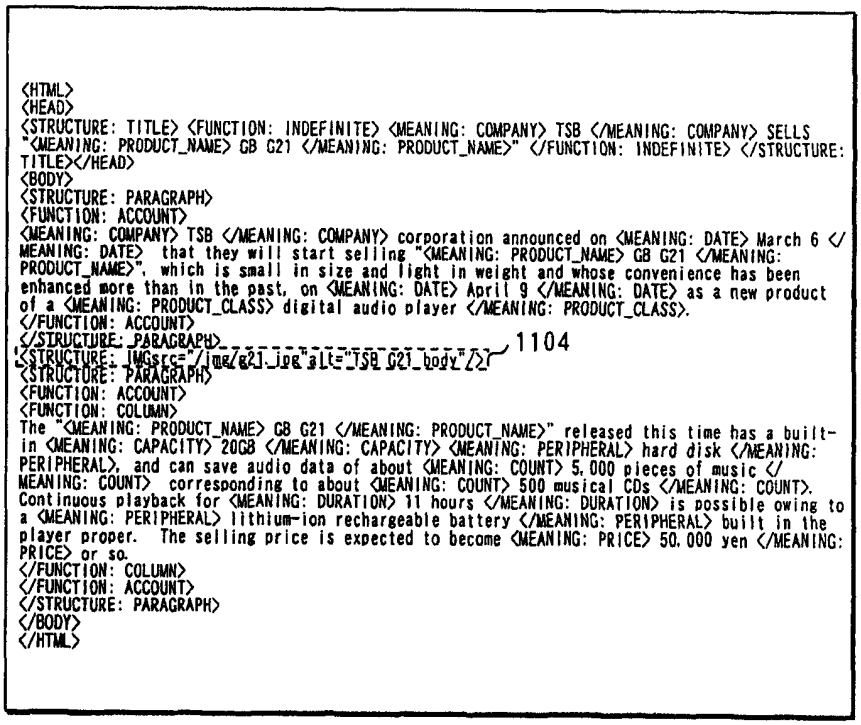
1001

图10




```
<HTML>
<HEAD>
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE>PRESS RELEASE</FUNCTION: INDEFINITE > </STRUCTURE:
TITLE></HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
<FUNCTION: ANNOUNCEMENT>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation has developed "<MEANING: PRODUCT_NAME>
GB G21 </MEANING: <PRODUCT_NAME>" as a new product of a <MEANING: PRODUCT_CLASS> digital audio
player</MEANING: PRODUCT_CLASS> , and will start selling the commodity on <MEANING: DATE>
April 9 </MEANING: DATE>. The "GB G21" has such enhanced convenience that, when musical data
are saved in a synchronized folder set on a <MEANING: PRODUCT_CLASS> personal computer </
MEANING: PRODUCT_CLASS>, the whole folder can be collectively transferred to the player
proper.
</FUNCTION: ANNOUNCEMENT>
</STRUCTURE: PARAGRAPH>
<STRUCTURE: PARAGRAPH>
<FUNCTION: INDEFINITE>
Audio data of about <MEANING: COUNT> 5,000 pieces of music </MEANING: COUNT> corresponding to
about <MEANING: COUNT> 500 musical CDs </MEANING: COUNT> can be saved in a <MEANING: CAPACITY>
20GB </MEANING: CAPACITY> <MEANING: PERIPHERAL> hard disk </MEANING: PERIPHERAL> being a
record medium. Owing to a <MEANING: PERIPHERAL> lithium-ion rechargeable battery </MEANING:
PERIPHERAL> built in the player proper, continuous playback for about <MEANING: DURATION> 11
hours </MEANING: DURATION> is possible in spite of a small-sized and light-weight body.
</FUNCTION: INDEFINITE>
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图11A




```
<HTML>
<HEAD>
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE> <MEANING: COMPANY> TSB </MEANING: COMPANY> SELLS
"<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>" </FUNCTION: INDEFINITE> </STRUCTURE:
TITLE></HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
<FUNCTION: ACCOUNT>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation announced on <MEANING: DATE> March 6 </
MEANING: DATE> that they will start selling "<MEANING: PRODUCT_NAME> GB G21 </MEANING:
PRODUCT_NAME>", which is small in size and light in weight and whose convenience has been
enhanced more than in the past, on <MEANING: DATE> April 9 </MEANING: DATE> as a new product
of a <MEANING: PRODUCT_CLASS> digital audio player </MEANING: PRODUCT_CLASS>.
</FUNCTION: ACCOUNT>
</STRUCTURE: PARAGRAPH>
<STRUCTURE: IMG src="/img/g21.jpg" alt="TSB G21 body" />
<STRUCTURE: PARAGRAPH>
<FUNCTION: ACCOUNT>
<FUNCTION: COLUMN>
The "<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>" released this time has a built-
in <MEANING: CAPACITY> 20GB </MEANING: CAPACITY> <MEANING: PERIPHERAL> hard disk </MEANING:
PERIPHERAL>, and can save audio data of about <MEANING: COUNT> 5,000 pieces of music </
MEANING: COUNT> corresponding to about <MEANING: COUNT> 500 musical CDs </MEANING: COUNT>.
Continuous playback for <MEANING: DURATION> 11 hours </MEANING: DURATION> is possible owing to
a <MEANING: PERIPHERAL> lithium-ion rechargeable battery </MEANING: PERIPHERAL> built in the
player proper. The selling price is expected to become <MEANING: PRICE> 50,000 yen </MEANING:
PRICE> or so.
</FUNCTION: COLUMN>
</FUNCTION: ACCOUNT>
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图11B

