(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 102855244 B
(45)授权公告日 2015.02.25

(21)申请号 201110179986.X

(22)申请日 2011.06.28

(73)专利权人 北大方正集团有限公司
地址 100871 北京市海淀区成府路298号方正大厦5层
专利权人 北京北大方正电子有限公司

(72)发明人 缪萍

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/22(2006.01)

(56)对比文件
CN 101436185 A,2009.05.20,
CN 102103573 A,2011.06.22,
CN 102103605 A,2011.06.22,
CN 1687926 A,2005.10.26,
周为.“B/S模式下的Web应用中基于XML生成PDF机制的研究与实现”.《上海师范大学硕士学位论文》.2008,1-6.
未知.“方正经典出版资源加工系统”.《(互联网网页)http://www.founder.com.cn/cn/fouser_product/2010-02/10/content_13167.htm》.2010,1-3.

审查员 马晓宇

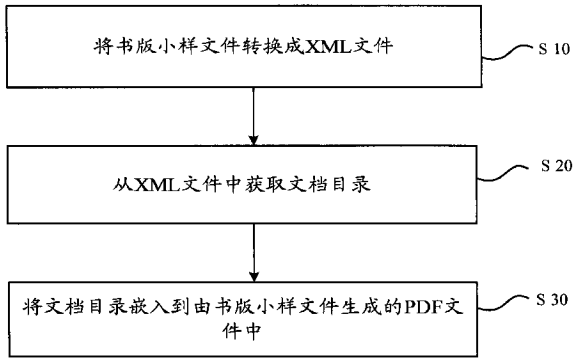
权利要求书2页 说明书5页 附图1页

(54)发明名称

文档目录处理方法和装置

(57)摘要

本发明提供了一种文档目录处理方法,包括:将书版小样文件转换成XML文件;从XML文件中获取文档目录;将文档目录嵌入到由书版小样文件生成的PDF文件中。本发明提供了一种文档目录处理装置,包括:转换模块,用于将书版小样文件转换成XML文件;获取模块,用于从XML文件中获取文档目录;嵌入模块,用于将文档目录嵌入到由书版小样文件生成的PDF文件中。本发明达到了提高目录处理效率的效果。



CN 102855244 B

1. 一种文档目录处理方法，其特征在于，包括：

将书版小样文件转换成 XML 文件；

从所述 XML 文件中获取文档目录；

将所述文档目录嵌入到由所述书版小样文件生成的 PDF 文件中；

其中，将书版小样文件转换成 XML 文件包括：

在由所述书版小样文件进行二扫排版生成书版大样文件的过程中，每生成一页大样内容，将其页号及其内容对应于所述书版小样文件中的起始、终止位置记录到临时文件中；

根据所述书版小样文件的注解，将所述书版小样文件的内容分成多个段；

确定每段内容在所述书版小样文件中的先后顺序；

确定每段内容中所用到的字体、字号及其作用范围；

根据所述临时文件，确定每段内容所在的页号；

将以上确定内容写入所述 XML 文件。

2. 根据权利要求 1 所述的方法，其特征在于，从所述 XML 文件中获取文档目录包括：

预先创建匹配规则；

对所述 XML 文件运行所述匹配规则，以获取所述文档目录。

3. 根据权利要求 2 所述的方法，其特征在于，预先创建匹配规则包括以下至少之一：

将所述书版小样文件中的标题注解加入到所述匹配规则中；

规定字体和字号组合表示章节标题；

规定序号表示章节标题；

接受用户自定义。

4. 根据权利要求 1 所述的方法，其特征在于，将所述文档目录嵌入到由所述书版小样文件生成的 PDF 文件中包括：

将提取的所述文档目录的章节目录内容、层次结构和页号按照书签格式嵌入到所述 PDF 文件中。

5. 一种文档目录处理装置，其特征在于，包括：

转换模块，用于将书版小样文件转换成 XML 文件；

获取模块，用于从所述 XML 文件中获取文档目录；

嵌入模块，用于将所述文档目录嵌入到由所述书版小样文件生成的 PDF 文件中；

其中，所述转换模块包括：

临时记录模块，用于在由所述书版小样文件进行二扫排版生成书版大样文件的过程中，每生成一页大样内容，将其页号及其内容对应于所述书版小样文件中的起始、终止位置记录到临时文件中；

