●傅 敏

谈数字图书馆信息组织的数据库技术

摘 要 数字图书馆信息组织的数据库技术,主要有面向对象数据库技术、关系数据库技术、多媒体数据库技术、非结构化数据库技术和数据仓库技术等。当前数字化信息组织技术中应注重解决的问题主要是信息组织的标准化问题、信息资源标引问题和数字资源的检索问题。参考文献7。

关键词 数字图书馆 数据库技术 信息组织 存在问题
分类号 G250.74

ABSTRACT Database technologies for the information organization of digital libraries include object-oriented database technologies, relational database technologies, multimedia database technologies, non-structured database technologies and data warehouse technologies, We should pay our attention to the standardization of information organization, the indexing of information resources and the retrieval of digital resources. 7 refs.

KEY WORDS Digital library Database technology. Information organization. Existing problem.
CLASS NUMBER G250.74

为了促进我国数字图书馆的发展,本文就数字图书馆信息组织技术及发展趋势等进行了探讨,希望能对数字图书馆信息资源的组织和建设起到积极的作用。

1 数字化信息组织的数据库技术

可以说数字图书馆就是存在于因特网上的巨型数据库。数字图书馆中信息的获取、存储、组织、检索和分析统计都离不开数据库技术。

1.1 面向对象数据库技术

面向对象方法是一种认识、描述事物的方法论,它起源于程序设计语言。它以客观世界中客观存在实体对象为基本元素,并以类和继承来表达事物间具有的共性和它们之间存在的关系,用一种与客观世界比较直接的映射方式很好地实现了抽象、封装、复杂性控制、信息隐蔽等机制。面向对象数据库是面向对象方法在数据库领域中的实现和应用,它既是一个面向对象的系统,又是一个数据库系统。

当前,面向对象数据库技术仍处于不断发展和创新的阶段,在概念、原理和实现上都还没有形成被广泛接受的共识。但对下述基本概念的支持是面向对象数据库所应该具有的:对象(Object)、类(Class)、继承(Inheritance)、封装(Encapsulation)等。目前已有一些成功的面向对象的数据库管理系统,但是其工具、环境和对面向对象方法的支持程度还有待一步完善。URION、IRIS、ONTOS、ObjectStore等是当前较有影响的OODBMS。许多主流的关系型数据库系统(如Oracle,Informix)也在新版本中加入了面向对象的特性,也就是对象关系型数据库。纯面向对象数据库管理系统自然对于面向对象方法具有较好的支持,对象关系型数据库管理系统一般也应支持SQL环境中的基本类型扩充、复杂对象处理、对象类属的继承和产生式规则的应用。

1.2 关系数据库技术

关系数据库技术是通过引入数学领域的关系模型及关系代数和关系演算,以关系概念为基础发展起来。经过几十年的应用和发展,在处理文本数据、管理事务等方面奠定了它的优势。在信息存储方面,关系数据库以二维表的方式管理数据,数据以一条条记录的方式进行管理,每一记录内部包括许多字段,字段名不可重复,对每一记录的每一字段具有唯一值,字段中不支持子字段。关系数据库是一个严格的二维表,在结构定义上有很强的限制,如对表中每个属性的长度是固定的、类型是事先定义好的。这样做是为了保证关系运算的准确性和检索的完备性,但是在另一方面又限制了数据库内容的变化。在信息(数据)检索方面,关系数据库的检索是在基于索引文件(IndexFile)基础上的SQL查询。关系数据库为每一个可检索的数据项建立一个索引文件,通过索引文件对字段进行检索。对属于不同表的数据项进行组合检索则需要对表进行链接操作。当数

据量太大时，对系统空间要求很高，且检索速度也不太理想。另外，关系数据库对检索词的索引是以整个数据项的内容为单位的，不能满足一些更深层次的索引要求，如全文检索。为了克服这一缺点，关系数据库生产商推出的新版 UniVersalDatabaseServer 或对象关系数据库中普遍提供了数据库扩充功能，使得全文检索引擎可以无缝集成到数据库中，例如 Oracle8.x 的 Cartridge 技术，Informix 的 DataBlade 技术，IBMDB2 的 Extender 等等。在多媒体信息的处理方面，关系数据库以处理文本信息见长，对于多媒体信息从一开始就没有将其纳入管理范围。随着因特网的兴起，大量多媒体信息的涌入，使得关系数据库生产商们不得不考虑对多媒体信息的处理问题，于是纷纷提供了对于一些超长文本、图像、声音等多媒体的以及面向对象的扩充，如 Informix 数据库允许用户在数据库中建立复杂的数据类型及用户自定义的数据类型，同时可对这些数据类型定义各种操作和运算以实现对象的封装。关系模型理论从提出到现在已发展了近 30 年，关系数据库技术已成为一种相当成熟的技术，特别是在结构化数据的处理方面有着极大的优势。在数字图书馆中还有着相当一部分的结构化信息，如各种统计数据、数值和事实数据库等都可以应用关系数据库技术进行管理。