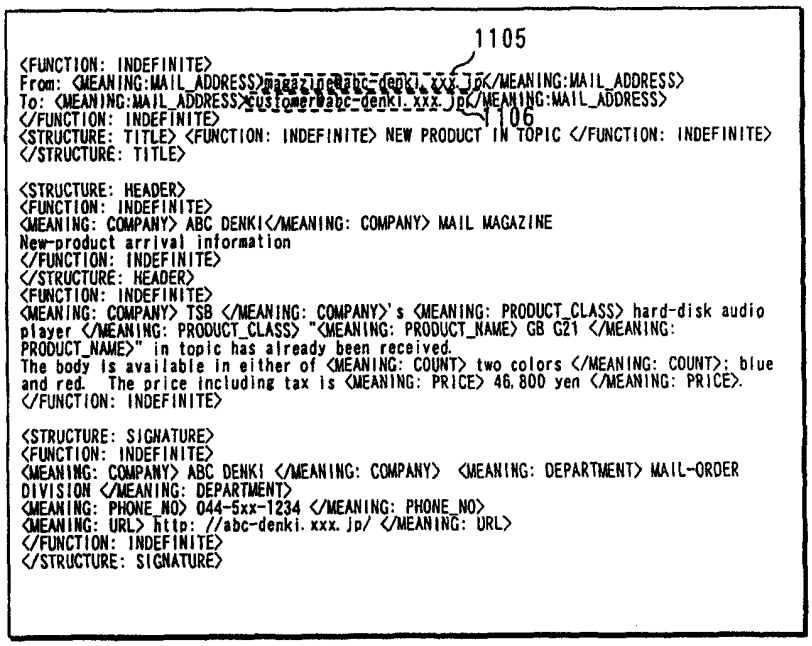

1105
<FUNCTION: INDEFINITE>
From: <MEANING:MAIL_ADDRESS>magazine@abc-denki.xxx.jp</MEANING:MAIL_ADDRESS>
To: <MEANING:MAIL_ADDRESS>customer@abc-denki.xxx.jp</MEANING:MAIL_ADDRESS>
</FUNCTION: INDEFINITE>
1106
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE> NEW PRODUCT IN TOPIC </FUNCTION: INDEFINITE>
</STRUCTURE: TITLE>
<STRUCTURE: HEADER>
<FUNCTION: INDEFINITE>
<MEANING: COMPANY> ABC DENKI</MEANING: COMPANY> MAIL MAGAZINE
New-product arrival information
</FUNCTION: INDEFINITE>
</STRUCTURE: HEADER>
<FUNCTION: INDEFINITE>
<MEANING: COMPANY> TSB </MEANING: COMPANY>'s <MEANING: PRODUCT_CLASS> hard-disk audio player </MEANING: PRODUCT_CLASS> "<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>" in topic has already been received.
The body is available in either of <MEANING: COUNT> two colors </MEANING: COUNT>: blue and red. The price including tax is <MEANING: PRICE> 46,800 yen </MEANING: PRICE>.
</FUNCTION: INDEFINITE>
<STRUCTURE: SIGNATURE>
<FUNCTION: INDEFINITE>
<MEANING: COMPANY> ABC DENKI </MEANING: COMPANY> <MEANING: DEPARTMENT> MAIL-ORDER DIVISION </MEANING: DEPARTMENT>
<MEANING: PHONE_NO> 044-5xx-1234 </MEANING: PHONE_NO>
<MEANING: URL> http: //abc-denki.xxx.jp/ </MEANING: URL>
</FUNCTION: INDEFINITE>
</STRUCTURE: SIGNATURE>

图11C

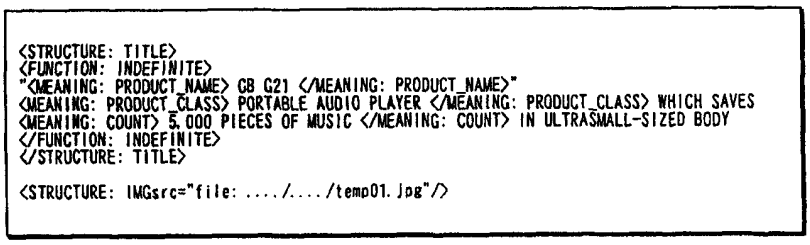

<STRUCTURE: TITLE>
<FUNCTION: INDEFINITE>
"<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>"
<MEANING: PRODUCT_CLASS> PORTABLE AUDIO PLAYER </MEANING: PRODUCT_CLASS> WHICH SAVES
<MEANING: COUNT> 5,000 PIECES OF MUSIC </MEANING: COUNT> IN ULTRASMALL-SIZED BODY
</FUNCTION: INDEFINITE>
</STRUCTURE: TITLE>
<STRUCTURE: IMGsrc="file: ..../..../temp01.jpg"/>

图11D

图1中部分化装置104的处理流程

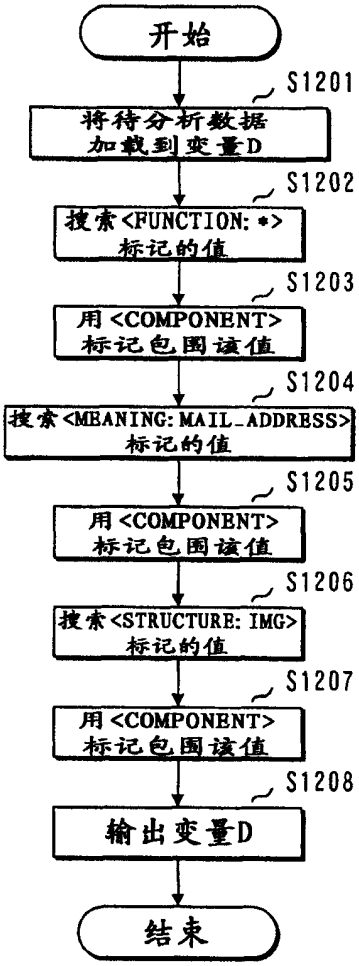
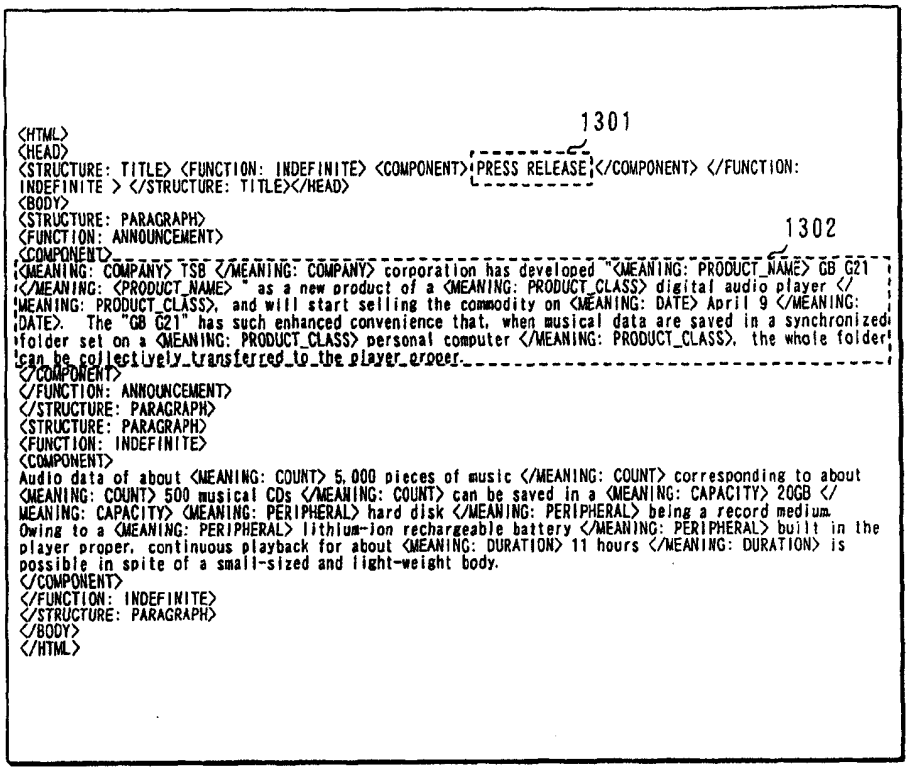


图12




```
<HTML>
<HEAD>
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE> <COMPONENT>PRESS RELEASE</COMPONENT> </FUNCTION:
INDEFINITE > </STRUCTURE: TITLE></HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
<FUNCTION: ANNOUNCEMENT>
<COMPONENT>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation has developed "<MEANING: PRODUCT_NAME> GB G21
</MEANING: <PRODUCT_NAME> " as a new product of a <MEANING: PRODUCT_CLASS> digital audio player </
MEANING: PRODUCT_CLASS>, and will start selling the commodity on <MEANING: DATE> April 9 </MEANING:
DATE>. The "GB G21" has such enhanced convenience that, when musical data are saved in a synchronized
folder set on a <MEANING: PRODUCT_CLASS> personal computer </MEANING: PRODUCT_CLASS>, the whole folder
can be collectively transferred to the player proper.
</COMPONENT>
</FUNCTION: ANNOUNCEMENT>
</STRUCTURE: PARAGRAPH>
<STRUCTURE: PARAGRAPH>
<FUNCTION: INDEFINITE>
<COMPONENT>
Audio data of about <MEANING: COUNT> 5,000 pieces of music </MEANING: COUNT> corresponding to about
<MEANING: COUNT> 500 musical CDs </MEANING: COUNT> can be saved in a <MEANING: CAPACITY> 20GB </
MEANING: CAPACITY> <MEANING: PERIPHERAL> hard disk </MEANING: PERIPHERAL> being a record medium.
Owing to a <MEANING: PERIPHERAL> lithium-ion rechargeable battery </MEANING: PERIPHERAL> built in the
player proper, continuous playback for about <MEANING: DURATION> 11 hours </MEANING: DURATION> is
possible in spite of a small-sized and light-weight body.
</COMPONENT>
</FUNCTION: INDEFINITE>
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图13A

