# 图书馆元数据信息发布平台及其应用

李 艳[1],郝大鹏[2],徐 行[3]
(1. 西安欧亚学院 物流贸易学院,陕西 西安 710065;
2. 西安航空学院 计算机系,陕西 西安 710077;
3. 西安航空学院 科技处,陕西 西安 710077)

**摘 要**:文中研究了图书馆信息服务。信息服务呈现多元化、个性化的趋势,根据需求定制服务十分重要,图书馆信息可以重复利用是服务定制的保障。提出元数据信息发布平台设计方案,发布平台利用中国机读目录抽取信息,并结合互熵-信息检索方法提高抽取信息的正确性,抽取的信息以元数据形式存储,通过 OAI 协议发布。给出发布平台的应用实例,说明信息发布平台如何为毕业设计过程管理系统提供数据服务。

**关键词**:信息服务;元数据;OAI
**中图分类号**:G250 **文献标识码**:A **文章编号**:1673-629X(2014)02-0234-03
doi:10.3969/j.issn.1673-629X.2014.02.058

# Study and Application of Library's Metadata Information Distribution Platform

LI Yan[1], HAO Da-peng[2], XU Xing[3]
(1. School of Logistics Trade, Xi'an Eurasia University, Xi'an 710065, China;
2. Department of Computer, Xi'an Aeronautical University, Xi'an 710077, China;
3. Department of Science and Technology, Xi'an Aeronautical University, Xi'an 710077, China)

**Abstract**: It studies library's information service. Information service is showed a trend of diversification and personalized. Customized service according to demand is very important, it is the guarantee of service customization when library information can be recycled. Put forward a new platform design project for metadata information distribution. The platform extracts information by using CNMARC, improves the correctness of information extraction through using pointwise mutual entropy and information retrieval. Extraction of information is stored in the form of metadata, releasing by using OAI protocol. Giving an application example, discuss how to provide data service for graduation project process management system by using the platform.

**Key words**: information service; metadata; OAI

## 0 引 言

自"服务科学"的内容引入图书馆学界后,使得原有的"以藏为主"的办馆理念发生了本质的变化。随着时代的发展,"以人为本"的思想在图书馆界根深蒂固。由于读者是图书馆存在的唯一理由,所以服务成为了图书馆的核心价值。创新可以使服务的内涵更深更广,从而更大程度地满足读者的需求。基于以上思想,大量的服务创新内容纷纷涌现,资源统一检索平台为读者提供跨平台检索服务[1];在资源不能满足读者需求时,提供文献传递服务[2];为提高与读者之间的交流沟通,提供 living books[3] 服务。但是这些服务具有一个共同的局限性,服务方式和服务内容是由图书馆设定的,这对读者或机构(例如学校内的系部、服务单位)需要根据自身需求获取图书馆信息资源形成了障碍。事实上,读者对信息的需求是根据读者从事的具体活动变换而变换,故而创新服务不断涌现,仍然无法避免用户对图书馆的抱怨[4]。文中提出一种图书馆元数据信息发布平台的设计框架,有了元数据信息发布平台,用户不仅可以直接访问图书馆检索平台进行元数据关联检索,更重要的是机构可以申请成为信息

**收稿日期**:2013-04-26 **修回日期**:2013-07-28 **网络出版时间**:2013-11-29
**基金项目**:陕西省教育科学"十一五"规划 2010 年立项课题(SGH10324);2011 年陕西省高等职业教育教学改革研究项目(11Z32)
**作者简介**:李 艳(1979-),女,河南孟县人,讲师,硕士,研究方向为无线传感器网络、电子商务。
**网络出版地址**:http://www.cnki.net/kcms/detail/61.1450.TP.20131129.0911.025.html