分段模块，用于根据所述书版小样文件的注解，将所述书版小样文件的内容分成多个段；

顺序模块，用于确定每段内容在所述书版小样文件中的先后顺序；

字体字号模块，用于确定每段内容中所用到的字体、字号及其作用范围；

页号模块，用于根据所述临时文件，确定每段内容所在的页号；

写入模块，用于将以上确定内容写入所述 XML 文件。

6. 根据权利要求 5 所述的装置，其特征在于，所述获取模块用于对所述 XML 文件运行预

先创建的匹配规则，以获取所述文档目录。

7. 根据权利要求 6 所述的装置，其特征在于，预先创建匹配规则包括以下至少之一：

将所述书版小样文件中的标题注解加入到所述匹配规则中；

规定字体和字号组合表示章节标题；

规定序号表示章节标题；

接受用户自定义。

8. 根据权利要求 5 所述的装置，其特征在于，所述嵌入模块将提取的所述文档目录的章节目录内容、层次结构和页号按照书签格式嵌入到所述 PDF 文件中。

文档目录处理方法和装置

技术领域

[0001] 本发明涉及数字排版领域，具体而言，涉及文档目录处理方法和装置。

背景技术

[0002] 常用的书版排版软件例如方正书版排版软件是一个流式处理排版软件，目前最新的版本为书版 2008 版，它采用 BD 注解描述排版内容的格式和样式，通过二扫解析、排版内容，形成页面描述文件展现排版结果，具有排版速度快、效率高、排版标准等特点，适合各类教材、教辅、辞书、公文排版，在各个出版社、排版中心和数字加工中心得到了广泛的使用，其排版结果已经成为出版行业的排版标准。

[0003] 随着信息化建设的加速，越来越多的出版社开始重视原始资源的多介质形式发布。目前国内出版社多数是以书版软件作为主要排版软件，因此都保留着大量的书版文件，出版社除了传统的纸质印刷之外，还输出成 PDF(Portable Document Format，便携文档格式）文件，以实现网络营销。

[0004] 书版文件可以以书版小样文件的形式存在，书版小样文件是指包含书版软件规定的 BD 语言注解信息的文本文件，这些 BD 注解描述了后续小样内容的排版属性和排版格式，暗藏着小样的章节目录。

[0005] 目前书版小样文件输出的 PDF 文件是没有章节目录书签的，用户在浏览这些 PDF 文件时需要手动翻转到指定的页进行阅读，当 PDF 文件比较大时，翻转很不方便。为了解决这个问题，目前采用的方法是在书版小样文件输出成 PDF 文件后，在 PDF 文件中手动添加章节目录书签：首先通过人工查看 PDF 文件，找出所有的章节目录内容和所有的页号；其次再把这些章节目录内容及页号信息作为书签嵌入到 PDF 文件中。这种操作效率较低，工作量较大，而且较易出错。

发明内容

[0006] 本发明旨在提供一种文档目录处理方法和装置，以解决现有技术目录处理效率较低的问题。

[0007] 在本发明的实施例中，提供了一种文档目录处理方法，包括：将书版小样文件转换成 XML 文件；从 XML 文件中获取文档目录；将文档目录嵌入到由书版小样文件生成的 PDF 文件中。其中，将书版小样文件转换成 XML 文件包括：在由所述书版小样文件进行二扫排版生成书版大样文件的过程中，每生成一页大样内容，将其页号及其内容对应于所述书版小样文件中的起始、终止位置记录到临时文件中；根据所述书版小样文件的注解，将所述书版小样文件的内容分成多个段；确定每段内容在所述书版小样文件中的先后顺序；确定每段内容中所用到的字体、字号及其作用范围；根据所述临时文件，确定每段内容所在的页号；将以上确定内容写入所述 XML 文件。

[0008] 在本发明的实施例中，提供了一种文档目录处理装置，包括：转换模块，用于将书版小样文件转换成 XML 文件；获取模块，用于从 XML 文件中获取文档目录；嵌入模块，用于

将文档目录嵌入到由书版小样文件生成的PDF文件中。其中，所述转换模块包括：临时记录模块，用于在由所述书版小样文件进行二扫排版生成书版大样文件的过程中，每生成一页大样内容，将其页号及其内容对应于所述书版小样文件中的起始、终止位置记录到临时文件中；分段模块，用于根据所述书版小样文件的注解，将所述书版小样文件的内容分成多个段；顺序模块，用于确定每段内容在所述书版小样文件中的先后顺序；字体字号模块，用于确定每段内容中所用到的字体、字号及其作用范围；页号模块，用于根据所述临时文件，确定每段内容所在的页号；写入模块，用于将以上确定内容写入所述XML文件。