```
<HTML>
<HEAD>
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE> <COMPONENT> <MEANING: COMPANY> TSB </MEANING: COMPANY> SELLS
"<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME> "</COMPONENT> </FUNCTION: INDEFINITE> </
STRUCTURE: TITLE>
</HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
<FUNCTION: ACCOUNT>
<COMPONENT>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation announced on <MEANING: DATE> March 6 </MEANING:
DATE> that they will start selling "<MEANING: PRODUCT_NAME> GB C21 </MEANING: PRODUCT_NAME>", which
is small in size and light in weight and whose convenience has been enhanced more than in the past, on
<MEANING: DATE> April 9 </MEANING: DATE> as a new product of a <MEANING: PRODUCT_CLASS> digital audio
player </MEANING: PRODUCT_CLASS>.
</COMPONENT>
</FUNCTION: ACCOUNT>
</STRUCTURE: PARAGRAPH>
<COMPONENT> <STRUCTURE: IMGsrc="/img/g21.jpg"alt="TSB G21 body"/> </COMPONENT>
<STRUCTURE: PARAGRAPH>
<FUNCTION: ACCOUNT>
<FUNCTION: COLUMN>
<COMPONENT>
The "<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>" released this time has a built-in
<MEANING: CAPACITY> 20GB </MEANING: CAPACITY> <MEANING: PERIPHERAL> hard disk </MEANING: PERIPHERAL>,
and can save audio data of about <MEANING: COUNT> 5,000 pieces of music </MEANING: COUNT>
corresponding to about <MEANING: COUNT> 500 musical CDs </MEANING: COUNT>. Continuous playback for
<MEANING: DURATION> 11 hours </MEANING: DURATION> is possible owing to a <MEANING: PERIPHERAL>
lithium-ion rechargeable battery </MEANING: PERIPHERAL> built in the player proper. The selling price
is expected to become <MEANING: PRICE> 50,000 yen </MEANING: PRICE> or so.
</COMPONENT>
</FUNCTION: COLUMN>
</FUNCTION: ACCOUNT>
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
```

图13B

```
<FUNCTION: INDEFINITE>
<COMPONENT>
From: <MEANING:MAIL_ADDRESS><COMPONENT>magazine@abc-denki.xxx.jp</COMPONENT></MEANING:MAIL_ADDRESS>
To: <MEANING:MAIL_ADDRESS><COMPONENT>customer@abc-denki.xxx.jp</COMPONENT></MEANING:MAIL_ADDRESS>
</COMPONENT>
</FUNCTION: INDEFINITE>
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE> <COMPONENT> NEW PRODUCT IN TOPIC </COMPONENT> </FUNCTION: INDEFINITE> </STRUCTURE:
TITLE>

<STRUCTURE: HEADER>
<FUNCTION: INDEFINITE>
<COMPONENT>
<MEANING: COMPANY> ABC DENKI </MEANING: COMPANY> MAIL MAGAZINE
New-product arrival information
</COMPONENT>
</FUNCTION: INDEFINITE>
</STRUCTURE: HEADER>
<FUNCTION: INDEFINITE>
</COMPONENT>
<MEANING: COMPANY> TSB </MEANING: COMPANY>'s <MEANING: PRODUCT_CLASS> hard-disk audio player </MEANING: PRODUCT_CLASS>
"<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>" in topic has already been received.
The body is available in either of <MEANING: COUNT> two colors </MEANING: COUNT>; blue and red.  The price including tax is
<MEANING: PRICE> 46,800 yen </MEANING: PRICE>.
</COMPONENT>
</FUNCTION: INDEFINITE>

<STRUCTURE: SIGNATURE>
<FUNCTION: INDEFINITE>
<COMPANY>
<MEANING: COMPANY> ABC DENKI </MEANING: COMPANY>  <MEANING: DEPARTMENT> MAIL-ORDER DIVISION </MEANING: DEPARTMENT>
<MEANING: PHONE_NO> 044-5xx-1234 </MEANING: PHONE_NO>
<MEANING: URL> http://abc-denki.xxx.jp/ </MEANING: URL>
</COMPONENT>
</FUNCTION: INDEFINITE>
</STRUCTURE: SIGNATURE>
```

图13C

```
<STRUCTURE: TITLE>                                                    1303
<FUNCTION: INDEFINITE>
<COMPONENT>
"<MEANING: PRODUCT_NAME> GB G21 </MEANING: PRODUCT_NAME>"
<MEANING: PRODUCT_CLASS> PORTABLE AUDIO PLAYER </MEANING: PRODUCT_CLASS> WHICH SAVES <MEANING: COUNT> 5,000 PIECES OF MUSIC
</MEANING: COUNT> IN ULTRASMALL-SIZED BODY
</COMPONENT>
</FUNCTION: INDEFINITE>
</STRUCTURE: TITLE>