服务的提供者,根据自身需求设定服务,对图书馆信息再利用有重要意义。

## 1　元数据信息发布平台的设计

图书馆拥有大量纸质图书和海量电子资源,但是这些信息以各种数据形式存储在数据库中,信息数据表异构,表内信息形式特殊,信息被提取后仍无法解读,抑制了信息资源的利用。文中所提出的信息发布平台可以提取图书馆各类型的信息,将信息中的知识进行抽取,形成元数据,并存储于元数据数据库中,利用 OAI 协议建立元数据信息发布平台,用户可以向发布平台接口发送数据请求。元数据信息发布平台原理图如图 1 所示。图 1 中虚线左侧为数据库线下操作,图书馆信息数据库与元数据数据库有效隔离保证信息数据库的安全性。实线表示服务器接口连接着服务器与用户,完成服务的请求与响应。下面针对图书数据信息的抽取、元数据生成、发布进行介绍。

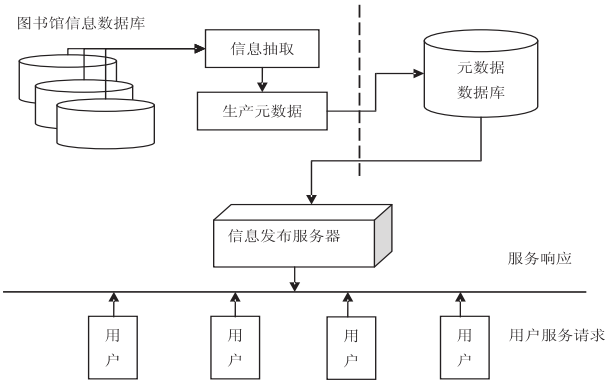


图 1　元数据信息发布平台原理图

### 1.1　信息抽取

虽然图书馆纸质图书管理系统比较多,但是图书信息将在编目阶段形成馆藏 CNMARC 信息,可以利用 CNMARC 的格式对所需信息进行抽取。CNMARC 在数据表中以字符串的形式存储[5],其中前 24 位是 MARC 记录头标区。MARC 记录头标区标识了记录字符串的长度、信息标识和地址。根据 MARC 记录头标区可以获取所需要图书信息的地址,从而获取所需图书编目信息。在信息抽取阶段抽取的信息包括:正题名(200 字段 $a)、第一责任人(200 字段 $f)、出版者(210 字段 $c)、内容提要(300 字段 $a)、主题词(606 字段 $a)、中图分类号(690 字段 $a)。在抽取信息中最重要的是内容提要,内容提要以自然语言形式对图书进行概要介绍,但是内容提要在著录时不是必选字段,往往被编目人员忽视[6]。文中假定图书著录信息中包括完整的内容提要。

### 1.2　生产元数据

元数据是数据的数据,是描述、检索、管理资源的结构化信息,Haslhofer 给出了元数据和现有元数据交互技术的详细综述[7]。元数据标准多种多样,其中 ONline Information eXchange(ONIX)与 MARC 21 Format for Bibliographic Data 标准适合与图书相关的元数据标准。

原则上信息抽取中获得的图书著录信息可以直接利用这两种标准形成元数据,但是图书的主题词不像期刊文献的主题词那么明确,当用户查找某个具体内容时,通过图书名称和主题词查询是不能获得所需结果的。例如,图书馆馆藏一本 ISBN 为 9787040292183、名为《实变函数与泛函分析基础》的图书,当用户需要查找与“可测函数”相关内容的书籍时,由于书名不包括“可测函数”,题词是“实变函数、高等学校、教材”也不包含“可测函数”,用户将不能得到想要的结果。但在内容提要中包括“可测函数”的内容。将图书摘要中有用信息提取到元数据中可大幅增加检索的准度。

根据上述分析,在生产元数据时,需根据语义分析图书内容摘要,提取主题。信息抽取技术方法多样,李保利等对信息抽取技术进行了综述[8]。对于图书内容摘要,一般都会包括“内容包括”、“本书介绍了”、“阐述了”、“主要涉及”等相关词句,需要提取的主题一般都在其后包括,并且用标点符号分割。图书内容摘要抽取方法属于命名实体识别类型(Name Entity Recognition)。

Nadeau 和 Sekine[9]对命名实体识别类型的发展过程和通过学习策略识别命名实体的方法进行总结。其中学习策略对于识别命名实体非常重要,包括监督学习、半监督学习、非监督学习。对于海量图书内容摘要信息,监督学习和半监督学习都需要大量的人工干预。近些年,非监督学习在识别命名实体方面成为了热点,其中 Etzioni 等[10]提出的逐点互熵-信息检索(Pointwise Mutual Entropy and Information Retrieval,PME-IR)最为出众,在无监督情况下,正确抽取率达到 80%。