[0009] 本发明实施例的文档目录处理方法和装置，因为采用XML文件来获取文档目录，所以克服了现有技术目录处理效率较低的问题，达到了提高目录处理效率的效果。

附图说明

[0010] 此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

[0011] 图1示出了根据本发明实施例的文档目录处理方法的流程图；

[0012] 图2示出了根据本发明实施例的文档目录处理装置的示意图。

具体实施方式

[0013] 下面将参考附图并结合实施例，来详细说明本发明。

[0014] 图1示出了根据本发明实施例的文档目录处理方法的流程图，包括：

[0015] 步骤S10，将书版小样文件转换成XML文件；

[0016] 步骤S20，从XML文件中获取文档目录；

[0017] 步骤S30，将文档目录嵌入到由书版小样文件生成的PDF文件中。

[0018] 现有技术因为无法从书版小样文件中直接确定文档目录，所以只能通过人工方式向PDF文件中添加文档目录。而本实施例利用XML文件可以记录书版小样文件的结构化信息的特点，从XML文件中获取目录信息，从而自动向PDF文件添加文档目录，这提高了生成文档目录的效率，而且能减少差错。

[0019] 优选地，步骤S10包括：

[0020] 在由书版小样文件进行二扫排版生成书版大样文件的过程中，每生成一页大样内容，将其页号及其内容对应于书版小样文件中的起始、终止位置记录到临时文件中；

[0021] 根据书版小样文件的注解，将书版小样文件的内容分成多个段；

[0022] 确定每段内容在书版小样文件中的先后顺序；

[0023] 确定每段内容中所用到的字体、字号及其作用范围；

[0024] 根据临时文件，确定每段内容所在的页号；

[0025] 将以上确定内容写入XML文件。

[0026] 书版软件采用BD语言编写的书版小样文件不是纯文本文件，其中除了文本内容，还包含各种注解。本优选实施例利用这些注解分析排版信息，从而可以记录到XML文件中。

[0027] 具体来说，步骤S10包括：对小样进行排版处理，依次输出书版大样文件和PDF文件，并且在生成书版大样文件过程（例如进行二扫排版）中，每生成一页大样内容，记录该大样页内容对应于书版小样文件中起始、终止位置及页号信息，当输出完所有小样内容后，

把这些信息记录到一个临时文件中。首先对书版小样文件注解进行分析，根据各个注解排版属性、注解最终排版效果是否独立成段、注解作用范围及它在书版小样文件中的位置，对小样内容进行切分，形成一段段内容；其次完成对整个小样的分析后，依据每段内容在书版小样文件中的先后顺序，把所有切分出来的段按顺序组织起来；再次对每段内容中所用到的字体、字号注解进行解析处理，提取出本段内容所用到的所有字体和字号属性及它的作用范围；再次根据生成的内容与页号关系临时文件，按顺序与切分出来的每段内容进行比较，确定每段内容排版所在的页；最后把经过上述加工操作后的内容输出出来，形成小样结构化的逻辑 XML 文件。

[0028] 优选地，步骤 S20 包括：预先创建匹配规则；对 XML 文件运行匹配规则，以获取文档目录。该匹配规则可以是正则表达式。本优选实施例很容易通过计算机编程来实现。

[0029] 优选地，预先创建匹配规则包括以下至少之一：将书版小样文件中的标题注解加入到匹配规则中；规定字体和字号组合表示章节标题；规定序号表示章节标题；接受用户自定义。章节标题通常具有与其他段落内容不同的排版格式或者特定的内容，具体来说，可以包括以下几步：

[0030] 1、提供一个匹配规则设置工具，在通过此工具进行规则设置时，如果书版小样文件中包含了标题注解，即书版小样文件中已经指明了哪些内容排版成章节内容，则在该匹配规则设置界面中默认增加“大纲提取”规则，它表示把书版小样文件中标题注解内容提取成章节内容。

[0031] 2、在匹配规则设置工具中还提供其它三种章节提取规则供用户选择，分别是一：字体、字号提取规则；二：序号提取规则；三：自定义提取规则；这三种规则概括了目前书籍出版物所有通用的章节目录排版方式。

[0032] 1) 字体、字号提取规则：指出什么样的字体和字号组合表示第几级章节标题，比如用户小样常采用一号黑体表示一级章节，即篇；二号黑体表示二级章节，即章；三号宋体表示三级章节，即节等等。

