●孔敬 李广建

学科信息门户：概念、结构与关键技术

摘要 以信息与应用高度集成、个性化和智能化为显著特征的新一代学科信息门户系统框架正在成为构建基于 WWW 的专业数字图书馆的理想模式。网络信息技术如门户构件、跨系统集成检索、元搜索引擎和智能代理等技术的迅速发展，为构建学科信息门户提供了关键技术支撑。图 4。参考文献 9。

关键词 学科信息门户 体系结构 门户构件 跨库检索
分类号 G254

ABSTRACT With the characteristics of high integration, personalization and intelligence, a new generation of subject information portals are becoming ideal patterns of WWW-based subject digital libraries. The rapid development of network information technologies, such as portlets, cross-database search, metabata search engines and intelligent agents, provides key technological supports for the construction of subject portals. 4 figs. 9 refs.

KEY WORDS Subject portal. Architecture. Portlet. Cross-database search.
CLASS NUMBER G254

1 学科信息门户的概念与发展

1.1 因特网环境下门户概念的兴起

近年来，随着因特网上网站数和网页数量的迅猛增长，门户网站(Web Portal)应运而生。门户网站提供个性化的单点网络接入，对本地和远程的信息进行集成，使信息高度组织化，以便用户查找和发现信息。它最大的特点是针对特定的用户，提供与他们相关的内容集成和信息推送，以及业务协作、社区服务等相关应用和服务的访问，为用户提供一种获取相关信息的简洁方式，同时又避免了信息过多的问题。它为当前海量分布式信息资源的采集、存储与管理提供了令人关注的解决方案。越来越多的机构正在将自己的网站转变成具有门户特征的网站。门户网站也已由最初的以内容汇集和搜索引擎为特征的初级门户发展为今天的以智能化、个性化、信息与应用高度集成为特征的高级门户概念。

1.2 学科信息门户概念的提出

专业信息资源建设和服务的网络化发展与新兴的门户网站概念相结合促成了学科信息门户的产生，使学科信息门户成为专业信息资源共建共享的新模式。简单说，学科信息门户(Subject Portal，简称 SP)是用户访问某学科资源与服务的一个单一入口或通道。它是一种网络服务，用以完成本学科网络资源内容的高度组织集成和网络应用程序的聚集，并将这些资源与应用集成在一个可定制个性化的界面中来满足每个最终用户的需要。它还提供一个统一协作的学术交流环境。从用户角度来看，它是某学科用户访问该学科网络资源和服务的起始站点或称入口。

1.3 学科信息门户的发展

学科信息门户的概念可追溯到由 T. Koch 等人提出的学科信息网关(Subject Gateway，简称 SG)概念。由于 IT 界 Portal 概念的兴起，学科信息网关的建设已逐渐转向为学科信息门户的形式。对学科信息门户的研究、试验与推广，在欧洲进行得最为深入和广泛。例如：欧洲范围内开展的 Desire 项目、Renardus 项目和英国的 RDN(Resource Discover Network)项目就是其中的典型。Desire 项目于 1998 至 2000 年间实施，该项目联合了来自荷兰、挪威、瑞典和英国 4 个欧洲国家的 10 个机构共同协作工作。Renardus 项目于 2000 年启动，联合了欧洲多个国家的国家图书馆、大学研究机构与技术中心，以及各个学科信息门户，提供了包括大部分学科领域 64000 个重要学术网站的资源检索与浏览服务。RDN 学科信息门户(The RDN Subject Portals Project)由英国的 JISC(The Joint Information Systems Committee)资助，目前共建立了社会科学类(SOSIG)、工程、数学与计算科学类(EEVL)、健康与生命科学类(BIOME)、物理科学类(PSIGate)、人文科学类(Humbul)、工艺美术类(ARTIFACT)、休闲娱乐体育旅游类(ALTIS)和地理环境类(GEsource)

等八大学科信息门户。澳大利亚建立了国家级的多机构联合的学科信息门户，美国也建立了众多的学科信息门户。美国威斯康星—麦迪逊大学还开发了SPT(Subject Portal Toolkit)学科信息门户免费软件包，已被广泛用于学科信息门户的开发。但项目规模较大，且对学科信息门户进行了全面系统研究的地区还是在欧洲，尤其是英国。我国国内一些专业图书馆也做了初步尝试。如中国科学院的国家科学数字图书馆门户及其子门户物理数学学科信息门户、化学学科信息门户、生物学科信息门户等，以及武汉理工大学图书馆的材料复合新技术信息门户。综合来看，当前国内外学科信息门户多处在功能不完善的中低级阶段，个性化、智能化和高度集成的高级信息门户还处在实验阶段。