PME-IR 为在线信息抽取方法,其核心思想是利用“问题”与“选择信息”的共现性(co-occurrence),通过互熵来决定选择信息的准确性,并将准确性最高的选择信息作为抽取信息。选择信息的准确性的逐点互熵为:

$$\mathrm{score}(\mathrm{choice}_i) = \log 2(A)$$

$$A = \frac{P(\mathrm{problem\&choice}_i)}{P(\mathrm{problem})P(\mathrm{choice}_i)}$$

文中在线下组织可能出现的提问,利用了 PME-IR 作为图书内容摘要的信息抽取方法,批量提取出所需主题。并将提取出的主题加入主题词中,最终形成元数据。元数据结构如下:

元数据的结构:={题名,作者,出版社,中图分类

号,主题词}。

### 1.3 数据服务发布

虽然数据服务发布协议多种多样,但是针对图书馆元数据,美国图书馆和信息资源委员会和数字图书馆联盟提出了一种高效的“收获式”发布协议,即元数据获取协议(OAI)[11-12]。OAI 工作原理如图 2 所示。

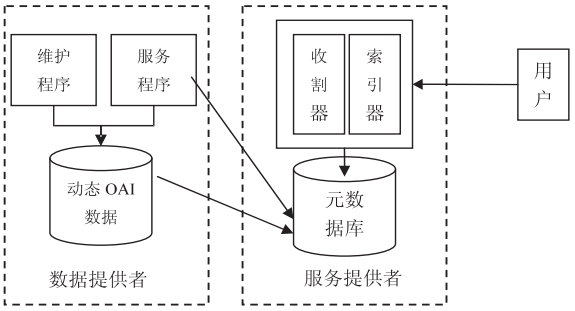


图 2 OAI 工作原理图

其中数据提供者为图书馆,服务提供者可以是图书馆也可以是其他机构,在面向用户服务时,机构可以申请成为服务提供者,获取数据方法的权力,根据自身需求针对元数据进行开发,这样图书馆的资源会更广泛地被利用。

由于图书馆的信息资源数据库对于图书馆而言非常重要,出于安全性的考虑,数据抽取和元数据生成为离线数据处理,并与元数据库分离,图书馆管理员人工定期动态更新元数据库。

针对元数据库的发布,UCAR 机构开发了基于 JAVA 的开放式 JOAI 开放包,在用途为非盈利目的时,可以免费使用,并且拥有完整的 API 文档,使用方便。JOAI 已成功应用于美国罗德岛大学海洋科学教育卓越中心(http://www.cosee.net)、新泽西地球科学教师联盟网站(http:// Njesta.org)等知名教育机构,文中的项目就使用了 JOAI 实现元数据的发布。

## 2 元数据信息服务平台的应用

信息服务将“资源是本,行动是纲,服务是魂”作为基本准则,图书馆拥有大量资源使它成为信息服务的有力保障,但是服务需求的多元性和灵活性,更多机构参与到信息合作服务中,才能使信息服务做到更好。

在设计研发毕业设计(论文)过程管理平台的过程中需要教务处、科研处的联合研发,但是毕业设计过程中无论学生还是指导教师都需要获取图书馆信息资源,并且每次毕业设计指导过程都会根据毕业设计的内容而不同。指导教师根据自己的需求检索图书馆数据库,提出毕业设计的选题,并根据自身经验获取一些学生必读的参考文献,最终将选题和与选题相关的参考文献放置在管理平台中等待学生的访问。学生也会在选择毕业设计题目后,根据自身需求检索图书馆数据库。但是检索的结果一般不能保存,并且直接检索图书馆数据库的准确度不高,对于一般的专科或本科生而言,检索所需资料比较困难。当图书馆提供元数据服务发布平台后,情况则会不同。图书馆处理好图书馆信息资源相对应的元数据,并将访问元数据的权力交给毕业设计(论文)过程管理平台,那么指导教师和学生都可以向过程管理平台申请自身所需元数据,间接地访问了图书馆信息资源。

