超媒体文档库协作写作系统的数据结构设计*

黄宜华　尤晓白　纪　元　杨文清　张福炎

（南京大学计算机科学与技术系　南京　210093）

摘要　本文主要讨论了一个面向大容量超媒体中文文档库的分布协作写作系统的数据和结构模型设计，着重介绍了分布式文档库、文档目录树、节点和超链链表、多媒体对象、文档检索索引等数据结构和整个超媒体系统的结构模型.

关键词　分布，协作，超媒体，文档库，数据结构.

(1)超媒体技术应用及国内外现状

近年来，多媒体电子出版迅速兴起，可望成为一个有巨大发展前景的电子信息技术产业. 大量多媒体电子出版物，尤其是电子图书、文档资料库类型产品的开发，促进了超媒体写作应用软件技术和系统的发展.

一个超媒体写作系统实际上就是一个超媒体信息的创建、组织、管理和展现系统，主要完成节点和链的输入编辑、组织建立、阅读展现、检索查询和管理等功能.

超媒体系统从早期的超文本模型系统，到目前已进入了第 2 代系统. 第 2 代系统的主要特征包括文档具有超链网状结构、节点多媒体化、良好的浏览导航工具、窗口化用户界面、节点和链的编辑管理等.[1]目前市场上推出的许多写作工具，都是一些实用的超媒体写作系统. 如国外推出的 Toolbook，Authorware，Microsoft Viewer 等. 国内也推出了一些比较有影响的面向中文的系统. 各个系统在系统功能、用户界面、多媒体支持、超文本链接、检索查询等方面都各有一定的特色，在多媒体光盘出版物的开发中，发挥了很大的作用.

但是，作为以面向大容量文本为主的中文多媒体光盘出版物的写作工具，现有的这些软件都有不足. 国外的一些系统是以面向多媒体为主的展示型作品的，对文本的组织、处理和检索等功能不足；而以面向文本为主的作品的系统如 Microsoft Viewer，虽然英文文本处理能力很强，但不能很好地兼容中文的处理，如文本阅读时的自动排版显示、英文全文检索等. 而国内推出的系统在整体功能和中文处理能力上也有不足，具有较强的面向大容量多媒体中文文档综合处理能力的系统，尚不多见.

* 本文研究得到国家 863 高科技项目资金和美国 INTEL 公司部分资助. 作者黄宜华，1962 年生，副教授，主要研究领域为多媒体技术，中文信息处理技术. 尤晓白，1971 生，硕士，主要研究领域为多媒体技术. 纪元，1972 年生，硕士，主要研究领域为多媒体技术. 杨文清，1973 年生，研究生，主要研究领域为多媒体技术. 张福炎，1939 年生，教授，主要研究领域为多媒体技术，图形处理技术，中文信息处理.

本文通讯联系人：黄宜华，南京 210093，南京大学计算机科学与技术系

本文 1996-04-03 收到修改稿

更重要的是,新一代超媒体系统还应考虑更强大的功能特性[2],如分布协作写作、超链的自动链接与超文本的自动转换、强大的检索查询等功能.而现有的系统大多是单机系统,难以支持大型作品的协作写作;在超文本链处理上,现有系统大多只能支持最基本的超链编辑功能,处理大量超链时工作量大、速度慢;在面向中文超媒体文档作品的阅读上,还没有一个系统在支持良好的浏览导航的同时,还支持强大的检索查询,尤其是全文检索.

(2)CCHMDOC 的基本设计思想

CCHMDOC 旨在研究实现一个实用化超媒体系统,它具有较全面的第 2 代超媒体系统特征,同时又具有第 3 代的部分功能特征,该系统着重研究 4 个方面的技术问题:中文超文本数据模型及超结构链接技术、超媒体系统中的中文处理技术(尤其是快速检索技术)、分布协作写作技术及多媒体技术,从而形成一个具有 CCHM(cooperation, Chinese, hyperstructure, multimedia)综合特征的超媒体系统.

CCHMDOC 基于 PC 的 Windows for Workgroups 窗口系统平台,提供一个窗口化的全集成交互写作和阅读环境,完成超媒体文档库的结构组织、节点和链的编辑、媒体混合、阅读展现、浏览导航、检索查询和分布协作管理等功能,主要面向以大容量中文文本为主的多媒体文档库作品,适用于大型中文电子图书和文档资料库类型的光盘出版物的制作.