<COMPONENT><STRUCTURE:IMGsrc="file:..../..../temp01.jpg"/></COMPONENT>
```

图13D

图1中索引装置105的处理流程

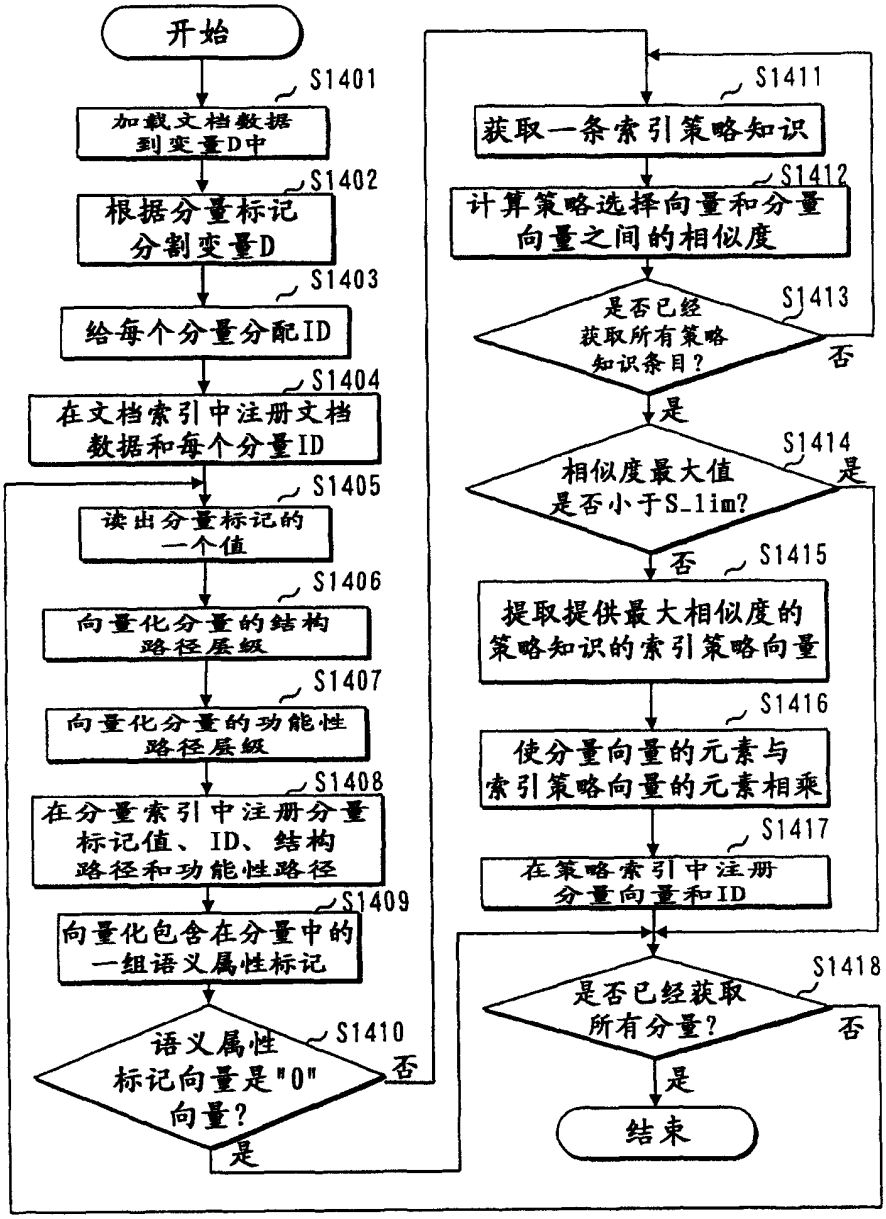


图14

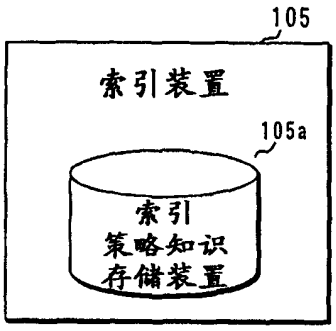


图15

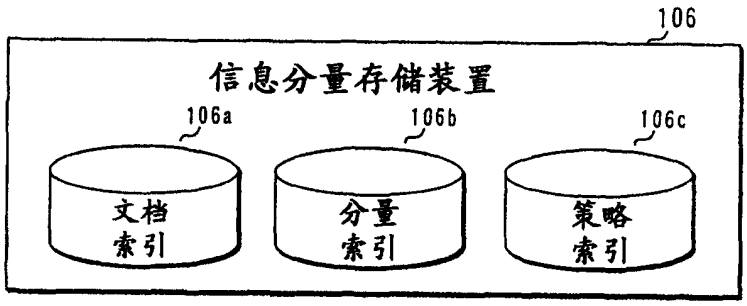


图16

索引策略知识

(PARAGRAPH, HEADER, TITLE, SIGNATURE, IMAGE)

(INDEFINITE, ANNOUNCEMENT, ACCOUNT, COLUMN)

(COMPANY, PERSON_NAME, PRODUCT_CLASS, PRODUCT_NAME, PRICE, DATE, COUNT, DURATION, LENGTH, WEIGHT, CAPACITY, MAIL_ADDRESS, URL, COUNTRY_NAME, PERIPHERAL)

图17A

文档结构	(0, 0, 0, 0, 0)
功能性作用	(0, 1, 0, 0)
语义属性	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0)
索引策略向量	(1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

901

文档结构	(0, 0, 1, 0, 0)
功能性作用	(0, 0, 0, 0)
语义属性	(1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0)
索引策略向量	(1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

902

文档结构	(0, 0, 0, 1, 0)
功能性作用	(0, 0, 0, 0)
语义属性	(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0)
索引策略向量	(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

903

图17B

图1中检索装置107的处理流程

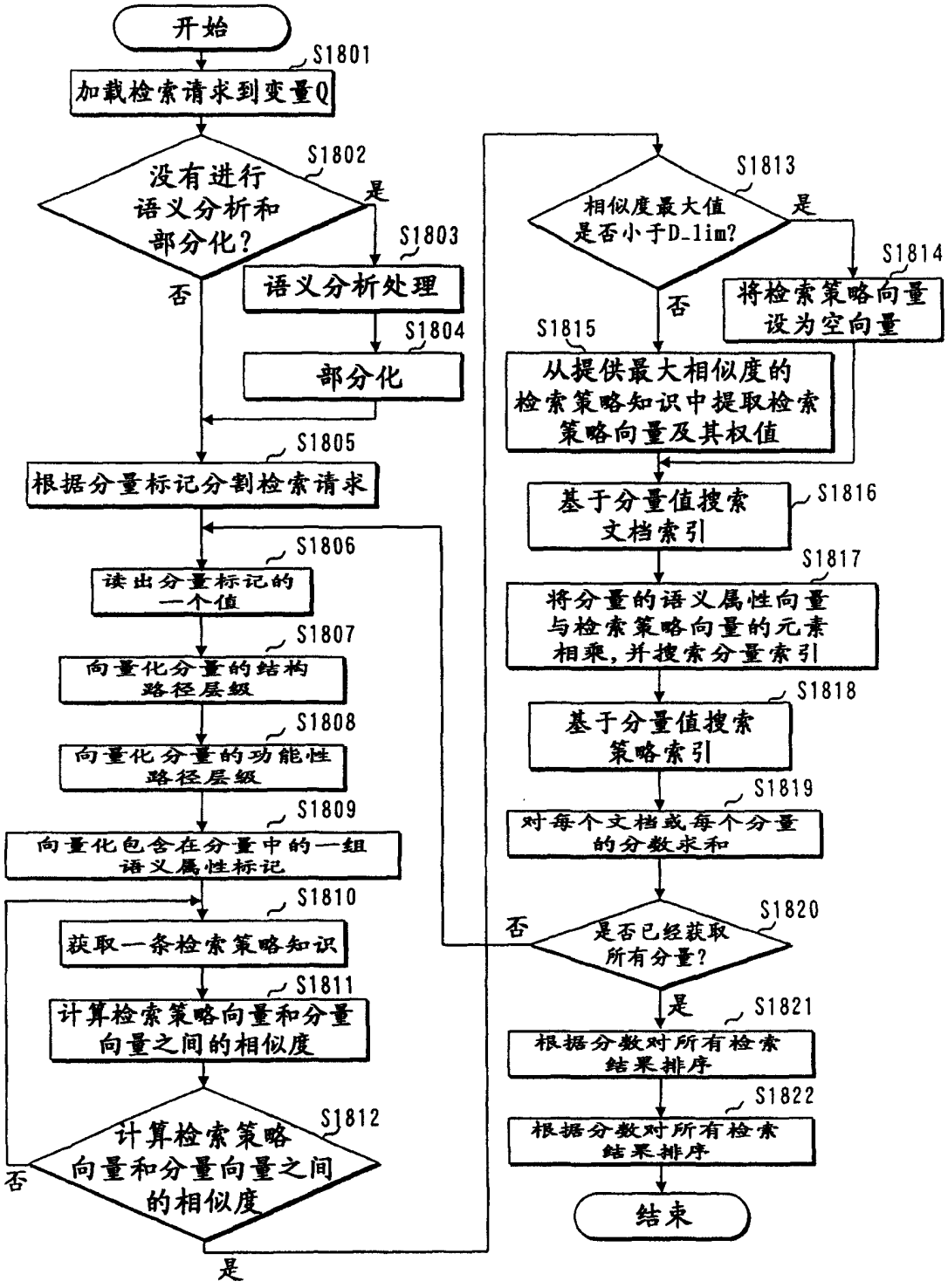


图18

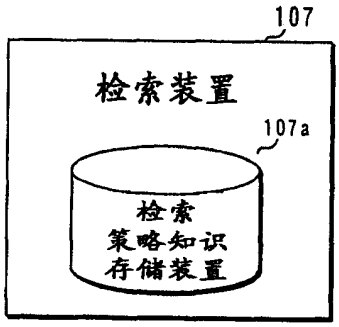


图19

检索策略知识

文档结构	(0, 0, 0, 0, 0)
功能性作用	(0, 1, 0, 0)
语义属性	(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0)
索引策略向量	(0.5, 0, 0.5, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

2001

文档结构	(0, 0, 1, 0, 0)
功能性作用	(0, 0, 0, 0)
语义属性	(1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0)
索引策略向量	(0.5, 0, 0, 5, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

2002

文档结构	(0, 0, 0, 1, 0)
功能性作用	(0, 0, 0, 0)
语义属性	(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0)
索引策略向量	(1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

2003

图20

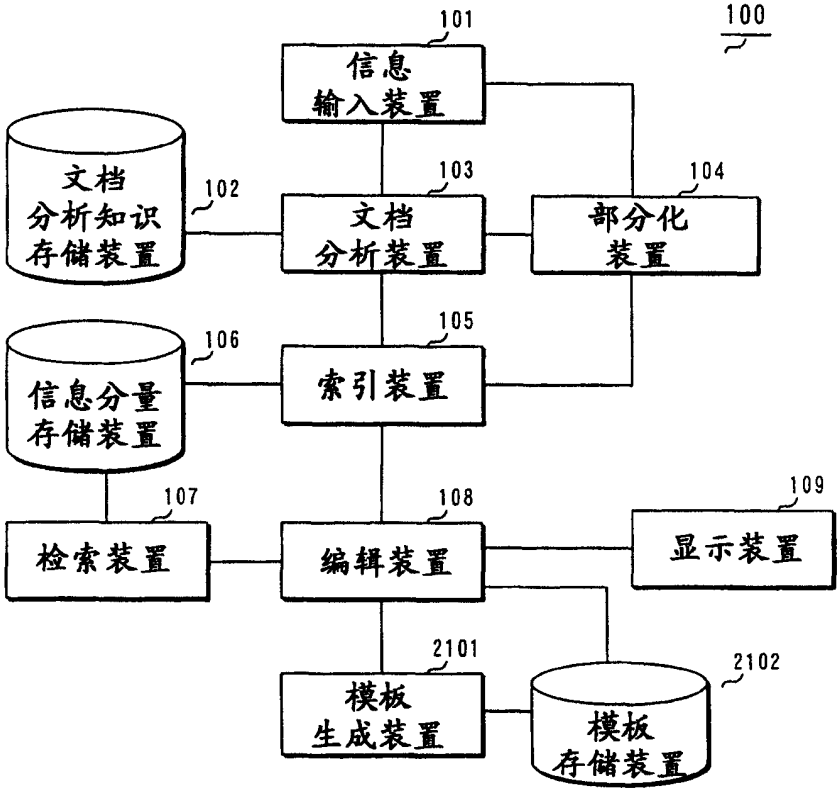

100
101
信息
输入装置
文档
分析知识
存储装置
102
103
文档
