

国家工程技术数字图书馆技术体系

□ 王星 / 中国科学技术信息研究所 北京 100038

摘要: 文章从资源遴选、资源采集和管理、资源数字化加工、数字资源存储与整合、资源揭示监控和数字信息资源服务等方面详细介绍了国家工程技术数字图书馆的技术体系,并对其中使用的关键技术进行了简要说明。

关键词: 数字图书馆, 技术体系, 关键技术

DOI: 10.3772/j.issn.1673-2286.2013.10.004

1 引言

数字图书馆(Digital Library)是用数字技术处理和存储各种文献的图书馆,它把各种不同载体、不同地理位置的信息资源用数字技术存贮,以便于跨越区域、面向对象的网络查询和传播。它涉及信息资源加工、存储、检索、传输和利用的全过程^[1]。随着计算机技术的发展,数字图书馆的研究和建设已经进行很长时间,从不同领域和方面促进了数字图书馆的理论和技术的发展,但是国际上至今并没有一个可以被广泛接受的准确的关于数字图书馆的概念描述。通俗地说,数字图书馆就是虚拟的、没有围墙的图书馆,是基于网络环境下共建共享的可扩展的知识网络系统,是超大规模的、分布式的、便于使用的、没有时空限制的、可以实现跨库无缝链接与智能检索的知识中心^[2]。数字图书馆的发展按照国内外的研究情况,可以概括为三个阶段:数字化图书馆、数字图书馆服务系统和知识网络为核心的数字图书馆^[3],这也表明需要涉及三个工作内容:一是将纸质资源转化为电子版的数字资源;二是电子版资源的存

储、交换、流通和服务;三是在信息资源建设的基础上强调知识发现与交流,构建以数字图书馆为核心的知识网络。

2 国家工程技术数字图书馆的技术框架

数字图书馆的技术体系结构是建设数字图书馆系统的基础,是数字图书馆在网络和计算机技术上的

具体实现,国家工程技术数字图书馆的技术体系结构参照国际主流的数字图书馆进行设计,分为数字业务管理和数字信息资源服务两大部分,如图1所示。

3 数字业务管理

3.1 资源遴选

出于合理利用资源建设经费

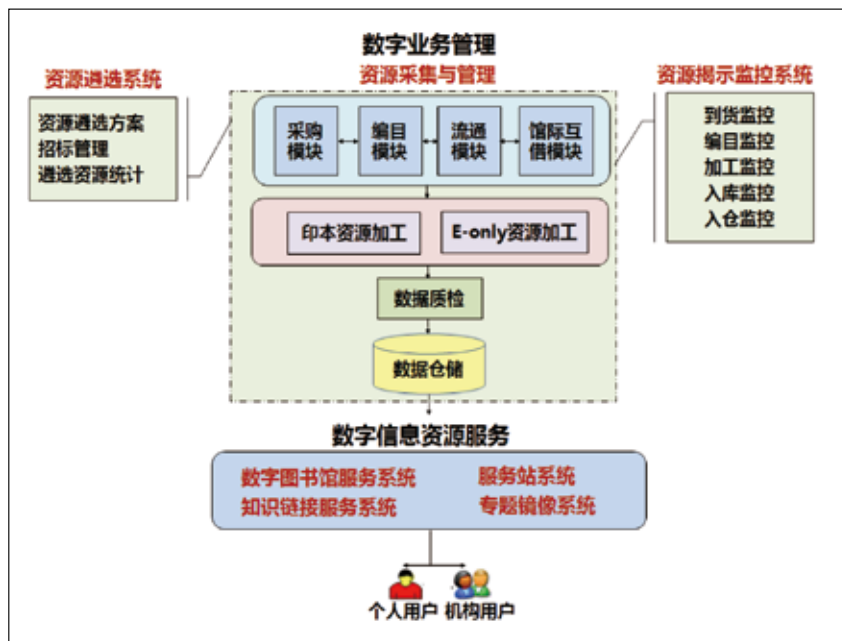


图1 国家工程技术数字图书馆的技术框架

的考虑,需要对资源进行遴选、评价,以便从中选出最有价值的资源,资源遴选系统在书商提供的清单、收录情况、资源价格等数据的基础上,提供自动规范、查重、比对的功能,形成资源遴选的基础数据,然后采用多种资源评价模型,自动计算资源的参考分值,使采购人员能够对资源进行科学的评价,并根据评价决定采购的品种,既降低了大量人工操作,又做到了采购品种选择的有理有据。