(3)CCHMDOC 的特点

系统具有以下主要特点:

a. 具有超结构的文档组织结构,支持顺序链接、交叉链接功能;

b. 支持大容量中文文档,提供较强的中文文本处理能力,尤其是快速的中文全文检索、关键词检索、节点主题属性检索功能;

c. 支持多媒体混合节点;

d. 图形文档目录提供可视化的目录编辑、管理和阅读查询时的浏览导航;

e. 强大的超文本链接处理技术,支持自动、半自动链接和基本的手工交互链接;支持文档内部、分布文档之间的链接;

f. 支持文档库的分布和协作写作及作品发行前的合成转换;

g. 支持常规文档与超文本文档的相互转换;

h. 支持多种浏览导航手段,防止迷航;

i. 支持动态子文档,可基于手工选择或各种检索结果动态生成子文档;

j. 窗口化的集成交互写作阅读环境,以友好的用户界面,提供强大的目录、节点和链的编辑管理功能和阅读检索功能,并能在写作与阅读层之间联机切换.

1　系统数据模型设计

1.1　主要数据模型

(1) 分布式文档库

文档库是中的 CCHMDOC 最大逻辑存储单位,它是指一类内容相关并具有整体逻辑结构的文档集合.在物理存储结构上,一个文档库可以用多个文件存放.我们可以把属于一个大型文档库中的每个文件视为是内容相对完整的子文档库,将它们分布在网络的多个站点上,由一组人员分别制作不同的子库,通过一定的逻辑连接关系来保持这些子库形成一个

逻辑上完整的大型文档库. 当然必要时(如总成制作光盘前)系统可以将分布完成的文档库合并起来.

为了支持作品的分布协作,系统进一步对文档库引入不同的层次——组文档库、成员文档库和单文档库. 所谓组文档库是指由工作组成员分头负责的所有子文档库构成的、逻辑上完整的总体文档库作品;而每个工作组成员又可以负责1个或多个内容与结构相对完整的子文档库,这些子库构成成员文档库;最终,每个以独立文件存放的子文档库称为单文件文档库. 一个单文档库可以独立工作,并在需要时加入或移出成员文档库,而成员文档库也可以单用户工作形式存在,需要时也可以加入或移出组文档库. 图1表示了1本图书《多媒体软件开发环境与工具》作品的结构内容和相应的分布文档库结构示例. 在该示例中,假设整个作品由4个成员构成的工作组协同写作,成员A到D依次负责作品第1~4篇,成员A为主编,负责整个作品的结构. 图中的每个圆形节点表示一个下属的分布子库文件,而方形节点的内容则直接包含在最近上层子库文件中.

