340-344

基于 SGML 的光盘文档库的数据组织*

武港山 李景春 张福炎

(南京大学多媒体计算机研究所 南京 210093)

摘要 本文设计了一种通用的光盘文档库数据组织,并在此基础上实现了基于 SGML 的光盘文档库生成系统,该系统利用通用的文档转换工具把 SGML(standard generalized markup language)文档转换成光盘文档的格式,并对其建立目录、索引等浏览机制以方便用户浏览和检索.光盘文档的数据组织充分考虑了光盘存储器的特点,提高了浏览与检索操作的效率.

关键词 SGML,CD-ROM,文档库.

SGML(standard generalized markup language)是用于描述结构化文档的一种国际标准通用置标语言.[1]基于 SGML 规范的文档现在已越来越广泛的得到采用,目前在 INTERNET 上非常流行的超文本置标语言(HTML)就是 SGML 标准的一种应用.[2]SGML 作为中间描述语言可用于各种类型的电子出版,它特别适合于文档库的文档描述.

随着多媒体技术的飞速发展,各类电子出版物大量涌现,其中 CD-ROM 光盘以其体积小、容量大、成本低而成为广泛使用的信息载体.本文提出一种通用的光盘文档数据组织并实现了基于 SGML 的光盘文档库生成及浏览系统,该系统运行于 WINDOWS95 环境,支持多字节文字编码系统.

1 光盘文档库的总体结构

本文考虑的光盘文档库是面向浏览器的可伸缩分布式结构,它可以按照浏览操作的需要从一张光盘上的文档库扩展到整个 Internet/Intranet 网上所有的文档库.

1.1 文档库的组成

文档库由成千上万的超文本文档组成,这些超文本可能分布在不同地域的不同的物理存储介质上,为了浏览的方便,系统在文档库和文档之间增加一个子文档库的概念.子文档库由文档的本体和库描述信息组成.因此对于浏览器来说,其最大的浏览单位是子文档库.如图 1 所示.其中每一个子文档库都有一个标识,对于光盘来说它是光盘的卷标,对于文件系统来说它是目录路径名,而对于网络而言它可以是一个网络文件地址.

* 本文研究得到江苏省自然科学基金资助.作者武港山,1967 年生,讲师,主要研究领域为计算机图形学,多媒体技术.李景春,1972 年生,硕士生,主要研究领域为多媒体技术.张福炎,1939 年生,教授,博士导师,主要研究领域为多媒体技术,计算机图形学.

本文通讯联系人:武港山,南京 210093,南京大学多媒体计算机研究所

本文 1997-03-18 收到修改稿

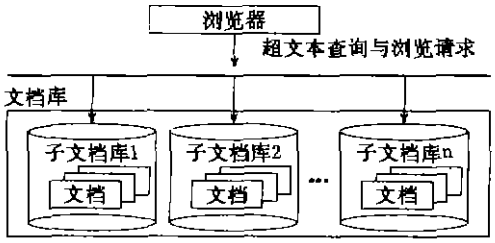

图1 光盘文档库的组成

对文档库中的文档浏览请求由3个部分组成：

〈文档浏览请求〉::=〈子文档库标识〉〈文档标题〉[〈开始数据块号〉〈数据块数〉]

〈子文档库标识〉::=〈光盘的卷标〉|〈目录路径名〉|〈URL〉

1.2 子文档库的数据组织

子文档库中包含库管理信息以及各超文本文档，其数据组织描述如下：

〈文档库〉::={〈子文档库〉}

〈子文档库〉::=〈库管理信息〉{〈超文本文档〉}

〈库管理信息〉::={〈超文本文档标题〉〈文档位置信息〉}

〈文档位置信息〉::=〈文档文件所在的目录名〉〈光盘上文档文件开始数据块号〉

〈超文本文档〉::=〈文档管理信息〉{〈文档数据〉}

〈文档管理信息〉::=〈管理信息头〉{〈文档数据种类标志〉〈开始数据块号〉〈数据块大小〉}

〈文档数据〉::=〈文档正文数据〉|〈文档索引数据〉|〈文档目录数据〉|〈文档媒体数据〉

〈文档正文数据〉::={〈文档表示单元〉}

〈文档表示单元〉::=〈文档文字数据〉|[〈文档表示标志开始〉〈文档表示单元〉〈文档表示标志结束〉]

〈文档表示标志〉::=〈段落标志〉|〈版面标志〉|〈文字修饰标志〉|〈媒体参照标志〉|〈超链标志〉

〈超链标志〉::=〈超链标志头〉〈文档正文数据中的地址〉|〈超链标志头〉〈本子文档库中的文档标题〉|
〈超链标志头〉〈子文档库标识〉〈文档标题〉

〈文档目录数据〉::={〈目录层次〉〈标题〉〈文档正文数据中的开始地址〉}

〈文档索引数据〉::=〈文档关键词索引数据〉|〈文档全文检索数据〉|〈文档媒体参照索引数据〉

〈文档媒体数据〉::={〈媒体描述头数据〉〈媒体本体数据〉}

子文档库可以包含多个超文本文档，这些文档之间可以通过超链相互连接，根据上述定义，不同子文档库之间也可以连接. 子文档库中的每个超文本文档由多种数据描述块组成，分别表示文档数据的正文、索引、以及媒体等数据，这些数据块相互之间存在参照关系，如图2所示.

