

特性化数字图书馆技术研究

Study of the Technology of Characteristic Digital Library

叶小榕/YE Xiao-rong

北京交通大学电子信息工程学院, 北京 100086

School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100086, China

[摘要] 针对国内数字图书馆的不足, 概述了国际数字图书馆技术人性化发展趋势, 从中得到可借鉴的经验。

[关键词] 数字图书馆; 视频图书馆; 语音图书馆; 移动代理图书馆; 网格图书馆

[中图分类号] G350

[文献标识码] A

[文章编号] 1000-7857(2006)10-0068-04

Abstract: This paper firstly presents an overview of the characteristics of the digital library. And then the current status of Chinese digital libraries is reviewed briefly and their deficiencies are pointed out. Finally, some innovative technologies of digital libraries around the world are introduced and the trend of foreign digital libraries is analyzed, which can be used for reference for the domestic ones.

Key Words: digital library; video digital library; spoken query digital library; mobile agent digital library; grid digital library

CLC Number: G350

Document Code: A

Article ID: 1000-7857(2006)10-0068-04

1 引言

数字图书馆具有以下几个特点。信息资源的数字化是建立数字图书馆的基础。数字化的信息资源具有保真度高、易于共享、便于检索的特性, 十分利于保存文献资料, 并且图像、视频、软件等资源只能以数字化的方式保存。信息资源网络化。网络化使信息资源的获取不受时空限制, 用户只需通过网络, 就可以方便快捷地随时获取信息资源。信息资源共享化。数字图书馆实现了真正意义上的资源共享, 可以跨行业、跨地域, 从全球范围内获取资源, 其资源之广泛、信息之丰富, 是传统图书馆远远不及的。数字图书馆是以促进知识共享、信息交流为目的, 利用数字技术、计算机技术和网络技术形成的无时空限制的信息系统, 使用户可以通过网络, 不受时间、空间限制地获取这些共享信息资源。

2 一般数字图书馆技术

自从 20 世纪 90 年代提出数字图书馆概念以来, 世界各国都投入了大量的人力、物力, 积极开展数字图书馆的研究工作; 国内也开展了数字图书馆的研究工作, 逐步建成了一些资源丰富、技术先进的数字图书馆。国内比较大的数字图书馆包括: 以中国国家图书馆为主建立的中国国家数字图书馆, 中国科学院建立的国家科学数字图书馆, 清华同方公司建立的 CNKI 知网数字图书馆等。再加上各个高校、科研院所自己

