第23卷 特刊 计 算 机 工 程 1997年12月
Vol.23 Special Issue Computer Engineering December 1997

·数据库技术及应用·

结构化彩色图文混排技术的研究

孙永学 夏宽理 李丽燕
(复旦大学计算机科学系 上海 200433)

摘要 该文提出了信息处理领域一类用途广泛的排版——数据库信息排版的需求以及结构化排版的概念。通过一个实际系统的设计对基于数据库的结构化排版系统进行了实现。该系统以自行设计的 LMDL 排版描述语言为核心，具有基于数据库排版的特色。该系统是数据库信息结构化排版的一个成功实例。

关键词 排版 结构化排版 数据库出版 排版描述语言

Research on Structured Color Hybrid Publishing Technology

Sun Yongxue Xia Kuanli Li Liyan
(Department of Computer Science and Engineering, Fudan University Shanghai 200433)

【Abstract】 This paper describes a structural publishing system based on database. It surveyes some typical publishing systems and compares their respective advantages and disadvantages.It also defines requirements of database information publishing and concepts of structural publishing. A system is designed and implemented to satisfy those requirements. The system has a publishing description language—LMDL—as its core, and uses database technology extensively.

【Key words】 Publishing; Structural publishing; Database-based publishing; Publishing description language

排版系统从处理信息的种类来分，可以分为通用排版系统和专用排版系统。流行的通用排版系统有 Aldus 公司的 PageMaker、MicroSoft 公司的 Word，以及北大方正办公排版系统等。专用排版系统有北大方正地图出版系统，简谱排版软件等。综合起来，这些系统具有如下几个特点：*

(1) 文件型——数据以文件形式存储，这意味着数据不能是海量的。

(2) 不支持数据库操作——存储于数据库中的大量数据信息无法以一种简便的方式得到排版的支持。

(3) 手工处理——版面组织靠手工进行，属于半自动化系统。

(4) 版面信息之间非结构化——版面信息之间上下文无关，即非结构的，信息之间的上下级关系、导出关系、绑定关系等复杂关系在这类系统中无表示措施。

而在信息处理领域，出版信息通常具有如下的一些特点：

(1) 信息量大——通常数据量在 GB 到数十 GB 之间。

(2) 有频繁的变更。

(3) 信息之间有复杂的紧密联系——例如上下级关系、导出关系等。

显然，交互式的排版系统无法方便地支持这类信息的排版工作。本文提出了一种新型的排版系统——基于数据库的结构化排版系统，该系统具有以下特点：

(1) 系统的数据输入输出均基于数据库系统，因此系统能够处理大容量的数据。

(2) 系统提供一种可以描述基于数据库信息的出版要求及信息间关联的结构化描述语言 LDML。

1 系统的设计

1.1 结构化排版系统结构

如图 1 所示系统主要由数据编辑子系统、设置子系统、数据分析和处理子系统、编辑输出子系统、数据库接口实现模块以及排版描述语言组成。

数据编辑子系统主要完成初始排版数据的选定、编辑，排版结果的各种检索索引的自动生成等工作。

* 本项目受到上海市科委"九五"攻关项目部分资助
孙永学 男，25 岁，研究生，研究方向：软件工程
收稿日期：1997-09-01

设置子系统提供一个使用排版描述语言的图形用户接口，让用户直观简洁、合法和快速地定义自己对数据库内容的排版要求。

数据分析和处理子系统是排版描述语言的一个解释分析器，通过对采用排版描述语言描述的排版要求和语义进行分析，完成取数据、依赖性数据的生成和提取、基于语义的字符串分裂、格式化、组版，产生包含丰富结构、格式、定位信息的排版结果。