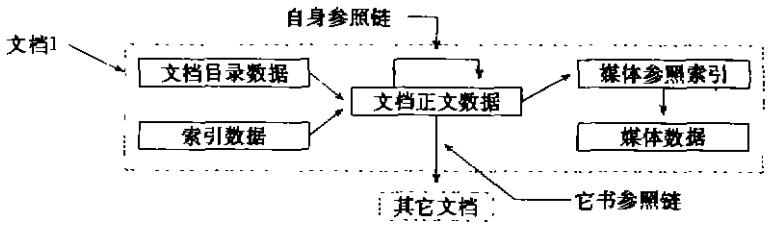

图2 文档中各数据块关系图

2 光盘文档库数据组织

设计以 CD-ROM 光盘为存储介质的文档库时，必须充分考虑光盘的数据存取特点以提高光盘文档的适用性.

2.1 光盘文档库的文件系统

为了既能满足软浏览器又能兼容专用硬件播放器，本系统中的光盘数据组织设计了两套数据查询方案：(1)基于 ISO9660 的文件查询；(2)基于光盘绝对地址的查询. 基于文件系统的查询可以保证光盘文档能够被目前的计算机系统所识别. 基于绝对地址的查询使得光盘文档能够被专用的硬件播放器所接受，它可以避开文件系统直接搜索到光盘文档的内容.

光盘文档的文件目录树形结构如图 3 所示.

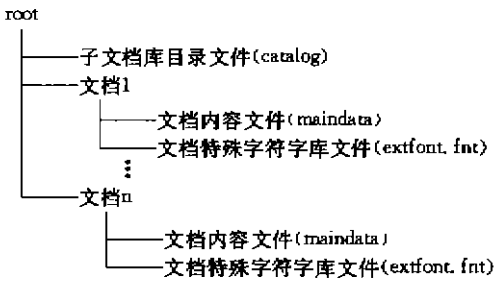

图3 光盘出版物的目录结构

光盘作为一个子文档库，其文档目录文件中存放着子文档库的管理信息，其中包括各文档的标题、所在的目录名及其在光盘上的绝对地址. 文档目录文件总是存放在光盘的某个固定位置上，以便硬播放器能直接找到.

相应文档目录名下的 maindata 文件是文档的本体，超文本的大多数数据块都集中在这个文件里，以方便硬播放器的连续读取. 由于光盘文档数据的存放受光盘的黄皮书规范制约，因而某些特殊的媒体数据如 CD-DA 音频数据必须存放在光盘的其他区域，光盘文档中仅记录其位置信息.

2.2 光盘文档的数据组织

maindata 文件主要由以下几个部分组成：文档目录、文档正文、关键词、基于关键词的索引、全文检索索引、媒体参照索引、媒体数据. 这些数据都是按光盘的物理数据存储单位块(BLOCK，块的大小为 2 048 字节)进行组织的，即数据的大小都是块的整数倍，而所有这些数据在 maindata 文件中的组织又是由一个称为文档管理块数据结构来控制. 其结构如图 4 所示. 其中正文数据中记录了加标志的实际文档数据的正文，这些标志包括文档结构定义、字形修饰、媒体参照等内容. 光盘文档中的标志主要是面向浏览而设计的，尽量体现源 SGML 文档的逻辑结构，如光盘文档中的标志可以表示出文档的段落章节.

检索数据由基于关键词检索和全文检索两种数据组成，全文检索数据只有一种，而基于关键词的检索数据可以有多种，检索时允许用户选择其中一种或几种参与查询. 全文检索采用基于字表的查询方法.

文档目录数据是浏览器的重要导航数据，它记录了光盘文档的逻辑结构，用户可以通过选择目录来浏览文档中相应的内容.

媒体数据由媒体参照索引数据和媒体数据组成，媒体参照索引中记录了文档中所有的媒体参照数据，本系统中的媒体参照源有 4 种描述形式：本地本书、本地它书、外部它书和外部非书类. 参照的媒体可以是系统类媒体(如常见的图形图象格式以及音视频动画格式)和用户自定义媒体格式，后者用户必须提供播放该媒体的 DLL 模块.