对于图书馆参与毕业设计(论文)过程管理平台是非常重要的,如果没有图书馆的参与,那么过程管理平台就成为了一个 OA 系统,不能起到应有的作用。相反对于图书馆而言设计元数据信息服务平台不仅仅可以为毕业设计(论文)过程管理平台提供有效的数据服务,而且一次开发,其他的合作机构也可以利用元数据信息服务平台访问优质元数据。

元数据信息服务平台如何作为数据的提供者为教师与学生提供服务,如图 3 所示。系统框图中元数据信息服务平台是数据提供者,毕业设计(论文)过程管理平台是服务提供者,学生与教师通过毕业设计过程管理平台向元数据信息服务平台请求文献查询服务。

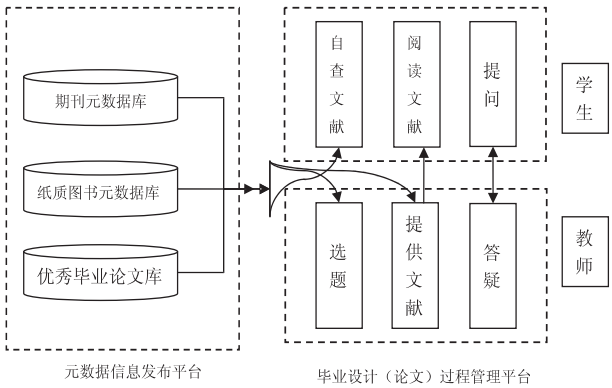


图 3 系统框图

## 3 结束语

图书馆是学校内部获取信息的重要场所,可以为教学与科研提供服务保障。虽然大多数高校图书馆都已有学科馆员或系部联络馆员,但是人员数量和服务质量还不能完全保障教学和科研,更无法达到图书馆“嵌入式”服务的目标。

文中设计了元数据信息发布平台,介绍了信息抽取、元数据生成,元数据发布的可行办法,给出了元数据信息发布平台与毕业设计(论文)过程管理平台相结合的案例。元数据信息发布平台作为数据服务的提供者可以与其他机构进行服务项目的合作,拓宽图书馆信息服务的领域,并且可以在一定程度上缓解学科馆员或系部联络馆员的人员不足。设计服务平台并与

(下转第 241 页)

间直线交点的方法近似地计算出目标物的空间三维坐标,再判断出目标人物是否在安全空间内。由于摄像头成像能力的限制,以及测量距离的误差,利用该方法所测得距离与实际距离之间会存在一定的误差,但是完全可以满足使用需求。并且该方法抛弃了传统双目视觉测距所需要的通过特殊标定板进行复杂的摄像头标定工作,使得此系统具有安装简单、高实时性等特点,具有一定的指导意义。

**参考文献:**


[1] Zhao Chunxia, Chang Wensen. Design and implement of multi-functional monitoring system for the suspension and guidance system of maglev train[C]//Proc of international conference on measuring technology and mechatronics automation. [s. l.]:[s. n.],2009:20-22.

[2] Yao Yachuan, Yao Yi, Song Hong. The remote monitoring system based on the OPC technology[C]//Proc of international workshop on intelligent systems and applications. [s. l.]:[s. n.],2009:18-20.

[3] 杨　武,王小华,荣命哲,等. 基于红外测温技术的高压电力设备温度在线监测传感器的研究[J]. 中国电机工程学报,2002,22(9):113-117.

[4] 祝晓辉. 基于图像处理的电力设备识别方法研究[D]. 保定:华北电力大学,2007.

[5] 张　浩. 图像识别技术在电力设备在线监测中的应用[D]. 北京:北京交通大学,2009.

[6] Bradski G, Kaehler A. Learning OpenCV[M]. USA: O′Reilly Media,2008:375-378.

[7] 舒　文. 实时场景下的运动目标检测技术研究[D]. 成都:西南交通大学,2007.

[8] 裴巧娜. 基于光流法的运动目标检测[D]. 北京:北方工业大学,2009.