2　学科信息门户的特征

从技术上说，学科信息门户就是指采用多种信息技术，诸如跨系统检索、元数据采集技术等，对分散的分布式的学科网络信息资源进行收集、分析、整理和合并，将整合后的内容集成到一个可定制的个性化的界面中呈现给用户。这个界面可以是Web浏览器，也可以是其他可能的方式。

学科信息门户的核心特征有以下几个。

信息和应用的集成整合：信息内容经过深层次组织加工，形成高质量的信息内容。这些信息与各种信息服务有机地集成在一个统一的界面中。

跨系统一站式检索：用户在一个搜索界面，将搜索请求一次性输入，就可实现对多种资源和数据库信息的查询。它将各个系统的检索结果汇集起来，以统一的界面展示给用户，使用户的搜索方便而高效。而普通网站通常并未提供这种跨系统检索功能，用户不得不分别进入各个本地的或远程的检索系统来进行检索。

简单统一界面：通过共同的表达和一致的用户界面，使门户更易于使用。由于界面统一并遵循用户习惯，用户无需进行培训就能方便地发现和搜索到所需信息。

单点登录，一次性认证：用户只需要一次登录，就能使用他已得到授权的各种资源和服务，而无须记住和输入众多不同资源与服务的账号和口令。

可定制：门户根据不同的角色预设了不同界面内容。可基于用户所属的角色来提供给用户相应的内容。

个性化：根据用户需求与偏好的描述信息，或通过用户信息访问行为的动态分析来推测用户意图，进行信息过滤和信息推荐，对不同用户提供不同的内容和用户界面。

协作性：一系列门户的协作服务，如即时消息传递和团队室的访问。帮助团队共享门户网站页面、应用程序、文档、消息传递和其他协作工具。

安全性：门户采用安全性策略管理以确保用户安全地进行各种活动。

工作流：实现学科信息业务流程的集成。如学术信息的网上发布、编辑、评注和信息分析研究等等。

3　学科信息门户的体系结构

学科信息门户的框架由用户界面层、个性化可定制门户构件层、信息与应用集成层、信息结构层和信息资源层组成。其层次结构如图1所示。

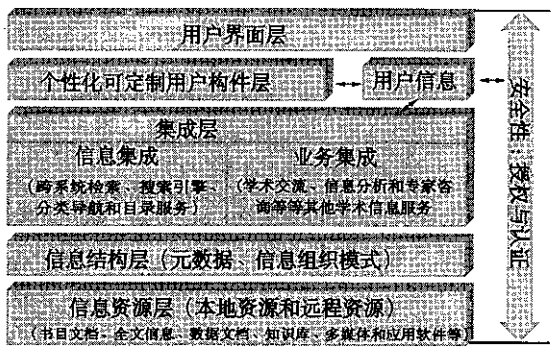

图1　图书馆门户的层次结构

用户界面层：它是学科门户网呈现给用户的表现形式。它通常包括这样一些功能组件，如跨系统数据库检索、搜索引擎、信息分类导航、学术交流、专家咨询、个性化可定制服务等。

个性化与可定制门户构件层：门户构件(Portlets)是一些能生成网页片段内容的，以Java技术为基础的Web组件。它运行在门户服务器(Portal Server)中，被插入运行于网页程序中，用来设计和构建聚合了大量内容的组合页面。门户构件层根据系统的用户信息库(User Profiles)中保存的用户信息需求和权限信息，调用不同的门户构件，针对不同用户产生不同的页面信息内容。

信息与应用集成层：在信息集成方面，常用技术和策略有跨系统数据库检索、元搜索引擎和分类导航。跨系统数据库检索着重于对分布式的异构数据

库进行信息检索与整合。元搜索引擎、智能代理和分类导航用于www信息资源的采集与整合。在应用集成方面,可考虑将学术交流、信息分析、专家咨询和个性化可定制服务等应用集成于此。学术交流用于动态交互地实现学术信息的网上发布以及对信息进行编辑批注和评述,它和其他业务逻辑组件如专家咨询组件等构成了数字信息服务和业务流程的集成。

信息结构层:即元数据格式和信息组织方式。元数据的格式有很多种类,如:通用元数据格式 Dublin Core 元素集,描述文献的 MARC 元数据格式,描述图像的元数据 MOA2,描述教育资源的元数据格式 IEEE LOM 和 GEM,政府信息资源元数据格式 GILS 等。在数字图书馆中最常用的是 MARC 和 DC 元素集。学科信息门户的资源组织应尽量遵循本学科的元数据规范和复用一些通用的元数据格式。