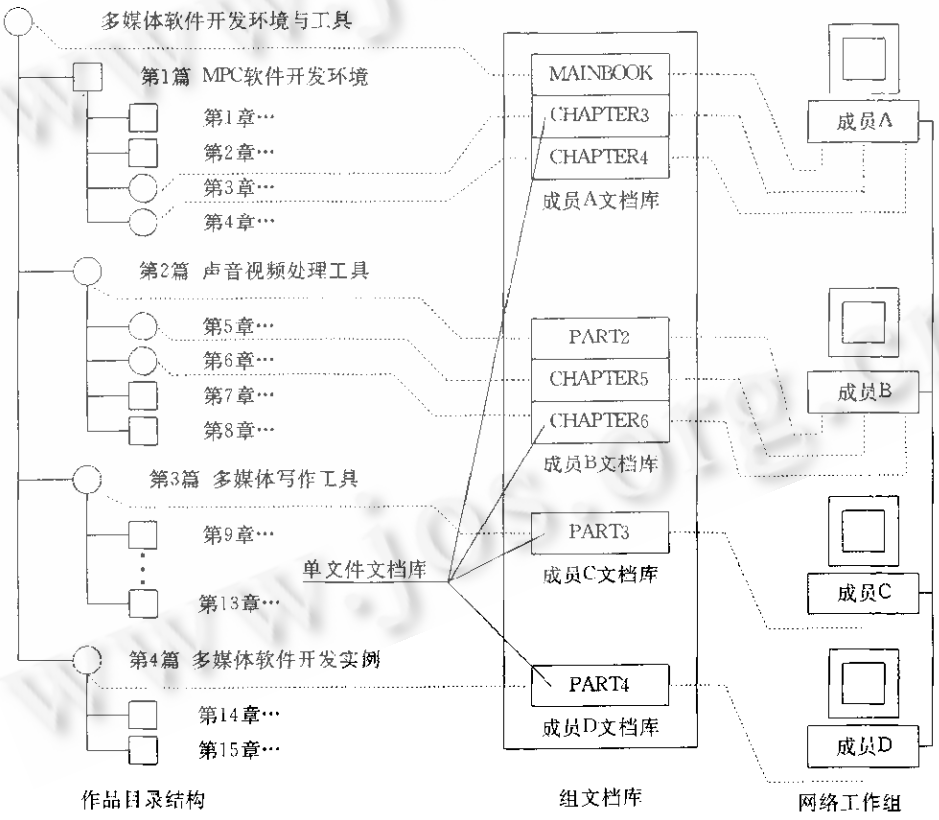

图1 分布文档库示例

(2) 文档目录树

大多数大型文档库在制作时都必须经过良好的分类组织,因而都具有清晰的目录层次结构(如图1所示的作品目录),文档的这种顺序结构在大多数情况下比交叉链关系更为重要,因为在文档制作和阅读时,人们更多地是先从文档的目录层次入手. 虽然这种目录层次

关系可以视为超链的一种特例,可以象现有的大多数系统那样,用常规的超链模型将其同交叉链等同表示和处理,但这样混合处理难以实现良好的图形浏览导航.为此,本系统将目录顺序关系独立出来,用目录树表示,为用户写作时的文档结构组织和阅读时的浏览导航提供了可视化的操作界面.

每个文档库的目录树由根元素(树根)、目录节点元素(树枝)和正文节点元素(树叶)构成.这种目录树可以表示1个大型文档库中文档的不同分类层次,或1个局部文档如1本书、1篇论文的章节目录.为了支持大型文档库的分布协作,目录树上的任一分支内容既可以集中在当前库文件中,也可以作为一个下属的分布子库存放在其它文件中,通过目录节点元素可以将分布子库逐层挂接到上层文档库文件上,用户可以通过一个完整的目录树访问整个组文档库中的任何内容,从而保证组文档库在逻辑上成为1个完整的文档库作品.

(3) 节　点

节点是超媒体系统中被分块的信息单位,而链用来将大量节点连接成网状结构的信息系统.本系统中支持2种节点:目录节点和正文节点.

目录节点用来表示目录树中的目录元素(树枝),其中包括主题信息描述——如目录标题、主题、作者等;子文档文件描述——若本目录分支内容分布在其它子文档库文件中,则本项记录子库文件路径名.正文节点是包含最终的正文内容的节点(树叶),除包含目录节点的信息外,还包含节点体索引指针.

节点体是不定长的混合型信息块,可以是书中的1节正文、论文集中的1篇论文、百科全书或字辞典中的1个条目,其划分方法和长度完全随内容而定,除含有文本外,还可以包含其它多媒体信息.图2表示了一个节点的基本描述信息和节点体结构信息.

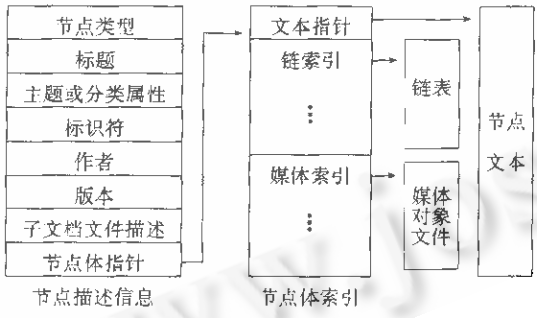

图2　节点描述和节点体

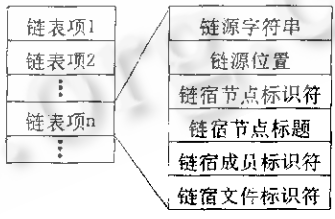

图3　链表及链表项

(4) 链和链表

本系统支持2种链:顺序链和交叉链.顺序链表示文档的基本目录顺序,而交叉链则记录节点间的交叉跳转.