光盘文档数据采用上述的结构主要出于以下考虑：

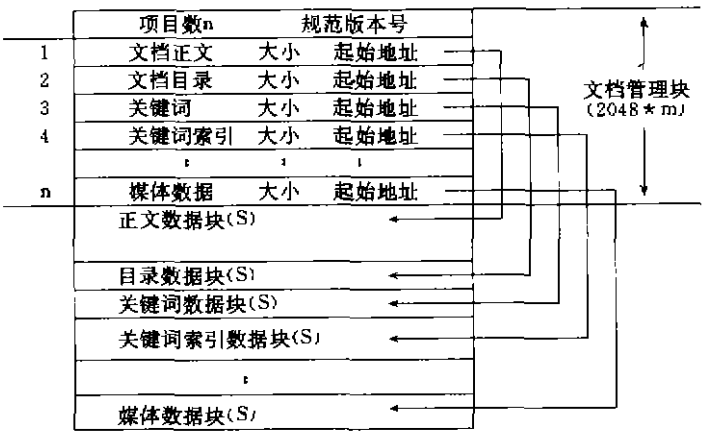

图4 光盘文档数据组成

首先，光盘中的数据存储不同常见的磁介质的数据存储，由于一次烧制的只读特性使得ISO9660规范中的数据存储在物理位置上都是连续的，且文件以块为单位对齐，因此光盘文档的各数据块集中存放并以起始地址和块数为定位手段，甚至寻址地址也采用块号和块内偏移的方式.

其次，文档内的各数据块集中管理存放有助于减轻浏览器的文件操作负担，而且特别适合于简化的浏览器使用，如硬件播放器等. 这种方式使得光盘文档在一定程度上摆脱了文件系统的束缚，增加了适应性.

最后，文档数据中的文档描述形式采用类SGML的形式，可以方便的把SGML文档转换成光盘电子文档，使之变成一个适合于浏览操作的文档格式，同时也可以方便地把光盘文档转换成HTML格式，供WWW服务器使用.

2.3 光盘文档库的文档作成

上述的光盘文档可以开发专门的文档编辑与转换工具来生成，本系统中采用SGML文档的通用转换系统生成光盘文档.

SGML规范是结构化文档的通用描述语言，也是文档内容保存得最为全面的一种描述形式，它不仅可以描述文档的结构，还可以描述其属性，因而特别适合于作为文档库的描述语言. 通过转换系统用户可以对其进行各种处理，如本文所述的光盘出版，转换成页面描述形式可以用于印刷出版，转换成HTML格式适用于INTERNET网络web服务器.

SGML文档通用转换系统是本系统中开发的面向一般文档转换要求的系统，用户的文档转换要求用转换描述语言来记录，用户可以用它描述SGML文档中要处理的标志以及处理方式等信息，转换描述语言由转换系统来解释、执行. 由于转换信息独立于转换系统，因此，用户可自由地设计各种不同的处理以适应不同的文档转换要求.

SGML文档到光盘文档格式的转换就是使用通用的转换系统完成的. 转换过程中用户可以将特定的文字作为关键词抽取出来，以便生成基于关键词的检索数据，同时利用转换系统中能够嵌入用户私有处理的功能，在转换处理过程中直接生成关键词的检索数据和全文检索数据. 图5给出了从SGML文档转换到光盘文档的处理过程.

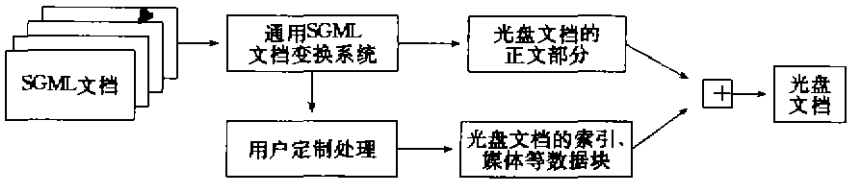

图5 从SGML文档到光盘文档的处理流程

4 结束语

应用本系统中的转换工具，我们成功地把一些 SGML 和 HTML 文档转换成光盘文档的描述形式，并对其建立了索引等机制，给浏览带来了很多方便. 光盘文档的应用在国外已很普遍，我国尚处于开始阶段，使用超文本、超媒体技术的多媒体光盘文档库有大量的社会需求，许多关键技术还需要进行深入的研究和探讨，特别是对基于网络环境的分布式文档库的数据组织和查询技术的研究，具有广泛的应用价值和理论意义.

参考文献

1 ISO 8879-1986, Standard Generalized Markup Language(SGML), 1986.

2 赵或等. HTML 文档规范及其应用实例. 多媒体世界，1996，6：19～21.

DATA ORGANIZATION OF A CD-ROM DOCUMENT LIBRARY BASED ON SGML SPECIFICATION

WU Gangshan LI Jinchun ZHANG Fuyan

(*Multimedia Computing Institute Nanjing University Nanjing* 210093)

Abstract A general document data organization for document library is presented in this paper, and according to it the authors develop a CD-ROM publishing system based on SGML. Using conversion tools SGML documents were transformed into CD-ROM documents on which catalog and index were also created. CD-ROM document organization takes advantages of CD-ROM data storage, so as to increase the efficiency of both browsing and searching.

Key words SGML, CD-ROM, document library.