编辑输出子系统提供了一个以所见即所得的方式处理排版结果的工具。该子系统支持排版结果和源数据的同步更新,支持打印输出及满足印刷工业标准的EPS 文件输出。

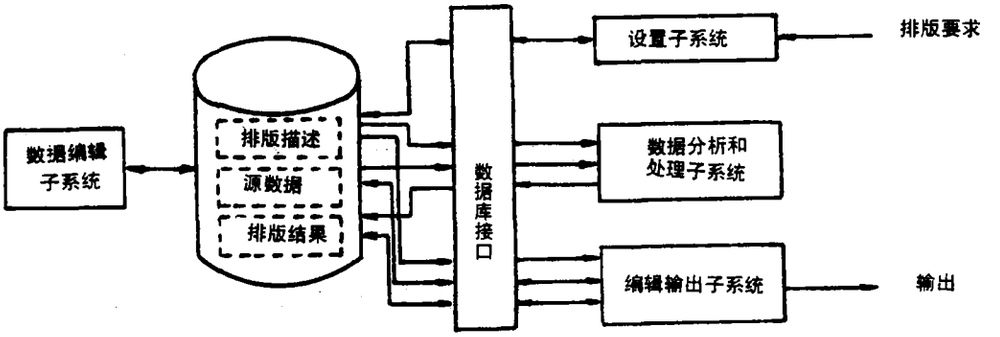

图 1　结构化排版系统结构

1.2 数据模型

一本图书可以按章、节、目进行分类，其文字材料、插图等均可嵌入到相应章节。而对电话号簿、辞典、词典等一类出版物来说，其信息类按行业、词条首字等进行分类，本系统的数据也按分类组织进行，表示为一棵树，如图 2 所示:

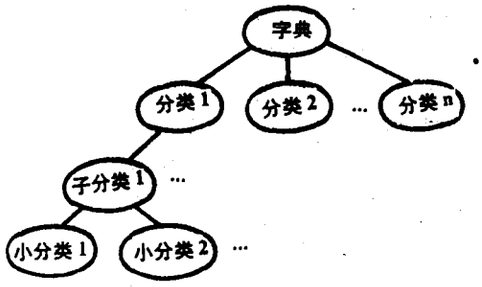

图 2　字典分类例

分类的组织及表示涉及正文、书眉的表示，在描述语言中都有相应的描述设施。数据除了有分类外，本身通常也有层次，也可以表示为一棵树。如一个企业用户的信息由一级用户、二级用户、三级用户、四级用户所组成的一个有层次的树来表示，在数据的每一级表示上均可有任意的数据项，这些用户级别及各数据项组成了一个词条或用户的出版项目，我们可将这一组有关联，排版时上下文相关的数据记录称为一个对象，排版就是对这类信息对象的排列及组合。

1.3 LDML 语言的设计

排版描述语言 LDML 是整个系统实现的基础，语言所具有的描述能力将直接影响到系统的排版处理能力。系统设计的思想为把一组具有相同组织结构和排列格式的对象的上下级关系、排版格式抽象为一个模式，称为对象模式，用 LDML 的记录描述。通过对数据库相联，对每个对象均有指定对象模式与之相对应。从而排版就成为对每个对象所对应的 LMDL 记录的语义进行解释，嵌入对象的数据，自动组合并填入版面。

排版描述语言主要描述的内容(这里所提供的是 LDML 的一个子集)有:

(1) 页面描述——对排版的页面尺寸，边距，页眉、脚、边，页中内容的密度等进行定义。

＜页面设置＞::=＜奇偶标志＞＜页面尺寸＞＜背景颜色编号＞＜顶边距＞＜底边距＞

＜外边距＞＜内边距＞＜顶基线＞＜底基线＞＜边基线＞＜顶边线宽度＞＜底边线宽度＞

＜外边线宽度＞＜内边线宽度＞＜栏数＞＜栏间距＞＜页眉＞＜页脚＞＜页边＞

(2) 分类描述——描述分类类目的刊登规格。

＜分类描述＞::={＜级描述＞}

＜级描述＞::=＜层次号＞{＜行＞}

(3) 对象刊登模式描述——对对象刊登的刊登格式、图片信息的刊登格式的详细描述。

＜对象模式＞::=＜模式号＞＜模式名＞{＜行＞}

＜行＞::=＜层次＞＜最大长度＞＜新行缩进标志＞＜绑定下行标志＞{＜项＞}

＜项＞::=＜文本项＞|＜数据库项＞|＜填充项

>

＜文本项＞::=＜正文＞＜项修饰＞|＜正文＞＜项修饰＞＜条件域＞

＜数据库项＞::=＜表名＞＜字段名＞＜项修饰＞|＜表名＞＜字段名＞＜项修饰＞＜条件域＞

＜填充项＞::=＜填充符＞|＜填充符＞＜条件域＞

＜条件域＞::=＜表名＞＜字段名＞

＜项修饰＞::=＜字体编号＞＜颜色编号＞＜最大长度＞＜分裂规则＞

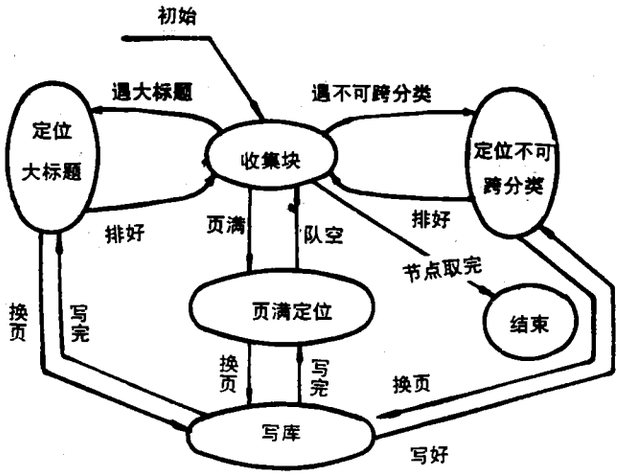

图 3　定位器状态转换图

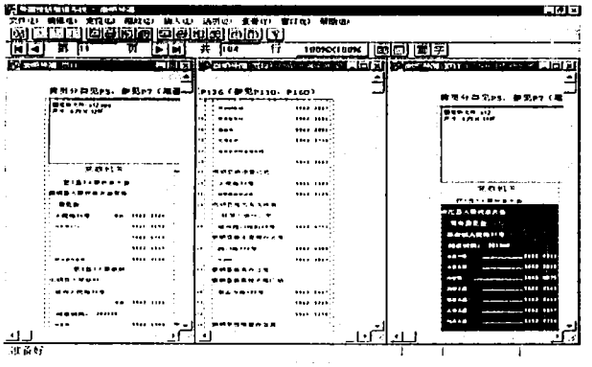

图 4　系统主界面

1.4 核心算法——定位算法

定位器从节点表中读取按照分类等次序排序的节点，恢复出记录块的结构，然后，以块为单位添加至定位器的内容队列中。由于上下文的相关性，定位器通过一个有限状态机来实现。通过定位器在不同状态间的转换完成号簿的定位工作。状态转换图 3。

如图 3 所示，当定位器处于收集状态时，不断读入需处理的块。当队列中未定位的对象尺寸超过页面空间时，定位器转入页满定位状态，当收集到大标题或不可跨分类时定位器转入对它们进行处理的相应定位状态。

1.5 用户界面举例

图 4 所示界面为系统对"上海市崇明电话号簿"排版时的 1 个界面。

2 结束语

本文介绍了基于数据库的结构化排版的有关概念的设计思路，并通过一个电话号簿排版系统的设计和实现对原型系统进行了实际设计，实践证明系统的设计思路是正确的，开发是成功的。

实例系统"天翼结构化排版系统"采用 Virsual C++开发，代码约为 6 万行。该系统目前已先后出版了亚洲第一本彩色号簿—"96'上海黄页号簿"等十多本号簿，在两年内使上海市电话簿公司的营业收入净增 2000 多万元。系统已于 1997 年 1 月通过专家鉴定，被认为"国内领先、达国际 90 年代先进水平"。

我们希望，随着研究的进一步深入，能够使排版描述语言得以标准化和更进一步的扩充，对原有系统的算法等进行改进，使之能够成为数据库信息结构化排版领域的一个强有力的支持工具。

参考文献

1 复旦大学. 天翼结构化彩色号簿排版系统设计文档. 1997-01

2 朱永和，高振魁，沈晓红. 计算机排版技巧及疑难问题解. 中国致公出版社，1995-03

3 王选. 排版系统若干重要方面的发展状况和展望(上、下). 今日印刷，1993-04

4 Mailoli, C. Anchors and Paths in a Hypertext Publishing System. Assoc. Italian Inf. & Calcolo Autom Milan. Italy.1992.146

5 殷步九.陈艺林. 中文版式设计语言文本简介. 北京新华照排研究实验中心. 1982-06

6 Heck A. From an Early Electronic-publishing Concept towards an Advanced Electronic Information Handling. http://cdsweb.ustrasbg.fr/heck