为了实现从一个节点向其它节点的跳转,系统中必须记录每个正文节点中的所有交叉链,形成交叉链表.在正文中引起跳转的信息块称为"链源",所要跳转到的目标节点称为"链宿",链宿指向整个目标节点.不能指到节点内部的局部信息.超链链表记录了每个链的链源和链宿节点地址.为了支持在分布文档库中不同库文件之间的交叉跳转,还要记录链宿节点所对应的成员信息、文档文件路径信息(当链宿是在一个文档文件内部跳转时,这2项都为空).图3表示了交叉链表的基本结构.

(5) 多媒体对象

本系统支持的媒体包括中英文文本、16色和256色的图形和图象、表格、波形声音、MIDI音乐、动画、AVI视频. 除文本外,其它媒体都借助于Windows的OLE或直接通过MCI驱动链接到正文文本中.

(6) 检索索引表

从本质上说,超文本是一种适用于在信息网中沿某个主题线索浏览航行的技术,它必须有一个要素,即已有"线索"后才能快速浏览得到所需的信息. 然而对大容量的文档,刚进入时仅仅依靠浏览是难以快速找到所要的主题线索的. 因此,第3代超媒体系统必须同时提供"粗定位"和"细定位"功能,亦即开始时利用检索功能以仅有的少量线索,在全系统中快速检索到更多相关线索,大幅度地缩小信息范围,然后再由超媒体浏览功能"细定位". 本系统提供了3种检索手段:全文检索、关键词检索和主题属性检索.

CCHMDOC首先实现了一种基于大容量超文本文档的全文检索技术,其主要数据结构包括全文字表——对文本中每个出现的字符建立一个字表,顺序记录该字在全文中的所有出现地址;如图4所示,当检索一个字串如"媒体"时,以字符"媒"和"体"在相应字表中扫描,若文本中有"媒体"一词出现,则必有Pim和Pjn,使Pjn−2=Pim(2是2字间的距离),每查到一对这样的位置值,就是检索出的一次出现,扫描2字的整个字表,可检索出所有出现. 但这样查到的位置仅仅是内部物理位置,在超文本系统中还要转换成超文本文档中的逻辑位置——节点索引信息,进一步转换成标题信息通过检索界面显示给用户. 因此,全文检索部分还要包含"节点逻辑地址转换表".

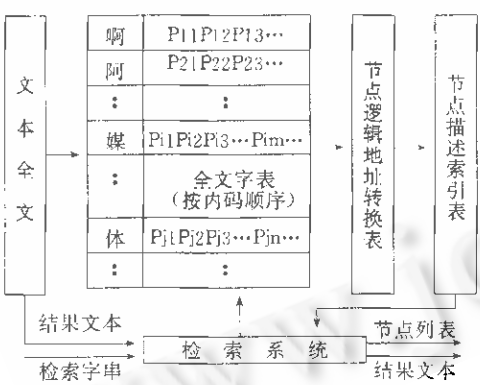

图4 全文检索数据结构示意图

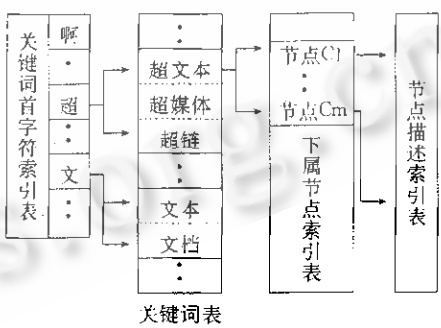

图5 关键词检索数据结构

关键词检索法的索引表由3级索引组成,如图5所示. 基本实现方法是对每个汉字字符建立一个"关键词首字符索引表",每个索引项中有一指针指向一个以该汉字开头的"关键词表",每个表项包含一个关键词和"节点索引表"指针,该指针指向所有包含该关键词的节点地址. 写作时对每个标注的关键词如"超文本",取出其首汉字"超",到首字符索引表中取出以其打头的词表,查出包含关键词"超文本"的所有节点索引,再通过节点描述索引表转换出具体的节点标题信息显示给用户.

对某些文档有时需要根据主题属性信息进行检索,如文章的标题、主题或分类、作者等. 在建立节点时,可将这些信息输入存放到节点描述表中,检索时就是根据描述表中的主题属性进行查找.

(7) 总体数据结构

图 6 给出了以上主要数据模型中的数据结构及相互间的关系.