3.2 资源采集和管理

资源采集和管理依靠图书馆自动化系统来实现,它主要解决印本、光盘、微缩胶片等资源的采购、编目、典藏、流通和馆际互借,分为文献采购模块、文献编目模块、文献流通模块、馆际互借模块、连续出版物管理模块、书目查询模块和参考咨询模块等,其中编目模块同时支持CNMARC和USMARC两种格式,通过Z39.50协议,图书馆自动化系统直接为数字图书馆服务系统提供元数据,与资源加工系统进行无缝对接,减少了传统图书馆工作中繁杂而又重复性强的手工劳动,提高了各种文献的加工处理速度,使图书馆提供的服务更加高效、方便和灵活多样。

3.3 资源数字化加工

资源数字化加工是在统一的元数据标准基础上对印本、光盘、微缩胶片和E-only网络版资源进行加工处理,形成相应的数字资源,使上述资源能够进行一次加工,多次使用,不同介质的资源使用不同的资源加工系统进行加工,每个资源

加工系统均有不同的加工流程。

印本资源加工系统解决印本资源数字化的问题,印本文献经过扫描、文字识别(OCR)、解析入库、全文合成等主要工序形成数字资源,印本资源以物理本为单位进行加工,印本资源加工系统采用流水线方式,不同工序为单独的客户端模块,并由专门人员负责,后一工序对前一工序进行监控,保证加工数据的质量和完整性,加工流程见图2。

网络版资源加工系统是将非结构化的文献信息从不同网站的网页中抽取出来保存到结构化的数据库中,与一般的信息采集系统不同,网络版资源加工系统采集的对象是文献,采集的任务是按照图书馆自动化系统中的登到信息进行定向采集的,采集的网页经过解析器解析成为元数据,再根据元数据的标准进行统一规范和整理才能形成文摘库。

数据在进入数字资源仓储系

统前需通过数据质检系统的检验,数据质检系统采用联机质检的方式,分为机器自动质检规范和人工质检规范两个步骤,机器自动质检规范针对全部数据使用多种质检规范的算法自动对数据进行初步质检和规范,人工质检规范按照计数抽样检验国家标准(GB/T 2828-2003)进行抽样,采用C/S模式以文摘与全文对照的方式进行一次及二次的质检工作,由服务器自动汇总检验结果并产生质检报告,保证数据质量符合元数据标准的要求。

3.4 数字资源存储与整合

数字资源仓储系统是数字资源的存储和整合的具体实现,它存储元数据和各种多媒体资源,数据的存储与管理是数字资源仓储系统的核心,同时具有异构资源的整合、数据查询统计与分析、数据发布等功能,数字资源仓储系统的数据结构、转换映射、仓储结构和控制信

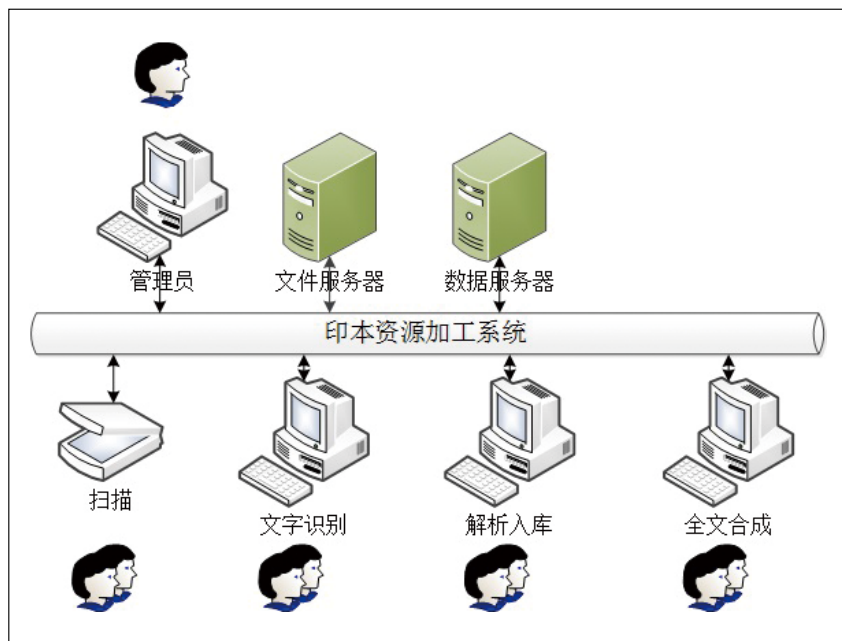


图2 印本资源加工系统工作流程

息基于各种文献类型的元数据标准进行制定,图3为数字资源仓储系统的体系结构。

经过数字化加工的各种文献资源的元数据和多媒体资源在经过质检、规范和抽取后装载到数据仓储中,通过联机分析处理(OLAP)服务器进行管理,从数据仓储中也可析出面向特定主题的数据集市,供特定的用户群体使用,而在数字资源仓储系统的前台可实现例如查询、分析、报表、发布和更为复杂的数据挖掘功能。