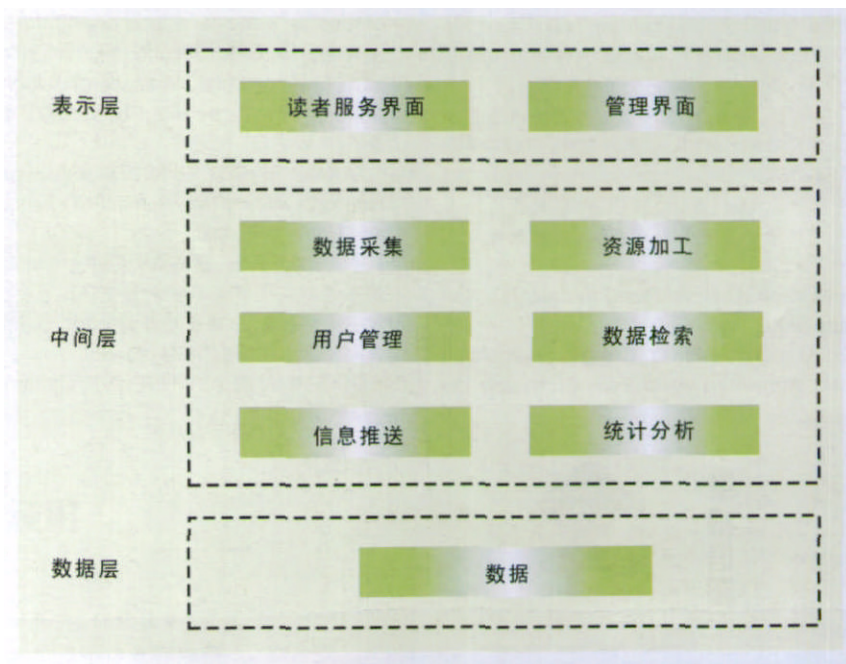


图 1 数字图书馆结构

Fig. 1 Hierarchical structure of the digital library

建立的数字图书馆, 国内的数字图书馆总数已近百家。这些数字图书馆极大地促进了国内数字图书馆的发展。

这些数字图书馆功能上各具特色, 在技术上有如下共有的特点: 基本都采用客户端/服务器模式, 可以大致划分为表示层、

中间层和数据层 3 层架构体系, 如图 1。

2.1 表示层

表示层为客户提供应用服务的图形界面, 负责向中间层发出服务请求, 并把接受传来的运行结果显示出来。表示层主要分为读者服务界面和管理界面。读者服务界

收稿日期: 2006-05-12

作者简介: 叶小榕, 女, 北京交通大学电子信息工程学院, 硕士生, 研究方向为计算机软件、数字图书馆; E-mail: yeelfine@sina.com

面为门户网站,为用户提供便利的、个性化的、安全可靠的服务,提供信息发布、资源检索、个性化定制等功能,如根据用户身份提供浏览、检索、下载,以及提供书刊简介、书摘预览、新书推荐等。管理界面用于管理员维护系统。

2.2 中间层

中间层位于表示层和数据层之间,提供客户应用程序和数据服务间的联系,所有的应用系统、应用逻辑、控制、业务处理都集中在这一层。中间层主要负责接收表示层的数据请求,对用户身份和数据库存取权限进行验证,运用服务器脚本,借助于中间件把请求发送到数据层,由数据层进行处理,然后将数据层返回的数据经过逻辑处理并转换成 HTML 及各种脚本返回给表示层。其主要功能如下。

2.2.1 资源采集

资源采集负责将各种资源信息进行数字处理,以满足数字图书馆的格式要求。比如对各种纸质资料、音像资料,要对其进行数字加工,转为数字格式的资源;对网络资源,要进行有目的的采集整理,进行有序化处理;对已有的各种数据库、电子书、电子期刊资源进行必要的格式转换、元数据标引等处理。总之,需要针对不同来源的信息,使用不同的技术手段,将它们数字化、有序化、规范化。

2.2.2 资源加工

对数字资源进行加工和著录,包括标引、分类编目、元数据管理、数字版权管理等。

2.2.3 用户管理

负责记录用户信息,并提供个性化管理,比如检索历史,还借书查询、预约,通过短信、电子邮件等方式通知用户到期书目等。

2.2.4 数据检索引擎

与传统图书馆相比,数字图书馆资源更加丰富,更加强调信息检索功能,包括目录检索、全文检索、组合检索、布尔逻辑检索等多种检索方式;支持基于后控词表的同义词、下位词扩展检索和智能全文检索,充分满足各种层次用户的检索需要。

2.2.5 信息推送管理

国内一些数字图书馆系统已经开始提供 RSS 聚合新闻服务,使系统可以主动将用户定制的信息内容,及时地送到用户端,用户可以通过本地的 RSS 阅读器在第一时间内接收自己感兴趣的信息,省去了用户登陆图书馆、查找信息的时间。

2.2.6 统计分析

提供数据挖掘功能,通过对各种记录数据,如点击浏览情况、个人定制情况的分析,得到相关的统计数据,为进一步改善系统、改进服务、优化资源提供决策支持。

2.3 数据层

数据层是 3 层模式中的最底层,它用来定义、维护、访问和更新数据,并管理和满足应用服务对数据的请求。

数据层存储大量的数据信息和数据逻辑,如加工标引后的正式数据、主题词表、用户数据、日志信息;还存储采集的原始数据,如 TXT、HTML 等文本文件,MP3、AVI、RM、WMV、MPG 等音视频文件,DOC、PPT、XLS 等 Office 系列文件,JPG、GIF、BMP、PNG、PSD 等图形图像文件,ZIP、PDF、SWF 等其他格式文件等。

3 特性化数字图书馆技术研究

特性化数字图书馆技术研究一方面继续进行信息资源的搜集整理入库工作,在数字图书馆的基础上,使信息服务更个性化;另一方面,结合计算机技术、网络技术、人工智能技术,使信息资源共享全球化。