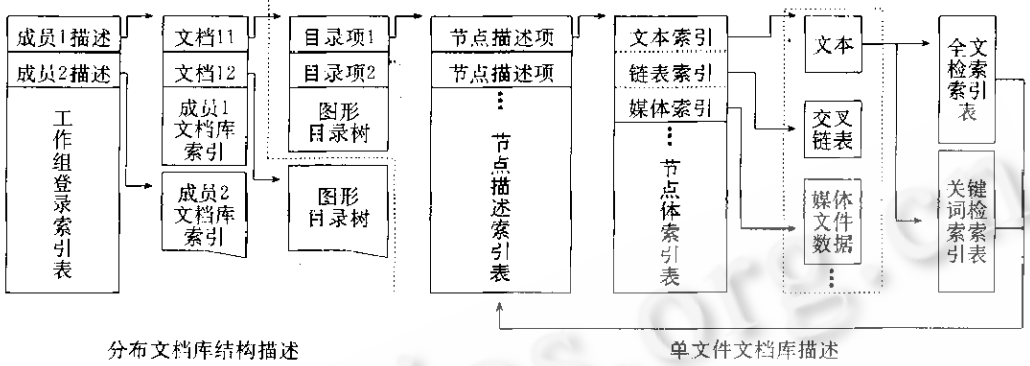

图6 系统总体数据结构

1.2 系统结构模型

超媒体系统至今仍没有一个标准的结构模型,但比较公认的是由 Compbell 和 Goodman 提出的结构模型[3],该模型中将超媒体系统划分为 3 个层次:

(1) 用户接口层:这是系统的最上层,包含了用户对系统的操作界面和工具,如节点和链的编辑、阅读展现界面和导航浏览工具等;

(2) 超媒体抽象机层:这是中间层,也是超媒体系统中的核心,包含了系统的基本数据模型,如节点和链的结构描述等;

(3) 主文件系统层:这是最底层,负责基本媒体数据文件的存取、数据共享等.

但是这只是一个基本的理论模型,本系统结合中文超媒体应用中的特有问题,建立了如图 7 所示的结构模型. 该模型仍保持了 3 个基本层次,其中用户界面层又分为写作层和阅读层,两界面下分别包含一些主要的写作和阅读应用工具. 抽象机层主要包含了前述的主要数据结构,而分布文档库文件系统层则包含了基本的文档文件存储管理、分布文档文件管理、共享并发控制、媒体服务驱动管理等,还包括分布式文档数据库.

2 系统实现

目前已实现了基本原型系统,系统是一个全集成化的交互式环境,主要分为写作层和阅读层,可以在写作层和阅读层之间相互切换. 对已制作好、准备最终发行的文档库作品,提供一个单独的阅读器. 系统总体软件结构如图 8 所示.

3 结束语

CCHMDOC 着重探讨研究了大型中文超媒体文档写作应用中的一些重要技术,尤其是中文超文本结构模型技术、中文快速检索技术、分布协作写作等技术,在进行基础研究的同时,从实际应用出发,已开发了一个有实用性的原型系统. 由于系统所涉及的技术内容很多,限于篇幅,难以将系统的主要技术内容同时容纳在本文中,因此本文主要仅讨论其中的主要数据结构模型问题,有关的主要技术内容将在后续文章中详细讨论.