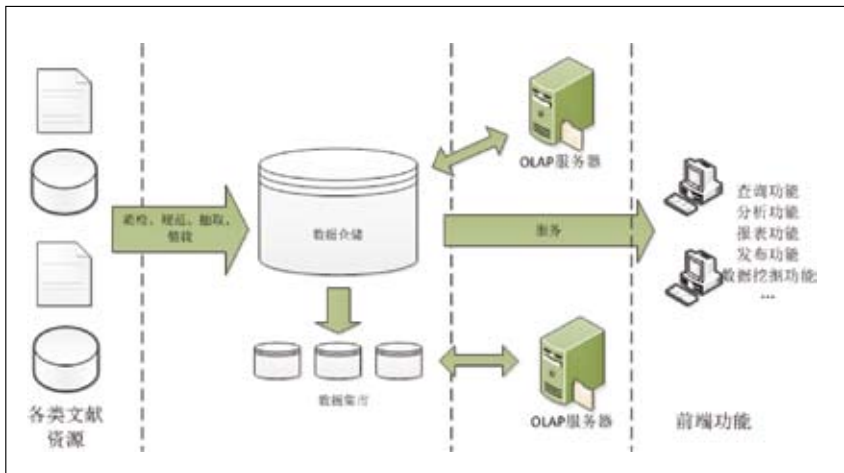


图3 数字资源仓储系统体系结构

3.5 资源揭示监控

为了保证数据在各个环节的完整性、及时性和一致性,需要在每个环节增加监控,资源揭示监控系统通过相应接口链接了图书馆自动化系统、印本资源加工系统、网络版资源加工系统、数据质检系统和数字资源仓储系统,将数据的统计信息统一汇总到一个系统中,通过资源揭示监控系统可以了解整体数据的情况,直观地看到数据流向,根据数据的情况及时调整相应的业务工作。

4 数字信息资源服务

4.1 数字图书馆服务系统

国家工程技术数字图书馆服务系统于2007年开始建设,2008年5月正式建成并对外进行服务,数字图书馆服务系统集合了国家工程技术图书馆所有本地馆藏资源、自建特色资源、采购的网络数据库的资源,为用户提供数字资源的全方位资源导航、馆藏检索、统一网络资源检索等各项文献资源服务,并

为用户提供文献原文传递、代查代借、科技查新、收录引证、参考咨询等增值信息服务。

在综合考虑易用性、开放性、安全性、可扩展性、可维护性等因素的基础上,数字图书馆服务系统使用开源的DNN(DotNetNuke)通用门户系统为基础^[4],构建了一个由图4所示的技术框架。该架构具有多层体系结构特征,采用分布式部署,WEB服务、数据库服务、全文检索服务及应用服务使用了服务器集群方式,可以满足更多复杂状况的需求。在统一检索中采用基于HTTP、OpenURL、SRW、RSS等协议的iSFX虚拟资源整合工具,实现了信息资源的整合集成和实时代理检索。在业务流程管理中使用了微软的WF(Windows Workflow Foundation)工具,使服务系统能够按照业务流程的改变而快速调整,避免业务流程变更带来的繁重的系统修改工作。

服务站系统,是为了数字图书馆服务系统业务拓展需要而开发的,使一些有条件的机构用户可以将数字图书馆服务系统部署到机构本地的服务器上,可自己管理机

构内的个人用户,服务站系统本身不包含文献数据及用户数据,系统通过调用数字图书馆服务系统的WEBSERVICE接口实现用户注册、登录、子站用户管理、文献检索服务、资源导航、服务中心、原文传递服务、网站动态管理等功能。