[0033] 2) 序号提取规则：有些小样采用诸如“一”、“二”、“三”或“一、”、“二、”、“三、”或“（一）”、“（二）”、“（三）”等等序号方式表示章节，通过设置各种序号方式对应的章节级别对小样进行章节目录进行提取。

[0034] 3) 自定义提取规则：有些小样采用诸如“第一章 XXX”、“第二章 XXX”或“第一节 XXX”、“第二节 XXX”或“章一：XXX”、“章二：XXX”等等方式表示章节，通过用户自定义规则，描述出各种类型章节组合提取规则对小样章节目录进行提取。

[0035] 在设置好匹配规则后，设置工具会输出形成匹配规则文件，用户还可以利用设置工具再打开该匹配规则文件，对上次设置的规则进行修改、完善。

[0036] 接下来，从逻辑 XML 文件中按顺序一段段读取出每段内容及其属性。

[0037] 取一段内容，利用匹配规则文件中每条章节目录提取规则按顺序一条条进行匹配，如果满足其中一条章节提取规则，则把该段内容作为章节目录内容提取出来，并记录下该段内容在 PDF 文件中的页号，同时终止该段的章节目录提取过程。

[0038] 循环读取下一段内容，重复执行上述匹配步骤，对每段内容进行章节目录匹配、提取，直到读取完所有段内容。

[0039] 把所有提取出来的章节目录内容及其页号信息按目录层次结构输出成小样章节

目录 XML 文件，完成小样章节目录提取过程。

[0040] 优选地，步骤 S30 包括：将提取的文档目录的章节目录内容、层次结构和页号按照书签格式嵌入到 PDF 文件中。PDF 文件提供了书签格式用于插入文档目录，本优选实施例与现有的 PDF 软件保持一致。具体来说，在本步骤中，可以利用生成的小样章节目录 XML 文件和由书版小样文件生成的 PDF 文件进行合并，把小样章节目录 XML 文件中章节目录内容、层次结构和页号信息按照 PDF 文件书签格式嵌入到 PDF 文件中，完成 PDF 文件书签添加过程。用户通过点击书签，可以自动跳转到该章节所在的页上进行浏览。同时，通过不同的、详细的匹配规则设置，提取出不同层次级别或详细级别的章节目录信息，并嵌入到小样生成的 PDF 文件中形成不同要求或形式的书签，从而实现在书版小样文件输出 PDF 时动态添加章节目录。

[0041] 本发明的优选实施例通过 PDF 文件输出、小样结构化加工、章节目录匹配规则设置、提取加工过程，可以提取出书版小样文件中完整的章节目录内容及其层次结构、页号，并通过自动嵌入到小样输出的 PDF 文件中实现动态生成 PDF 章节目录书签过程。同时，本发明优选实施例中，用户可设置章节目录内容提取规则，可以实现不同级别、层次结构的章节目录提取，实现根据不同需要动态生成不同要求的 PDF 文件章节目录书签。

[0042] 图 2 示出了根据本发明实施例的文档目录处理装置的示意图，包括：

[0043] 转换模块 10，用于将书版小样文件转换成 XML 文件；

[0044] 获取模块 20，用于从 XML 文件中获取文档目录；

[0045] 嵌入模块 30，用于将文档目录嵌入到由书版小样文件生成的 PDF 文件中。

[0046] 本装置提高了生成文档目录的效率，而且能减少差错。

[0047] 优选地，转换模块 10 包括：

[0048] 临时记录模块，用于在由书版小样文件进行二扫排版生成书版大样文件的过程中，每生成一页大样内容，将其页号及其内容对应于书版小样文件中的起始、终止位置记录到临时文件中；

[0049] 分段模块，用于根据书版小样文件的注解，将书版小样文件的内容分成多个段；

[0050] 顺序模块，用于确定每段内容在书版小样文件中的先后顺序；

[0051] 字体字号模块，用于确定每段内容中所用到的字体、字号及其作用范围；

[0052] 页号模块，用于根据临时文件，确定每段内容所在的页号；

[0053] 写入模块，用于将以上确定内容写入 XML 文件。

[0054] 本优选实施例通过分析书版小样文件的注解得到书版小样文件的结构化信息。

[0055] 优选地，获取模块 20 用于对 XML 文件运行预先创建的匹配规则，以获取文档目录。本优选实施例很容易通过计算机编程来实现。

[0056] 优选地，预先创建匹配规则包括以下至少之一：将书版小样文件中的标题注解加入到匹配规则中；规定字体和字号组合表示章节标题；规定序号表示章节标题；接受用户自定义。利用这些丰富的规则，可以很灵活地从 XML 文件中获取文档目录。