3.1 视频数字图书馆技术

数字图书馆在文字资料方面和网络技术结合得比较紧密,能够提供文字资料的查询、下载等功能。但在视频资料方面,技术却相对落后,网页上仅仅是文字介绍,无法通过互联网浏览视频,并且如果想获取视频资料,用户仍然不得不到图书馆去借阅,在视频方面没能发挥出数字图书馆资源共享的优势和特点。

视频数字图书馆结合了先进的网络技术,建立了一个视频资源共享信息库,使读者可以通过网络查阅、预览和下载视频资料。

视频数字图书馆技术具有以下特点。

3.1.1 采用了第二代互联网分布式存储技术

分布式存储技术使用磁盘阵列在多个服务器间分布的方式,来存储多媒体数据,服务器间具有很高的带宽。系统通过管理层实现单一存储映像(Single Storage Image)技术,使所有服务器相当于一个巨大的存储池。分布式存储技术解决了海量数据快速读取的问题,是视频图书馆有利的技术保障。

3.1.2 数据库采用了 MySQL 数据库

MySQL 是快速、多线程、多用户的 SQL 数据库服务器。并且 MySQL 作为开源数据库,有很低廉的商用使用价格,对国内很多资金并不充裕的图书馆而言,能够节省大量的建设费用。

3.1.3 支持高级检索功能

除了常见的关键字组合检索和分类检索外,还针对视频资料的特点,提供视频特征检索。用户可以根据视频文件的类型、时间长短、声音和提供的组织来检索视频文件。

3.1.4 提供视屏预览

这是视频图书馆特有的功能,这样用

户可以先了解视频的大致内容,再决定是否下载,极大地方便了用户。提供的预览方式有快速浏览、内容摘要图片和前 7 s 浏览等功能。这些功能完全是系统自动生成的,无需人工干预。而且视频图书馆提供了不同清晰度的视频下载,用以满足不同用户的需求。

3.1.5 支持用户提供数字视屏资料

传统的图书馆主要由图书馆自己收集或者购买资料,而视频图书馆具有更加先进的理念,其资料的搜集整理完全面向大众开放,任何个人或者组织均可以直接通过网络提交视屏资料,经过审核后就会公开发布。这种从用户来又服务于用户的思想,最大限度地扩大了资料的来源途径,体现了资源共享的精神。

视频数字图书馆可以说是“看得见的图书馆”,使图书馆不再仅仅局限在文字资料方面,而扩展到视频领域,极大地满足了读者对获取视频信息的需求。

3.2 移动代理数字图书馆技术(PDLS)

为能够充分利用网络上大量的信息资源,移动代理图书馆技术应运而生,即将人工智能领域的代理技术,以及网络搜索技术应用于图书馆技术中,在功能上使图书馆成为具有从网络上自动搜集信息、自我整理信息的智能数字图书馆,弥补了人工处理时效慢、搜集信息范围窄的缺陷。

移动代理图书馆技术具有以下特点。

3.2.1 代理技术

代理(Agent)是指穿梭在网络中,执行各种特定任务、搜寻远端工作站上的信息资料,或者与其他 Agents 交互的系统平台。它具有智能搜索、信息过滤、信息保存、资源共享等功能。而且其更加先进的移动代理技术(Mobile Agent),可以主动对网络环境变化做出反应,能够通过网络完成自身的定位、检索、统计、更新等功能,特别适合大规模复杂多变的网络环境。

移动代理的开发平台是在传统的分布式代理开发工具 DECAF (Distributed Environment Centered Agent Framework)基础上,增加了 Voyager(一种分布式移动代理)和多代理协商算法而形成的。它结合了 DECAF 的分布式特点,同时支持移动代理。其功能包括通信代理、计划安排、监控、协调、诊断、学习等,支持代理的自动发现和互相协商机制。最大的特点就是能主动激活远端设备来建立连接,并且可以充分地利用带宽资源,十分适合数字图书馆大信息量交换的需要。