4.2 知识服务系统

知识服务系统的基础数据来源于国内6000余种中文期刊,从1998年以来全部的文摘和引文数据,数据总量上亿条,其中文摘数据按照论文、作者、机构和基金分别进行了唯一、规范和规整工作,引文数据按照文后参考文献著录规则(GB/T 7714—2005)进行了切分,并在切分的基础上进行了规范和聚类,并与中外期刊论文、会议论文、学位论文、专利等文献类型进行了链接工作,知识服务系统的主要功能有论文引文的关联检索和专项分析(按期刊、按作者、按机构和按基金),并全面展示了论文/引文的引用、被引、同引及同被引关系,知识服务系统的结构见图5。

知识服务系统可以帮助读者寻

找交叉学科,揭示不同学科之间的自然联系,同时可以用来探寻某领域最优秀的期刊、作者、机构及基金资助单位,追溯某一研究课题的起源与历史、最新进展与动态,并揭示了中文期刊与学位论文、会议论文、专利、标准,乃至报纸、图书以及英文文献之间的引证关系。

4.3 专题镜像系统

专题镜像系统是为了方便互联网物理隔离的机构将国家工程技术数字图书馆资源与服务融入本地用户使用环境而推出的一种个性化服务方式。该系统是根据数字图书馆服务系统模式及局域网环境要求设计并开发的,扩展了国家工程技术数字图书馆的服务方式,使国家工程技术数字图书馆的丰富资源和便捷服务可以满足更多用户对科技文献信息的需求,图6为专题镜像系统的技术构架。

专题镜像系统的全文检索模块、订单模块、分词模块均采用成熟开源软件,采用了对Windows系统经过特别优化的全文检索引擎和高性能关系型数据库,在普通服务器上即可实现千万级数据库的全文检索,该系统为绿色系统,服务器只需要简单配置IIS即可,而不需要安装任何数据库系统及其他软件。