1.3　多媒体数据库技术

数字化图书馆要求数据库具有管理图像、文本、声音、视频等多种媒体信息的能力，关系数据虽然可以通过引入抽象数据类型来支持对多媒体信息的处理，但这种支持仅停留在简单的输入输出上，对于其他操作和深层次检索要求必须由用户自行定义。因此，多媒体数据库概念应运而生。

目前，多媒体数据模型仍处在探索阶段。通常多媒体数据库管理系统（Multimedia Database Management System 简称 MDMS）分三个层次。用户界面层 UIL（User Interface Layer）为第一层，它完成系统和用户之间的信息交换；第二层是多媒体数据库管理层 MDBML（Multimedia Data Base Management Layer），是实现 MDMS 的核心部分，它不但管理格式化数据，而且还管理非格式化数据；第三层是多媒体数据库层 MDBL（Multimedia Data Base Layer），它由四种类型的库组成，它们是字符数值库、文本库、图像库和声音库，可以是任何现有的数据库，并不要求面向对象。用户通过 UIL 向系统提交查询命令，MDBML 一方面将 UIL 送来的数据译成操纵语言，另一方面将各 DBML 获得的数据组装成一个统一的数据对象然后再送给 UIL。

1.4　非结构化数据库技术

非结构化数据库就是字段数据及字段长度可变的数据库。非结构化数据库观点认为信息大体上可分为两类：一类信息能够用数据或统一的结构加以表示，称为结构化数据，如数字、符号；而另一类信息根本无法用数字或者统一的结构表示，例如文本、图像、声音乃至网页等，称为非结构化数据。结构化数据是非结构化数据的特例。关系型数据库就是一种结构化数据库，它很难处理网络中千变万化的非结构数据，必须采用子字段、多值字段以及变长字段的机制，允许创建许多不同类型的非结构化的或任意格式的字段，以突破关系数据库非常严格的表结构。非结构化数据库技术将非结构化和结构化数据都定义为资源，使得非结构化数据库的基本元素就是资源本身，即数据库中的资源可以同时包含结构化的和非结构化的信息，所以，非结构化数据库能够存储和管理各式各样的非结构化数据。通过这种对资源的管理方法，非结构化数据库实现了数据库系统从数据管理到内容管理的转化。非结构化数据库最大的特点在于它突破了关系数据库结构定义不易改变和数据定长的限制，支持重复字段、子字段以及变长字段并实现了对变长数据和重复字段进行处理和数据项的变长存储管理，在处理连续信息和非结构信息中有着传统关系型数据库所无法比拟的优势。在信息检索方面，关系数据库是通过建立索引而实现快速检索的，而非结构化数据库则通过倒排文档来实现记录的快速定位。灵活高效的倒排文档技术不仅能满足传统的按整字段和子字段进行逻辑组配查询的需求，而且还能进行全文任意词的单项及组配检索，检索速度快且不受文献量的影响。在多媒体信息的处理方面，非结构化数据库的记录是不定长的，可以存储各种信息，如文字、图像、视听资料等。非结构化数据库可以很轻松地处理多媒体信息。

1.5　数据仓库技术

数字图书馆不仅要提供一次信息，还必须提供经过深层次开发的二次、三次信息。数字图书馆的功能不仅应包括信息导航和信息提供还应包括信息分析和决策支持。因此，数据仓库技术也应成为数字图书馆信息组织的关键技术之一。数据仓库是集成的面向主题的数据库集合，它是用来支持决策支持功能的。其中每个数据单位都与时间有关。数据仓库中的数据应该是良好定义的、一致的和不变的。

其数据量应该足够支持数据分析、查询、报表生成和与长期积累的历史数据的对比，数据仓库技术就是一种能满足上述问题的方法，数字图书馆中存在着大量的历史数据，用数据仓库将它们组织起来，可以在更高的层次上充分利用这些数据。