3.2.2 系统结构

整个系统由 2 部分组成:客户集和服务器集。

其中客户集分为 3 个模块:用户接口模块,给用户显示查询、控制等界面;用户描述控制模块,负责保存用户资料 and

以前的查询信息,并且能够根据用户查询单词出现的频率,不断地修正用户的描述信息,这些描述信息会起到辅助查询和信息过滤功能;PLA(Personal Library Agent)模块,负责多代理任务,并且根据用户的请求查询、搜集信息,并保存返回给用户结果。PDLs查询速度快、命中率高,并且随着代理搜索到的服务器的增加,PDLs的多代理优势就越发明显。

服务器集是数字图书馆系统利用移动代理的支持,自动搜索网络信息,自动整理过滤,添加到图书馆的数据库中。这方面类似“google”、“百度”等搜索引擎的功能,将搜索到的信息保存到本地系统中,可以自动扩充图书馆的信息量。

移动代理数字图书馆显示了“智能化图书馆”的发展趋势,充分利用了人工智能技术,目的在于实现图书馆自动搜索信息、整理信息的功能,这样图书馆就不仅仅是信息的被动接受者,而更是信息的主动探索者。

3.3 语音查询数字图书馆技术

随着数字图书馆的建设,已经有越来越多的人可以通过互联网访问数字图书馆,但是对于特殊人群,比如盲人,以及一些使用电脑有困难的老人和儿童,其需求却被无意中忽视了,他们在使用数字图书馆中有明显的不便,从图书馆获取信息很困难。语音查询数字图书馆技术弥补了这一不足。语音查询数字图书馆使数字图书馆能够支持语音查询功能,它就像是给数字图书馆戴上了一只耳朵,在不影响现有功能的情况下,使图书馆能够“听”懂用户的需求,做出正确的处理,然后将结果通过语音或者其他用户指定的方式返回给用户。当然,由于条件的限制,不可能通过语音查询方式获取详细的信息资料,但是可以进行简单的查询,满足用户最基本获取信息的需要。

语音查询数字图书馆主要采用了语音查询过程(SQP: Spoken Query Processing)系统作为技术保障。SQP技术提供了一个强大的语音查询平台,能够将用户的语音查询转换为数字图书馆系统可以识别的标准查询语言,从而可以直接使用现有系统的查询接口。SQP是从语音识别技术上发展起来的,但是SQP有更高的要求:要求查询过程是实时的查询,能迅速返回结果;查询的内容比语音文件更短,但是要求更加精确;用户声音差异很大,语音的识别技术难度更大。

提高SQP技术可以大大提高语音查询的准确性,其技术关键点有如下几个方面。

相关性反馈(RF: Relevance Feedback)技术。利用此反馈技术来提高系统的语音查询效率,特别是短词汇的查询准确性。关键词重读识别技术。如果某个单词被重读,

则可以认为此单词为用户认为的关键词汇,在查询的算法中,相应增加此单词的权重。综合语义学。系统在收到用户提交的查询词汇后,不仅仅查询匹配的词汇,还查询语义相关的词汇,在一定范围进行查询,提高查询的命中率。

语音查询功能现在虽然还相对简单,无法做到与人的自然沟通,但方便的语音查询功能将是现代化数字图书馆必不可少的功能,显示了对特殊人群的关注,使他们能方便快捷地查询所需要的信息。并且此功能已有一些试验系统投入使用,用户通过拨入指定的电话号码,就可以来查询报刊新闻标题,并通过电话返回简单的查询结果,十分地简单快捷。

3.4 网格数字图书馆技术

数字图书馆的不断涌现,一方面给用户提供了不同信息来源扩大的信息获取量,但另一方面却使用户在获取信息时,不得不登陆不同的数字图书馆分别查询,使获取信息的效率十分低下。同时,很多数字图书馆在建立数据库时,由于各个数字图书馆系统互不兼容,无法有效利用已有数字图书馆的信息资源,不得不进行重复建设、重复投资。这些对于数字图书馆的建设都是很不利。

为从根本上解决此问题,国外提出了网格数字图书馆,使各个不同的数字图书馆组成一个巨大的全球网格数字图书馆GDLG(Global Digital Library Grid),彻底解决不同的数字图书馆系统互相兼容从而彼此之间最大限度地共享资源、共享检索处理能力、共享存储资源等问题,使读者只需登陆任意一家数字图书馆,就可以访问网格系统中的所有图书馆,从全球范围内获取丰富的信息资源。并且在建设新的数字图书馆时,无需进行重复的数据资源采集、录入工作,从网格中提取需要的信息资源进行再加工,即可实现数据库的建设。