5 国家工程技术数字图书馆的关键技术

在国家工程技术数字图书馆各个系统的建设过程中,需要一些关键技术的支持,下面分别介绍一些主要的关键技术。

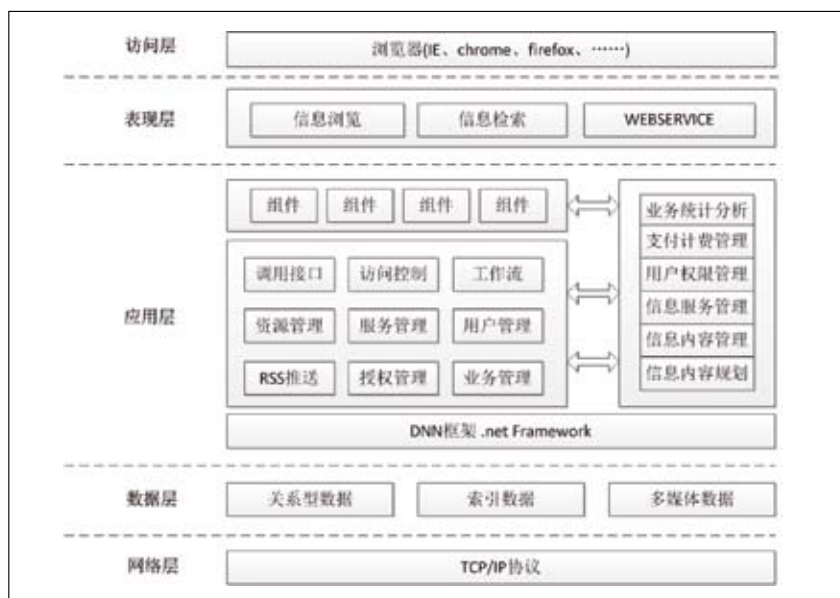


图4 数字图书馆服务系统技术架构

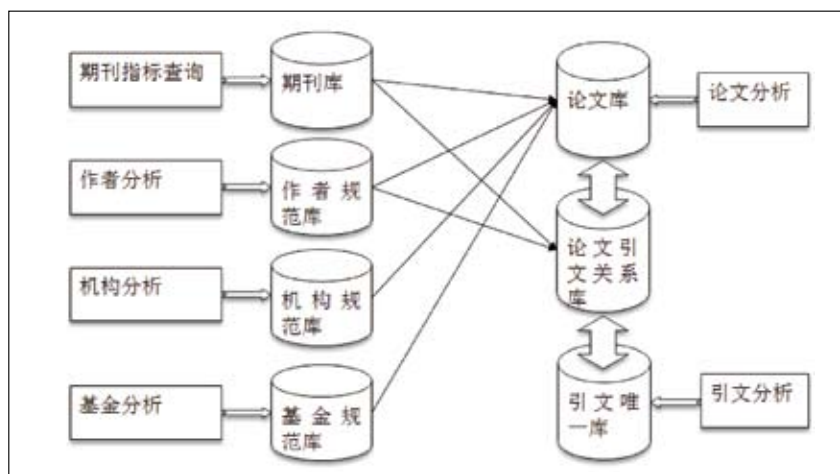


图5 知识服务系统结构



图6 专题镜像系统技术构架

5.1 全文检索技术

用户检索资源是数字图书馆提供的核心服务,全文检索技术是数字图书馆服务必备的核心技术,在国家工程技术数字图书馆的技术体系中,数字图书馆服务系统、知识服务系统和专题镜像系统中都使用了全文检索技术,根据不同系统的特点分别使用了不同的全文检索引擎来满足不同系统的功能需求。

(1) SQL Server全文检索引擎

作为关系型数据库的代表,微软的SQL Server从2005版本开始提供了全文检索功能,其优势是与SQL Server无缝结合,内置分词器并支持断字符,在SQL Server关系型数据库中添加的记录自动填充全文索引,在很短时间内即可做好索引并在服务系统上进行服务,由于是建立在关系型数据库的基础上,其计算、统计和分析功能比一般的全文检索系统要强大得多,知识服务系统使用的就是SQL Server的全文检索技术,利用其特点可以即时计算期刊的各项复杂统计指标,如影响因子、被引半衰期等,其缺点是检索效率不高,需要进行分布式检索部署以提高检索效率。

(2) 开源的Lucene全文检索引擎

Lucene是最广泛使用的开源全文检索引擎,为Apache软件基金会有一个项目,是一个开放源代码的全文检索引擎工具包,即它不是一个完整的全文检索引擎,而是一个全文检索引擎的架构,提供了完整的查询引擎和索引引擎,因此要使用它开发完整的全文检索系统,则需要自行编写数据导入、分词和索

引工具,其优势是检索效率高,基于GPL协议,可免费使用,因此在专题镜像系统中使用了Lucene作为全文检索引擎,实现了单台服务器上千万的文献毫秒返回结果的检索速度。

5.2 中文分词技术

分词是建立索引、进行检索的前提,英文句子中词与词之间有空格,不存在分词问题,而中文只是字、句和段能通过明显的分界符来简单划分,唯独词没有一个形式上的分界符,虽然英文也同样存在短语的划分问题,但是在词这一层上,中文比之英文要复杂和困难得多,中文分词的主要瓶颈是切分排歧和未登录词识别,切分排歧和未登录词降低了自身正确切分的可能性,同时还干扰了其相邻词的正确处理,而且未登录词往往和切分排歧交织在一起,进一步增加了分词的难度,在各种中文分词系统中,基于层叠隐马模型的汉语词法分析系统ICTCLAS较好地处理了这些问题^[5],同时该系统可进行人名、地名和组织机构名的识别,支持多级词性标注,将汉语主题词表作为自定义分词库后,分词的准确性得到了进一步的提高。

5.3 叙词表辅助自动标引技术

自动标引是指利用计算机系统从文本中自动生成若干标引词的过程,又叫做关键词抽取,在国内已经研究和发展了很多年,自动标引从技术角度可分为四类:统计方法、语言学方法、机器学习方法和其他方法,难点集中在分词和抽词

后关键词的筛选上,国家工程技术数字图书馆利用汉语主题词表进行分词,使分词结果较为准确,并利用分词后的词性、汉语主题词表中的属分和用代关系等来进行关键词的聚类 and 筛选,在工程领域的科技期刊自动标引上取得了较好的效果。

5.4 信息可视化技术

在知识服务系统中,统计了上千万的作者合著数据,为了直观地显示作者合著关系的复杂网络,使用基于力导向算法的Flash技术,在线显示了基于特定作者的关系网络,利用Flash的互动功能,可任意切换中心作者,并可点击关系连线以查看合著文献的具体信息,效果见图7。

5.5 服务器虚拟化技术

在数字图书馆建设中,为了提高服务器的利用率,加快应用部署的速度,提供高可靠性、高可用的应用服务^[6],在部署中使用了服务器虚拟化技术,服务器虚拟化(Server Virtualization)^[7]是指多个操作系统在同一时间一起运行在同一台主机上,将服务器物理资源抽象成逻辑资源,让一台服务器变成几台甚至上百台相互隔离的虚拟服务器,虚拟化软件主要有Citrix XenServer、Hyper-V Server 2008、VMware ESX Server等,国家工程技术数字图书馆所使用的是VMware ESX Server,其特点是运行在服务器裸机上,在同一台服务器底层硬件上,划分出若干虚拟机,进行集中管理,并具有集群、负载均衡、热迁移等功能。