[0057] 优选地，嵌入模块 30 将提取的文档目录的章节目录内容、层次结构和页号按照书签格式嵌入到 PDF 文件中。本优选实施例实现了对 PDF 文件加入文档目录。

[0058] 从以上的描述中可以看出，通过本发明实施例提供的方案，快速、高效、准确、自动地实现了输出 PDF 时动态添加章节目录书签，特别是当书版小样文件中已经明确使用了标

题注解指明了哪些内容为章节目录内容时，整个处理过程不需要人工干预。同时由于章节目录内容的提取自动完成，而且能自动定位到该章节目录在 PDF 文件中的页号，可以保证章节目录内容的正确性，减少了人工手动添加时的错误。

[0059] 显然，本领域的技术人员应该明白，上述的本发明的各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而可以将它们存储在存储装置中由计算装置来执行，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本发明不限制于任何特定的硬件和软件结合。

[0060] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

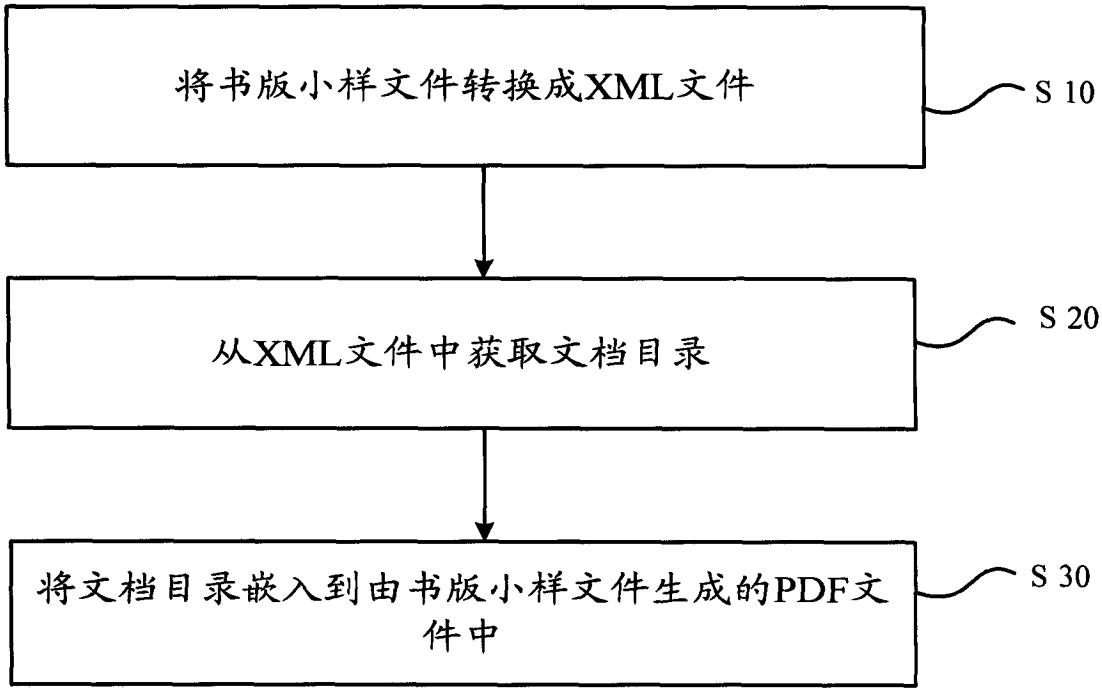

将书版小样文件转换成XML文件
S 10
从XML文件中获取文档目录
S 20
将文档目录嵌入到由书版小样文件生成的PDF文件中
S 30

图 1

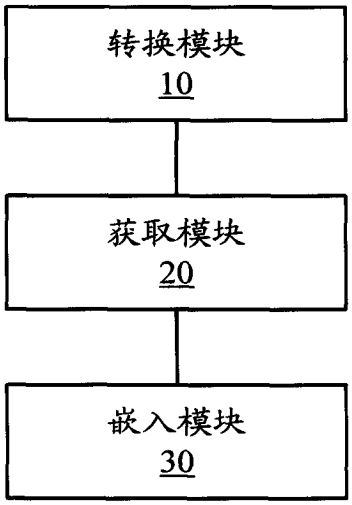

转换模块
10
获取模块
20
嵌入模块
30

图 2