网格数字图书馆技术难度比较高,它包括以下几个方面。

3.4.1 网络技术

网格技术出现于20世纪90年代,是利用互联网,把网络中的电脑组织成一台“虚拟的超级计算机”,实现计算资源、存储资源、数据资源、信息资源等的全面共享。目的是利用互联网,把分布于不同地理位置的计算机、数据库、存储器、软件等资源连成整体,为用户提供一体化信息服务。其核心思想是:“整个互联网就是一台计算机”。现在的网格技术基本都采用由全球网格论坛(GGF)开发的Globus工具包作为标准开发平台。Globus提供了安全管理、资源管理、数据管理、信息服务等网格技术。

3.4.2 GDLG技术

GDLG的使用方式是将基于Web浏览器的界面,分为应用入口和用户入口。前者

允许对网格系统进行特殊操作,主要用在专业性很强的应用方面,如天体物理模拟实验等等;后者提供对大众的服务,应当尽量简化,使其与普通数字图书馆的查询方式没有什么差别。

现在已经有了多个GDLG的技术解决方案:使用GridSphere技术,采用Tomcat Application Server作为开发平台,采用Javabeans, Servlets和JSP技术来实现;GridPort技术,采用Perl为开发工具,主要用于Unix系统;采用NMI OGCE(Open Grid Computing Environments Collaboratory)技术,通过中间件提供接入服务。

3.4.3 系统设计

网格数字图书馆系统最复杂的工作,就是要革新现有的图书馆系统,使网格技术深入到图书馆系统内部,使网格技术和图书馆的数据库系统、检索系统、管理系统能够协调工作,达到网络共享的目的。

GDLG的建设步骤如下:各个数字图书馆建立网格环境和安装Globus工具包;建立共享的数据元(Metadata)仓库,分析已有的数据资料;建立GridFTP,这是Globus工具传输文件的标准协议;各个网格成员间联通,实现网格技术下的资源共享。

网格技术将能彻底解决图书馆互联互通问题,使全球的信息资源成为一个整体,虽然此技术还没有完全成熟,但其影响将是十分深远的。

4 结论

特性化数字图书馆技术研究是围绕享受信息全球化、提供信息服务人性化而展开的,因此它们的个性化主要表现在以下几方面。

1) 突出创新。这些数字图书馆紧密结合科学技术的发展,计算机、网络、人工智能、语音识别、视频处理都有所涉及,力图满足不同用户的多种信息获取需求,其定位明确、特色突出,一定程度上代表了数字图书馆技术的多种发展方向。

2) 网络互联信息共享。数字图书馆的最大特点就是资源的共享,因此国外数字图书馆的建设很注重互联互通问题,一般是先进行网络的建设,保证资源能够共享,然后再进行数字图书馆本身系统的建设,从而避免出现信息孤岛。

3) 以用户为中心。比如语音查询等一些技术,虽然功能简单,但真正从用户角度着想,解决了易被忽视人群(如盲人)的使用困难,使数字图书馆能够为所有用户提供人性化的服务。

注重技术创新、强调信息共享、关注特殊人群,应当是今后数字图书馆发展的方向。

参考文献(References)

- [1] HIROSHI M. Next generation digital library architecture and implementation [C]. Kyoto: Kyoto international conference on digital libraries. 2000.
- [2] CHEN H. Digital library development in the Asia pacific [C]. Bangkok: Lecture notes in computer science, 8th international conference on Asian digital libraries. 2005.
- [3] PAEPCKE A, BALDONADO M, CHANG

- C- C. K, et al. Using distributed objects to build the Stanford digital library Infobus[J]. Computer, 1999, 32(2):80- 87.
- [4] IOANNIDIS Y, MAIER D, ABITEBOUL S. Digital library information - technology infrastructures [J]. International Journal on Digital Libraries, 2005(5) : 266- 274.
- [5] 黄锦. 基于信息 Agent 技术的个性化学术资源信息系统 [EB/OL]. (2004- 12- 28)[2006- 03- 01]. <http://www.lib.sjtu.edu.cn/chinese/teaching&research/501.doc>.