[9] 王　欢. 运动目标检测与跟踪技术研究[D]. 南京:南京大学,2009.

[10] 杨叶梅. 基于改进光流法的运动目标检测[J]. 计算机与数字工程,2011,39(9):108-110.

[11] Talukder A, Matthies L. Real-time detection of moving object from moving vehicles using dense stereo and optical flow[C]//Proc of IEEE conference on intelligent robots and systems. [s. l.]:[s. n.],2004.

[12] Wixson L. Detecting sailient motion by accumulating directionally consistent flow[J]. IEEE trans on pattern analysis and machine intelligence,2000,22(8):774-780.

[13] Lipton A, Fujiyoshi H, Patil R. Moving target classification and tracking from real-time video[C]//Proc of IEEE workshop on application of computer vision. Princeton, NJ: [s. n.], 1998:8-14.

[14] 王孝艳,张艳珠,董慧颖,等. 运动目标检测的三帧差法算法研究[J]. 沈阳理工大学学报,2011,30(6):83-86.

[15] 朱明旱,罗大庸,曹倩霞. 帧间差分与背景差分相融合的运动目标检测算法[J]. 计算机测量与控制,2005,13(3):215-217.

[16] 曹丹华,邹　伟,吴裕斌. 基于背景图像差分的运动人体检测[J]. 光电工程,2007,34(6):107-111.

[17] Gupte S, Masoud O, Martin R F K, et al. Detection and classification of vehicles[J]. IEEE transactions on intelligent transportation systems,2002,3(1):37-47.

[18] Hearn D, Baker M P. Computer graphics with OpenGL[M]. 3rd ed. London: Prentice Hall,2003:438-455.

[19] 郭　星,刘政怡,李　炜,等. 一种大屏幕人机交互系统的实现方法[J]. 计算机工程与应用,2012,48(1):176-179.

[20] 徐　杰,陈一民,史志龙. 双目视觉变焦测距技术[J]. 上海大学学报(自然科学版),2009,15(2):169-174.


(上接第 236 页)

元数据信息发布平台相关联,可以成为图书馆为教学科研提供服务保障的一种途径。

**参考文献:**


[1] 张晓娟,望俊成,张洁丽,等. 我国信息资源整合的研究进展分析[J]. 情报科学,2009,27(10):1545-1550.

[2] 杨　敏,张　斌. 浅议文献传递服务[J]. 图书馆学研究,2005(4):80-82.

[3] 李　菲,徐恺英,孙　岩,等. 基于“Living books”的图书馆潜在知识转移模型构建[J]. 情报科学,2011,29(12):1889-1891.

[4] 金业阳. 高校图书馆用户抱怨行为研究[J]. 情报理论与实践,2005,28(4):406-408.

[5] 毛有桂. 21 世纪的 MARC 格式-MARC2[J]. 图书馆建设,2003(3):43-45.

[6] 贾宇群. CNMARC 格式中 330 字段的著录[J]. 大学图书情报学刊,2009,27(5):54-56.

[7] Haslhofer B, Klas W. A survey of techniques for achieving metadata interoperability[J]. ACM computing surveys,2010,42(2):1-42.

[8] 李保利,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用,2003(10):1-6.

[9] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae investigationes,2007,30(1):3-26.

[10] Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the Web: An experimental study[J]. Artificial intelligence,2005,165(1):91-134.

[11] Lagoze C, van de Sompel H. The making of the open archives initiative protocol for metadata harvesting[J]. Library hi tech,2003,21(2):118-128.

[12] Maslov A, Creel J, Mikeal A, et al. Adding OAI-ORE support to repository platforms[J]. Journal of digital information,2010,11(1):1368-1376.

# 图书馆元数据信息发布平台及其应用



作者： 李艳， 郝大鹏， 徐行， LI Yan， HAO Da-peng， XU Xing

作者单位： 李艳,LI Yan(西安欧亚学院 物流贸易学院,陕西 西安,710065)， 郝大鹏,HAO Da-peng(西安航空学院 计算机系,陕西 西安,710077)， 徐行,XU Xing(西安航空学院 科技处,陕西 西安,710077)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(2)