2 我国数字化信息组织技术发展中的问题

2.1 信息组织的标准化与协调问题

为了便于用户获取有效信息，进行信息共享，数字图书馆在进行信息组织时就要有一个统一的标准，以便使组织起来的信息在各数据库、各网络平台之间自由流动。标准化是数字图书馆信息组织的关键。遗憾的是我们目前还没有相应的信息资源数字化制作标准。目前可行的措施是采用国际标准的数据格式，至少是开放的标准，如文本：纯文本、SGML、XML、HTML、PDF；图片：TIFF、JPEG、GIF；图像：JPEG、JPEG2000；声音：MPEG、AC3、MP3；视频：MPEG等。同时，国家有关部门组织信息产业界、图书情报界和国内软件开发商参与讨论制定电子书刊标准、各种元数据标准、多媒体信息等标准，尽快实现数字化信息资源的标准化建设，以便在统一的协议下，开展分布式海量信息资源建设与检索应用。

2.2 数字化信息资源的标引问题

大量的文献必须经过标引、分类才能更好的为用户所使用。传统的文献标引方法采用CNMARC或UNMARC标准。MARC标准是针对纸质文献而设计的，而大量数字化文献的出现，使过于精细RICH格式的MARC标准已经无法适应。对此，国际上制定并试行了针对数字信息资源新的标引方法，即DUBLIN CORE。经过一段时间的使用，DUBLIN CORE已经成为数字信息资源标引事实上的标准。采用DUBLIN CORE为数字信息资源的缺省标引方法，进一步保证了数字信息资源建设的标准化。在提供标准化标引方法的同时，信息资源数字化建设与应用软件系统也可提供用户创建自己的标引方法，即系统提供标引模板，用户可以根据自身的情况设计标引方式。除了提供DUBLIN CORE标引信息，还应针对文献的种类（书、刊、文集等）提供各具特点的目次信息，并解决期刊文献跳页的问题、文集的页号问题、按文章标引的问题。在这方面还远远不能适应信息化资源的标引需要，应当进一步开展研究。

2.3 数字化信息检索技术问题

如何对多媒体信息建立有效索引，是当前研究的热点之一。现代的检索技术已经引入了超文本和超媒体的概念，由字符匹配向概念匹配发展。当前检索技术研究的热点是如何综合利用两种或多种媒体的特征，以便使用户容易达到较高的检索效率，以及如何结合多类特征（音频、视频、文本等）抽取语义和结构，在多个层次上组织信息内容。此外，用户查询接口、多媒体内容描述标准的研究制定、高维索引技术以及对合成媒体如动画、VRML数据进行检索等，都是需要进一步研究的问题。

2.4 数字化信息资源的组织问题

在进行网络信息组织时，应遵循选择性原则、多维揭示原则、非线性组织原则、标准化原则、完备性原则等。应同时使用几种数据库技术，故必须解决好异种数据库之间的接口问题，要使对象服务器能够无缝完成查询指令的接收和对数据库的访问与检索操作。DUBLIN CORE只是一种标引方法，如处理不好标引信息的保存，DUBLIN CORE的作用也会被削弱。目前国际上广泛采用XML对数字化信息资源进行组织，同时使用XML作为DUBLIN CORE信息载体。XML被广泛使用在数据交换领域，有优异的跨平台、跨语言等特性，DUBLIN CORE也推荐XML作为首选信息载体。通过XML可将数字文献、目次信息及其标引信息有机地组织在一起，以便于读者使用及系统之间交换数据。所以全面支持XML是信息资源数字化建设与应用软件系统必须考虑的问题。

参考文献

1 玉英，索传军.网络环境中信息检索理论与实践的发展.图书情报知识，2001(1)
2 王群，敬卿.基于数字图书馆的信息组织研究.高校图书馆工作，2003(5)
3 张敏君.数字图书馆的特征及信息资源的组织.图书馆学研究，2002(7)
4 吴叶葵.数字图书馆信息资源整体化组织的实现.情报杂志，2003(12)
5 赵一丹.论数字图书馆基于内容的多媒体数据查询和检索技术.中国图书馆学报，2001(3)
6 孙一钢.数字图书馆的技术体系结构.现代图书情报技术，2001(5)
7 陈文翠.数字图书馆建设中面临的技术挑战及解决方案.情报科学，2004(1)

傅 敏 深圳大学图书馆馆员。通信地址：广东深圳。邮编518060。
（来稿时间：2004-06-20）