- [6] 张怀寿, 索传军, 代根兴. 网络环境与图书馆信息资源[M]. 郑州大学出版社, 2002.
- [7] 往建文. 数字化图书与数字图书馆应用研究 [M]. 北京工业大学出版社, 2005.
- [8] 刘海英, 冯文秀, 叶辉. 基于 J2EE 的三层 B/S 信息系统研究 [EB/OL]. (2004- 06- 16)[2006- 04- 20]. http://www.yesky.com/SoftChannel/72342371961929728/20040616/1815917_2.shtml.

(责任编辑 王 芷)

·图片科技新闻·



年会开幕式



丁肇中教授答记者问



国家副主席曾庆红参加“全国科普日”活动



“科技出版与科技创新”分会场

科技盛会 群英荟萃

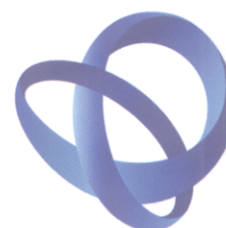
——2006 中国科协年会在北京隆重召开

金秋时节, 主题为“提高全民科学素质, 建设创新型国家”的 2006 中国科协年会在北京隆重召开。大会开幕式于 9 月 16 日在人民大会堂举行, 党和国家领导人王兆国、刘琪、路甬祥、韩启德、徐匡迪, 中国科协名誉主席周光召, 以及来自全国各地科技战线的 6 000 余名代表出席了开幕式。开幕式由中国科协副主席、书记处第一书记邓楠主持。

王兆国副委员长在开幕式上致词, 他代表党中央向辛勤工作在科技战线的广大科技工作者致以亲切的问候和崇高的敬意, 他号召各级科协组织和广大科技工作者: 充分发挥学术交流对自主创新的促进作用, 把开展学术交流与进行科学前沿的探索活动结合起来, 与解决我国发展中的重大问题结合起来, 与发现培养创新人才结合起来; 充分发挥科普主力军的作用, 大力提高公民的科学素质和创新意识; 充分发挥科技社团的桥梁纽带作用, 努力构建面向科技工作者的服务体系, 积极开展建言献策、专家论证、决策咨询等活动, 依法参与科技工作和社会事务管理, 促进我国经济社会协调发展。

开幕式上还颁发了中国科协“求是杰出青年奖”和香港求是科技基金会“求是杰出科学家奖”。山东大学特聘教授王小云因成功破解了一直在国际上广泛应用的两大密码算法 MD5 和 SHA-1 而获“求是杰出科学家奖”。开幕式结束后, 徐匡迪院士、王岐山市长、张启发院士以及丁肇中博士分别作了精彩的大会特邀报告。

9 月 16 日下午至 9 月 17 日, 年会各分会场和专题论坛分别进行了学术交流。本届年会共设“管理创新与创新性国家建设”、“农业现代化与新农村建设”、“疾病防治与人类健康”等专题分会场 15 个, 围绕



中国科协年会
China Association for Science and Technology Annual Meeting

年会永久性标识

“新北京·新奥运”、“北京产业发展”、“首都可持续发展”、“创新型北京”等专题设立了 13 个专家论坛, 共收录论文 2 081 篇。9 月 16 日, 全国科普日活动也同时在全国 31 个省(直辖市、自治区)举办, 并组织安排了“健康生活进社区活动”、“趣味科普活动进校区”、“大手拉小手——科技专家进百校”等 8 个系列、170 多项科普活动。

中国科协学术年会制度始于 1999 年, 主要目的是通过组织综合性、跨学科、跨地域、系列性的学术交流活动, 促进学科交叉融合, 促进科技与经济的结合, 促进经济与社会协调发展, 推动形成尊重知识、尊重劳动、尊重人才、尊重创造的社会风尚。从 2006 年起, 中国科协学术年会转型为中国科协年会, 并具有以下几个特点: 启用了年会永久性标志——数学中的双面拓扑图形; 年会的科普内容更加突出, 服务对象除面向科技工作者外, 还面向社会公众、政府部门和企业; 会务的组织工作由职能部门进一步向社会化的开放型转变。

(本刊记者 李慧政)