信息资源层:包括本地和远程的书目文档、全文信息、数据文档、知识库、多媒体和应用软件等多种类型信息资源。

学科信息门户的体系结构如图2所示。

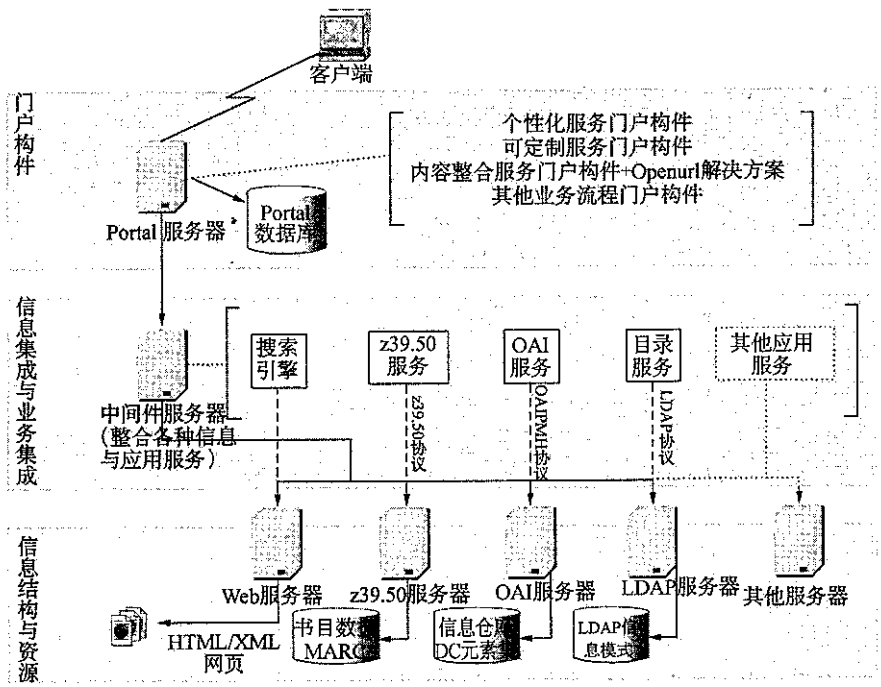

图2 学科信息门户体系结构

4 学科信息门户的关键技术

从学科信息门户的层次结构来看,涉及的技术主要有门户构件技术、数字资源集成整合技术、数字资源内容管理技术、安全技术等,其中又以对现有各种分布式异构信息系统的集成整合为重点。因为这些分布式异构系统通常已有各自的数字资源内容管理技术,所以本文将重点讨论构建学科门户最为关键的技术,即集成技术,如:面向用户界面集成整合的门户构件技术,面向各个数据库系统的集成检索技术,面向Web信息源集成搜索的元搜索引擎技术和智能代理等。

4.1 门户构件技术

门户构件技术使得用户界面的组件化成为可能,它建立在Web服务技术体系基础之上,用于共享和组合网页页面内容,从而实现本地与远程的数据与应用集成。

门户构件在许多方面都类似于Servlet。门户构件是用Portlet API来编写,就像Servlet用Servlet API来编写一样,不同的是运行在Portal容器(Portlet Container)中,而Servlets运行在服务器端的Servlet容器(Servlet Container)中;门户构件只生成网页代码片段,不生成整个网页文件;另外,Servlet直接与客户端通信,而门户构件则通过门户服务器的应用来调用。

进一步说,门户构件是运行在Portal服务器的Portal容器中的Web组件。Portal服务器将这些包含

了本地和远程的数据源或应用的组件即 Portlets，集成为 Web 服务。这些包含了各种数据和业务的 Portlets 可能来自于其他的数据库和业务系统、内容供应商或远程 Web 站点。Portal 服务器一方面将这些 Portlet 以 Web 服务方式在网络上进行注册、发布、请求和调用；另一方面将这些 Portlet 综合起来形成复杂页面，以容易被用户接收的形式返回给用户，是用户从不同位置访问不同信息和应用的焦点。图 3 展示了通过门户构件技术集成本地和远程网页信息内容的机制。