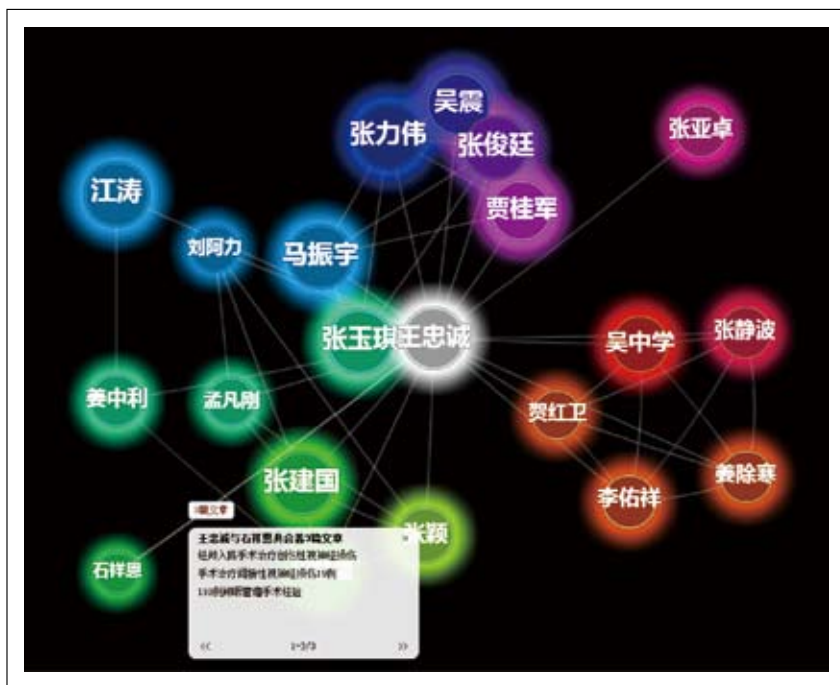


图7 作者合著关系可视化

6 展望

国家工程技术数字图书馆从2007年建设至今,基本建成了完整的数字图书馆技术体系,在全文检索、中西文分词、虚拟化技术、自动标引分类、信息可视化等关键技术方面也有了一定的技术积累,在此基础上,为了进一步增强图书馆的服务能力、优化服务流程、改善服务效率,并逐渐从信息服务转变为知识服务,从信息指引者演变为信息创造者,国家工程技术数字图书馆的技术体系将不断进行改进和完善,在云图书馆、语义检索、知识发现和大数据等方面继续深入研究和探索,以形成知识网络、知识服务为核心的数字图书馆。

参考文献

- [1] 魏蜀华,钱波,陈华.智能检索体在数字图书馆系统中的应用研究[J].情报杂志,2004,23(8):97-98.
- [2] 王卫琴.网络环境下的图书馆与图书馆服务方式[J].理论与现代化,2008(1):122-124.
- [3] 孙一钢.数字图书馆的技术体系结构[J].现代图书情报技术,2001(5):8-10.
- [4] 赵捷,苏静,王星,等.一种数字图书馆集成服务平台的设计与实现[J].情报杂志,2012,31(6):162-167.
- [5] 刘群,张华平,俞鸿魁,等.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004,41(8):1421-1429.
- [6] 陈武,王平.负载均衡技术在数字图书馆服务中的应用模式及实例分析[J].现代图书情报技术,2004(11):1-5.
- [7] BORT J. Server Virtualization: Controlling Server Sprawl [J]. IEEE Communications Magazine, 2003, 41(2): 126-133.

作者简介

王星,男,工程师,研究方向:程序开发、数据挖掘。E-mail: wangxing@istic.ac.cn

Technology System of National Engineering and Technology Digital Library

Wang Xing, Zhao Jie, Liu Minjian / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: The article discusses the selection, acquisition and management, digital processing, storage and integration, monitoring of digital resources, and reveals other aspects of digital information resources services of the National Engineering and Technology Digital Library technology system, and some key technologies used are briefly described.

Keywords: Digital library, Technology system, Key technology

(收稿日期: 2013-09-09)