分析装置
104
部分化
装置
106
105
信息分量
存储装置
索引装置
107
108
109
检索装置
编辑装置
显示装置
2101
2102
模板
生成装置
模板
存储装置

图21

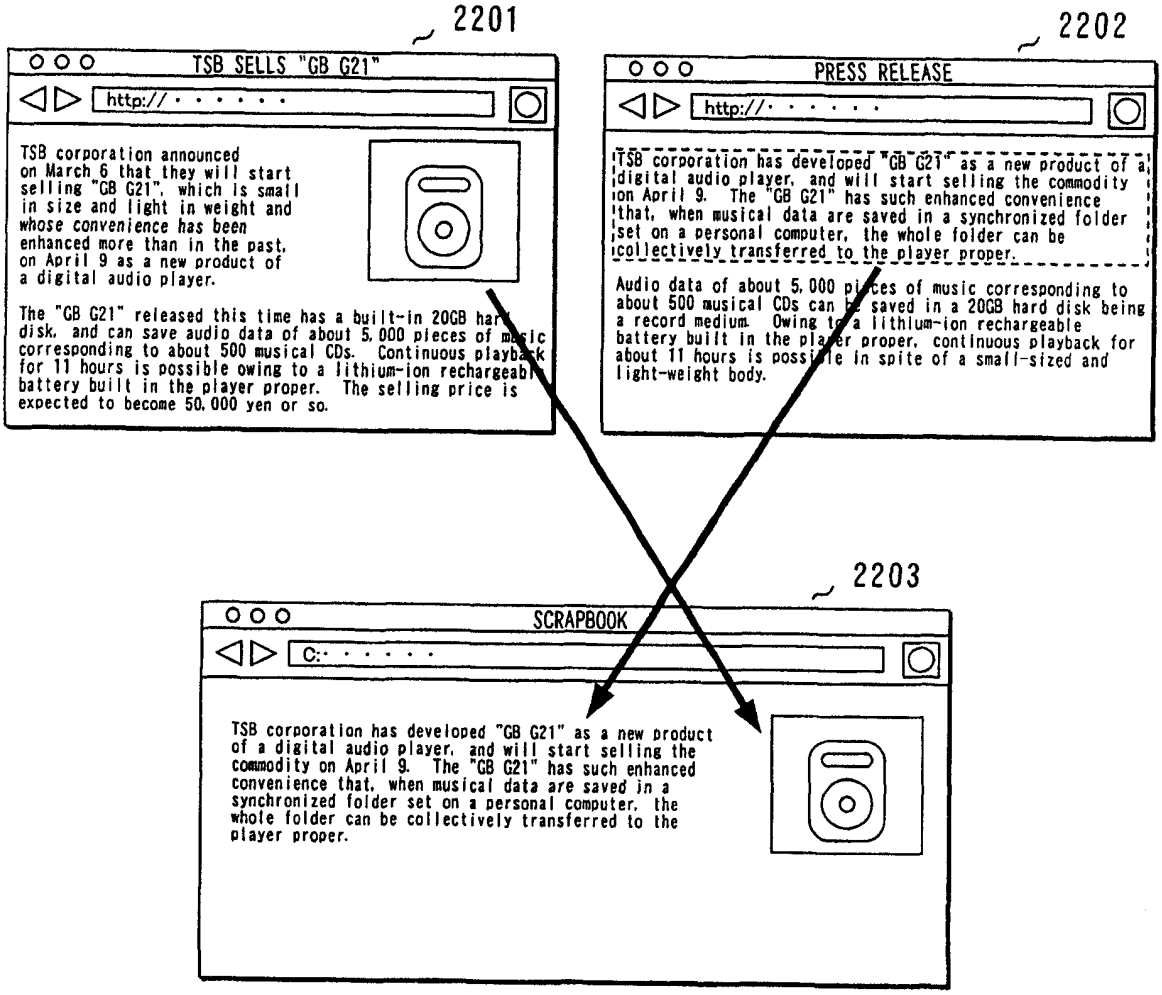

2201
TSB SELLS "GB G21"
http://
TSB corporation announced on March 6 that they will start selling "GB G21", which is small in size and light in weight and whose convenience has been enhanced more than in the past, on April 9 as a new product of a digital audio player.
The "GB G21" released this time has a built-in 20GB hard disk, and can save audio data of about 5,000 pieces of music corresponding to about 500 musical CDs. Continuous playback for 11 hours is possible owing to a lithium-ion rechargeable battery built in the player proper. The selling price is expected to become 50,000 yen or so.
2202
PRESS RELEASE
http://
TSB corporation has developed "GB G21" as a new product of a digital audio player, and will start selling the commodity on April 9. The "GB G21" has such enhanced convenience that, when musical data are saved in a synchronized folder set on a personal computer, the whole folder can be collectively transferred to the player proper.
Audio data of about 5,000 pieces of music corresponding to about 500 musical CDs can be saved in a 20GB hard disk being a record medium. Owing to a lithium-ion rechargeable battery built in the player proper, continuous playback for about 11 hours is possible in spite of a small-sized and light-weight body.
2203
SCRAPBOOK
C:
TSB corporation has developed "GB G21" as a new product of a digital audio player, and will start selling the commodity on April 9. The "GB G21" has such enhanced convenience that, when musical data are saved in a synchronized folder set on a personal computer, the whole folder can be collectively transferred to the player proper.

图22