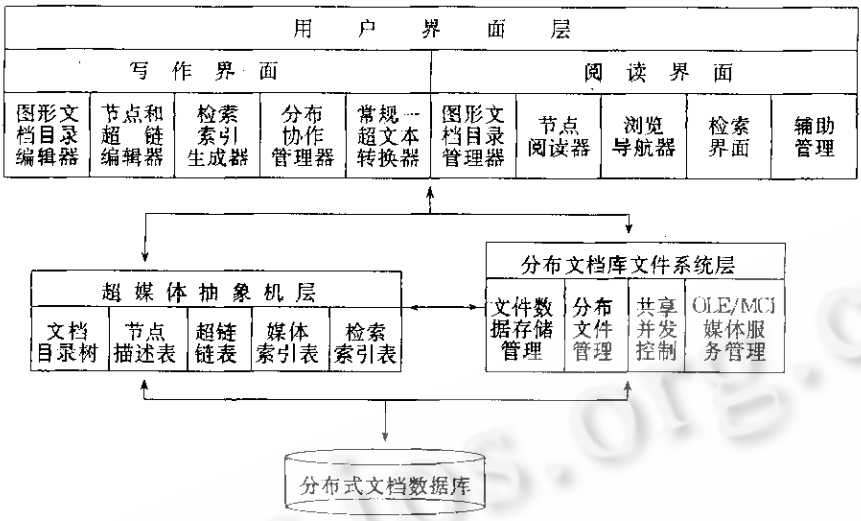

图7 系统结构模型

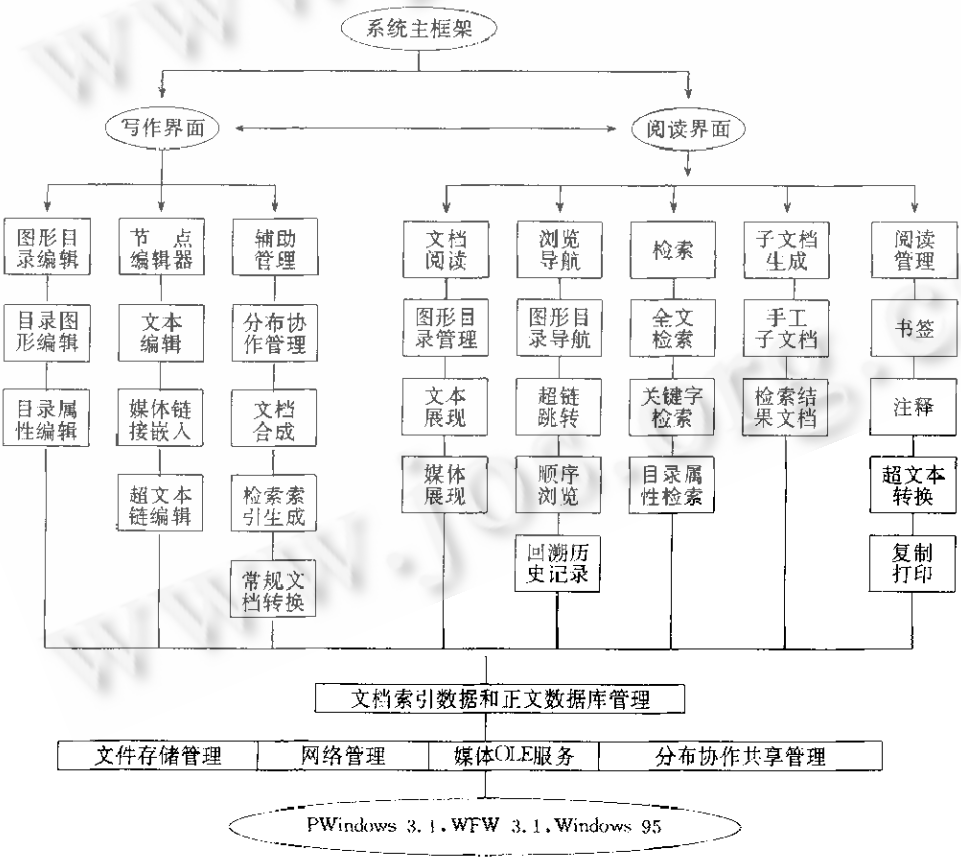

图8 系统总体软件结构

参考文献

1 老松杨等. Hypertext 系统的结构与特征. 计算机世界,1993-11-03.

2 Halasz Frank G. Reflections on notecards: seven issues for the next generation of hypermedia system. CACM, July 1988, **31**(7):836~851.

3 Brad Campbell & Joseph. HAM: a general purpose hypertext abstract machine. CACM, July 1988, **31**(7):856~861.

4 Goodman M, Akscyn Robert M, Mccracken Donald L. KMS: a distributed hypermedia system for managining knowledge in organizations. CACM, July 1988, **31**(7):820~835.

5 Emily Berk, Joseph Devlin. Hypertext/Hypermedia Handbook. USA: Intertext Publications, 1992.

DATA AND STRUCTURE DESIGN OF A COOPERATIVE AUTHORING SYSTEM FOR HYPERMEDIA DOCUMENT BASE

HUANG Yihua　YOU Xiaobai　JI Yuan　YANG Wenqing　ZHANG Fuyan

(Department of Computer Science and Technology　Nanjing University　Nanjing　210093)

Abstract　This paper discusses the design of the data and structure model for a cooperative authoring system for large-storage document based on Chinese hypermedia. It emphasizes on some key data and structure models such as distributed document base, document catalog tree, node and link table, multimedia object, document retrieval index and structure model for whole hypermedia system.

Key words　Distributed, cooperation, hypermedia, document base, data structure.