```
<scrapbook>
</scrapbook>
```

图23A

```
<scrapbook>
<COMPONENT ID=" 0103adcsdaddgt3t56d1" >
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation has developed "<MEANING: PRODUCT_NAME> GB G21 </
MEANING: <PRODUCT_NAME>" as a new product of a <MEANING: PRODUCT_CLASS> digital audio player </MEANING:
PRODUCT_CLASS>, and will start selling the commodity on <MEANING: DATE> April 9 </MEANING: DATE>. The
"GB G21" has such enhanced convenience that, when musical data are saved in a synchronized folder set on
a <MEANING: PRODUCT_CLASS> personal computer </MEANING: PRODUCT_CLASS>, the whole folder can be
collectively transferred to the player proper.
</COMPONENT>
<COMPONENT ID=" 4163fecsefdr2t6r5gdw" ><STRUCTURE:IMGsrc="/img/g21.jpg" alt="TSB G21body"/></COMPONENT>
</scrapbook>
```

图23B

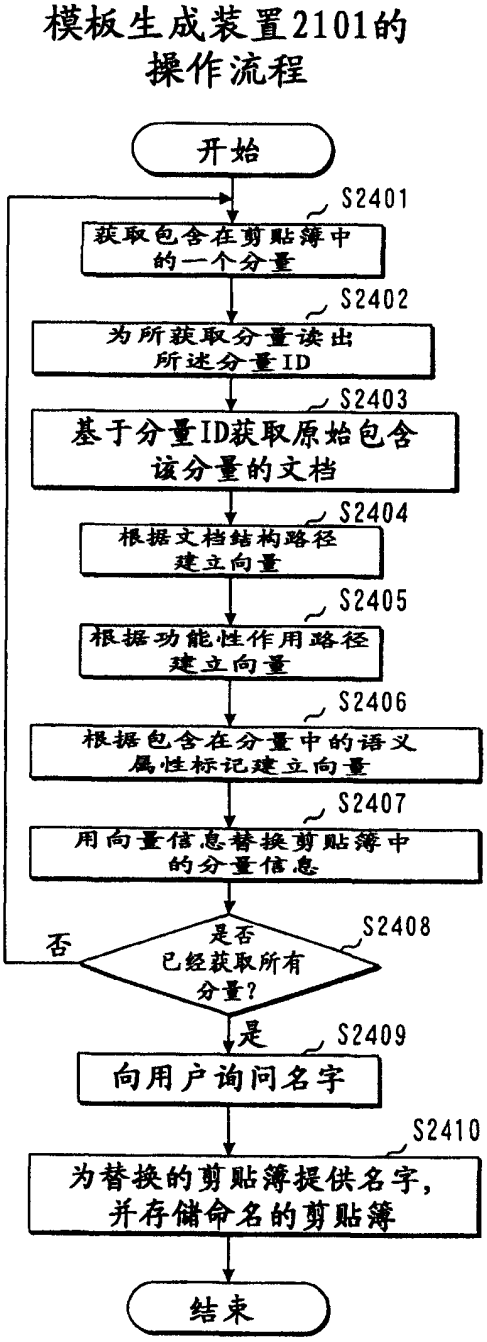

模板生成装置2101的
操作流程
开始
S2401
获取包含在剪贴簿中
的一个分量
S2402
为所获取分量读出
所述分量ID
S2403
基于分量ID获取原始包含
该分量的文档
S2404
根据文档结构路径
建立向量
S2405
根据功能性作用路径
建立向量
S2406
根据包含在分量中的语义
属性标记建立向量
S2407
用向量信息替换剪贴簿中
的分量信息
S2408
是否
已经获取所有
分量?
否
是
S2409
向用户询问名字
S2410
为替换的剪贴簿提供名字,
并存储命名的剪贴簿
结束

图24

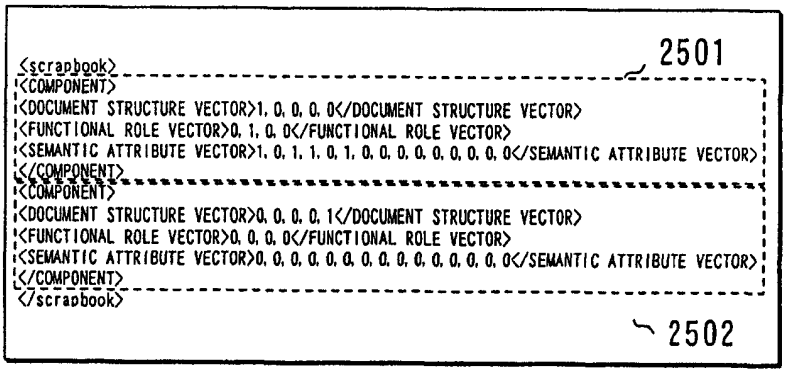

<scrapbook>
2501
<COMPONENT>
<DOCUMENT STRUCTURE VECTOR>1, 0, 0, 0, 0</DOCUMENT STRUCTURE VECTOR>
<FUNCTIONAL ROLE VECTOR>0, 1, 0, 0</FUNCTIONAL ROLE VECTOR>
<SEMANTIC ATTRIBUTE VECTOR>1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0</SEMANTIC ATTRIBUTE VECTOR>
</COMPONENT>
<COMPONENT>
<DOCUMENT STRUCTURE VECTOR>0, 0, 0, 0, 1</DOCUMENT STRUCTURE VECTOR>
<FUNCTIONAL ROLE VECTOR>0, 0, 0, 0</FUNCTIONAL ROLE VECTOR>
<SEMANTIC ATTRIBUTE VECTOR>0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0</SEMANTIC ATTRIBUTE VECTOR>
</COMPONENT>
</scrapbook>
2502

图25

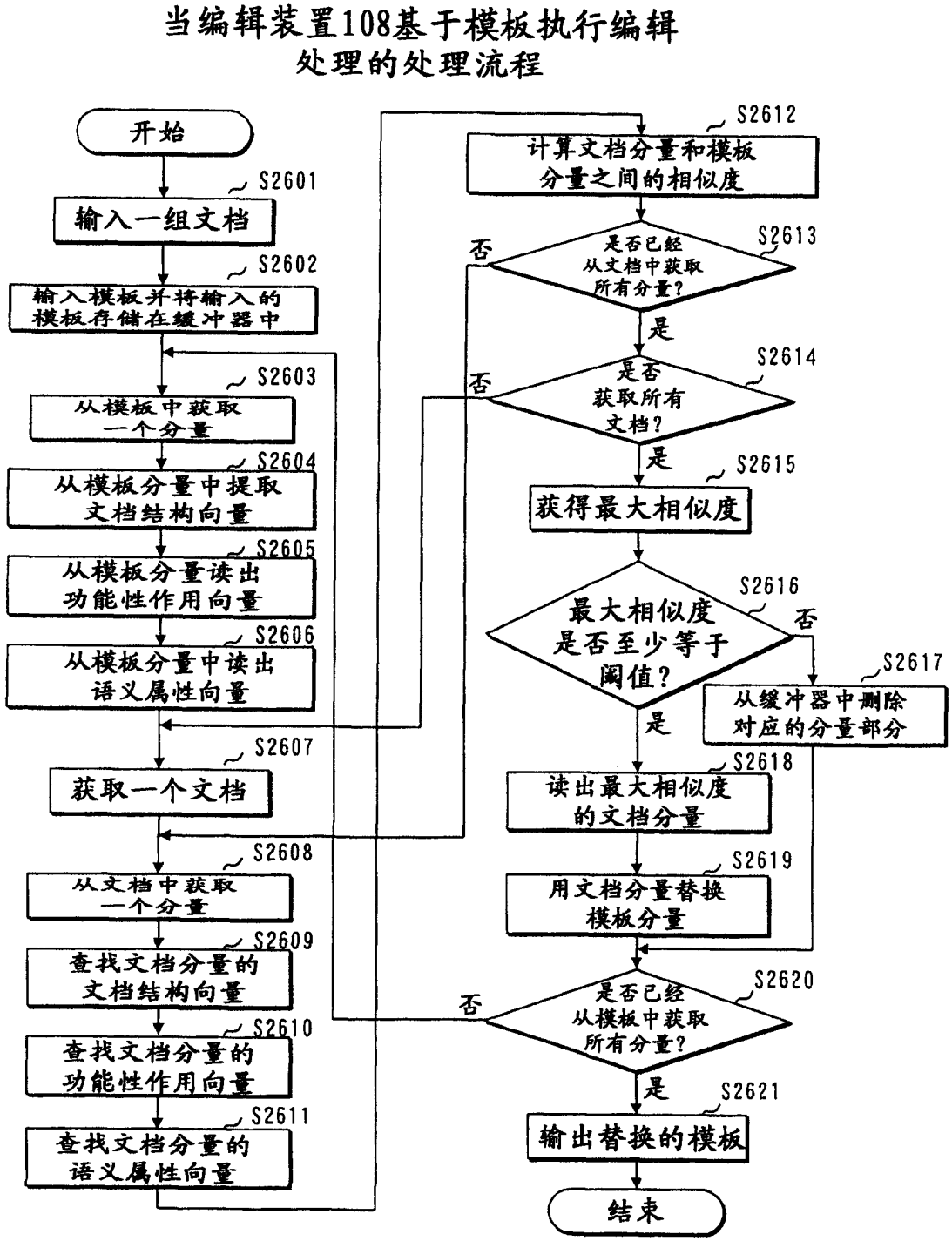

当编辑装置108基于模板执行编辑处理的处理流程
开始
S2601
输入一组文档
S2602
输入模板并将输入的模板存储在缓冲器中
S2603
从模板中获取一个分量
S2604
从模板分量中提取文档结构向量
S2605
从模板分量读出功能性作用向量
S2606
从模板分量中读出语义属性向量
S2607
获取一个文档
S2608
从文档中获取一个分量
S2609
查找文档分量的文档结构向量
S2610
查找文档分量的功能性作用向量
S2611
查找文档分量的语义属性向量
S2612
计算文档分量和模板分量之间的相似度
S2613
是否已经从文档中获取所有分量?
否
是
S2614
是否获取所有文档?
否
是
S2615
获得最大相似度
S2616
最大相似度是否至少等于阈值?
否
是
S2617
从缓冲器中删除对应的分量部分
S2618
读出最大相似度的文档分量
S2619
用文档分量替换模板分量
S2620
是否已经从模板中获取所有分量?
否
是
S2621
输出替换的模板
结束

图26

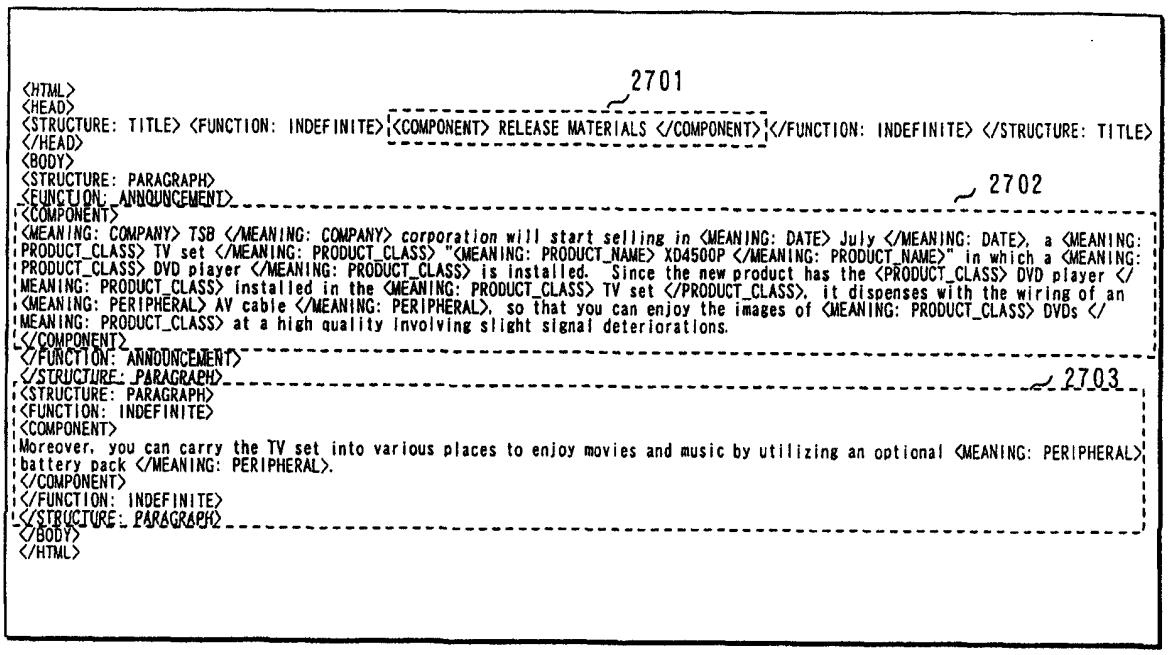

<HTML>
<HEAD>
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE> <COMPONENT> RELEASE MATERIALS </COMPONENT> </FUNCTION: INDEFINITE> </STRUCTURE: TITLE>
</HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
<FUNCTION: ANNOUNCEMENT>
<COMPONENT>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation will start selling in <MEANING: DATE> July </MEANING: DATE>, a <MEANING: PRODUCT_CLASS> TV set </MEANING: PRODUCT_CLASS> "<MEANING: PRODUCT_NAME> XD4500P </MEANING: PRODUCT_NAME>" in which a <MEANING: PRODUCT_CLASS> DVD player </MEANING: PRODUCT_CLASS> is installed. Since the new product has the <PRODUCT_CLASS> DVD player </MEANING: PRODUCT_CLASS> installed in the <MEANING: PRODUCT_CLASS> TV set </PRODUCT_CLASS>, it dispenses with the wiring of an <MEANING: PERIPHERAL> AV cable </MEANING: PERIPHERAL>, so that you can enjoy the images of <MEANING: PRODUCT_CLASS> DVDs </MEANING: PRODUCT_CLASS> at a high quality involving slight signal deteriorations.
</COMPONENT>
</FUNCTION: ANNOUNCEMENT>
</STRUCTURE: PARAGRAPH>
<STRUCTURE: PARAGRAPH>
<FUNCTION: INDEFINITE>
<COMPONENT>
Moreover, you can carry the TV set into various places to enjoy movies and music by utilizing an optional <MEANING: PERIPHERAL> battery pack </MEANING: PERIPHERAL>.
</COMPONENT>
</FUNCTION: INDEFINITE>
</STRUCTURE: PARAGRAPH>
</BODY>
</HTML>
2701
2702
2703

图27A

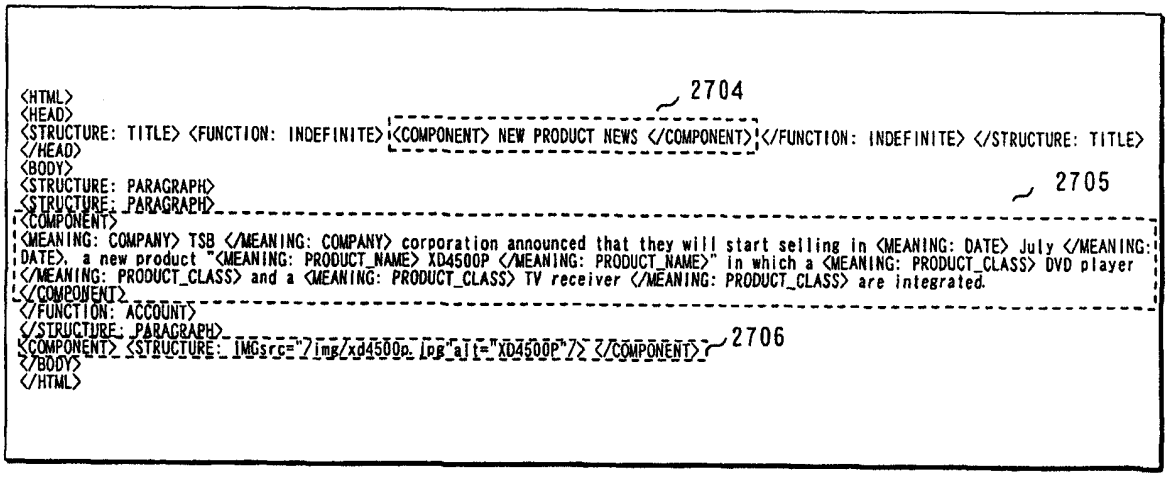

<HTML>
<HEAD>
<STRUCTURE: TITLE> <FUNCTION: INDEFINITE> <COMPONENT> NEW PRODUCT NEWS </COMPONENT> </FUNCTION: INDEFINITE> </STRUCTURE: TITLE>
</HEAD>
<BODY>
<STRUCTURE: PARAGRAPH>
<STRUCTURE: PARAGRAPH>
<COMPONENT>
<MEANING: COMPANY> TSB </MEANING: COMPANY> corporation announced that they will start selling in <MEANING: DATE> July </MEANING: DATE>, a new product "<MEANING: PRODUCT_NAME> XD4500P </MEANING: PRODUCT_NAME>" in which a <MEANING: PRODUCT_CLASS> DVD player </MEANING: PRODUCT_CLASS> and a <MEANING: PRODUCT_CLASS> TV receiver </MEANING: PRODUCT_CLASS> are integrated.
</COMPONENT>
</FUNCTION: ACCOUNT>
</STRUCTURE: PARAGRAPH>
<COMPONENT> <STRUCTURE: IMG src="/img/xd4500p.jpg" alt="XD4500P" /> </COMPONENT>
</BODY>
</HTML>
2704
2705
2706

图27B

```
<scrapbook>
<COMPONENT>
<MEANING: COMPANY> TSB corporation </MEANING: COMPANY> will start selling in <MEANING: DATE> July </MEANING: DATE>, a
<MEANING: PRODUCT_CLASS> TV set </MEANING: PRODUCT_CLASS> "<MEANING: PRODUCT_NAME> XD4500P </MEANING: PRODUCT_NAME>" in
which a <MEANING: PRODUCT_CLASS> DVD player </MEANING: PRODUCT_CLASS> is installed. Since the new product has the
<MEANING: PRODUCT_CLASS> DVD player </MEANING: PRODUCT_CLASS> installed in the <MEANING: PRODUCT_CLASS> TV set </MEANING:
PRODUCT_CLASS>, it dispenses with the wiring of an <MEANING: PERIPHERAL> AV cable </MEANING: PERIPHERAL>, so that you can
enjoy the images of <MEANING: PRODUCT_CLASS> DVDs </MEANING: PRODUCT_CLASS> at a high quality involving slight signal
deteriorations.
</COMPONENT>
<COMPONENT> <STRUCTURE: IMGsrc="/img/xd4500p.jpg"alt="XD4500P"/> </COMPONENT>
</scrapbook>
```

图28A

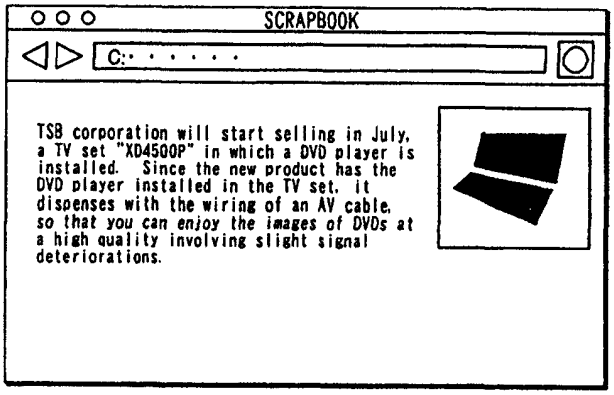


图28B

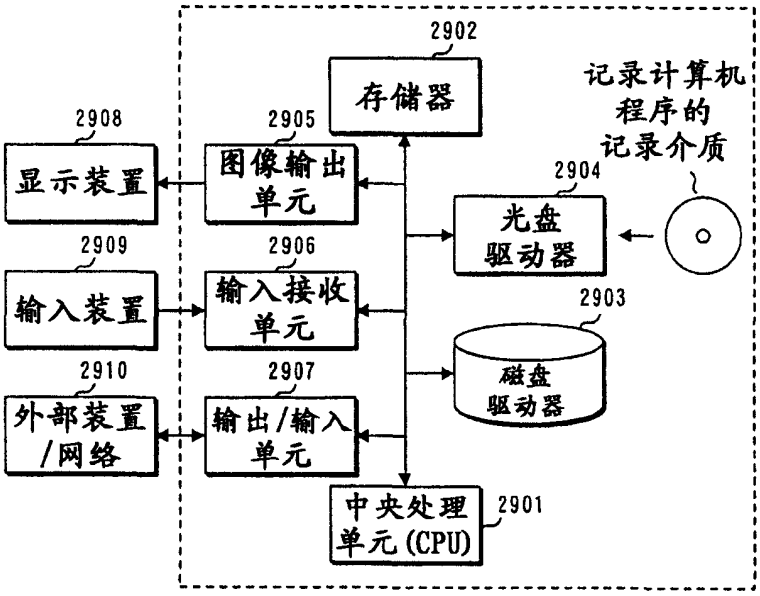

2902
存储器
记录计算机
程序的
记录介质
2908
显示装置
2905
图像输出
单元
2904
光盘
驱动器
2909
输入装置
2906
输入接收
单元
2903
磁盘
驱动器
2910
外部装置
/网络
2907
输出/输入
单元
中央处理
单元(CPU)
2901

图29