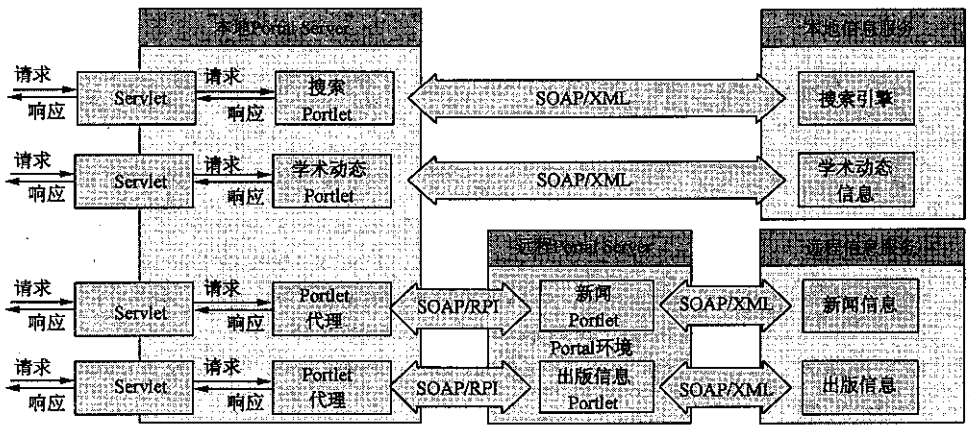

图 3　Portal 服务器、Portlets 集成本地和远程信息内容

4.2　跨库检索技术

跨库检索是门户的重要组成部分。跨库检索以分布式异构数据源为对象，通过统一的检索接口接收用户查询，将用户查询转化为不同数据源的检索表达式，并发检索本地和广域网上的多个分布式异构数据源，在经过去重和排序等操作后，以统一的格式将结果呈现给用户。

目前用于跨库检索的互操作协议标准有：Z39.50、OAIPMH、SOAP、OPENURL 和 LDAP，以及即将确定的用于分布式检索的标准 W3C XML Query Language。Z39.50、OAI、SOAP、OPENURL、LDAP 和 XQuery 各协议的关系可用图 4 表示。由图可知，LDAP 检索的信息为 LDAP 树状结构信息模式；传统的 Z39.50 是基于 TCP/IP 协议传输，主要用于检索 MARC 格式数据；OAI、OPENURL 和 XQuery 以及新一代 Z39.50；ZING 的改良方案都是基于 HTTP 和 XML 而构架的，可用于检索都柏林核心元数据集和其他元数据格式描述的信息。其中 OPENURL 提供了导引用户从当前资源链接到相关参考资源的详细内容的解决方案，其核心内容上下文对象（Context Object）可通过 HTTP GET/POST、SOAP、OAI-PMH 3 种方式传输。

上述这些标准和协议主要解决了分布式异构数据集成检索以下几个层次的互操作性：一是信息编码的互操作性，如 XML、DC 元素集、MARC 和其他元数据格式规范等；二是信息传输与通信的互操作性，如 TCP/IP、HTTP 和 SOAP 协议等。三是信息检索的互操作性，如基于元数据检索的 OAI、Z39.50 及其改良方案 ZING，基于关联链接检索的 Openurl，用于 XML 数据库的 XQuery 和自成体系的轻量级目录访问协议 LDAP 等。

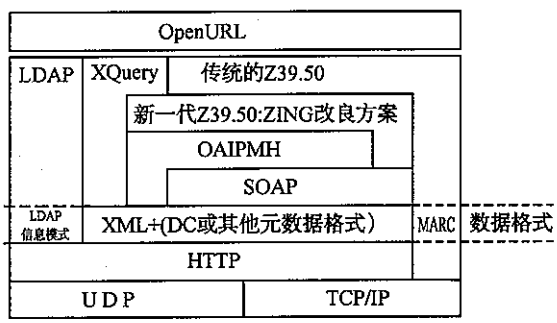

图 4　各种资源整合和互操作协议关系

4.3　元搜索引擎和智能代理

元搜索引擎和智能代理技术是构建学科信息门户的重要技术。这两项技术的目标是实现对因特网网页的信息集成检索和智能化个性化检索。

元搜索引擎就是对多个独立搜索引擎的整合、调用、控制和优化利用。相对元搜索引擎，可被利用的独立搜索引擎称为“源搜索引擎”　（下转第 90 页）

参考文献

1 原宏盛. 影响中国21世纪图书馆事业发展与变革的重大举动. 图书馆理论与实践,2000(5)
2 维普"中文科技期刊数据库". 苏州电大敬德图书馆. http://www. sztvu. com/lib(查询于2005-05-04)
3 吴慰慈. 从传统图书馆学向现代图书馆学的转型与过渡——1996~2000年的中国图书馆学基础理论研究. 图书馆,2001(1)
4 原宏盛. 论修订《普通高等学校图书馆规程》的历史背景和重大意义. 图书馆理论与实践,2004(2)
5 原宏盛. 标志中国图书馆事业发展的里程碑. 图书馆理论与实践,2001(1)
6 刘兹恒,张久珍. 国内外虚拟图书馆研究综述. 中国图书馆学报,2000(3)
7,17 吴慰慈. 图书馆学基础理论研究述评(1995-2004年). 中国图书馆学报,2005(2)
8 朱强. 高校图书馆当前形势和任务. 在全国高职高专图书馆工作经验交流会上的报告. 南京,2005-04-06
9 张玉霞,王元忠. 传统图书馆与虚拟图书馆比较研究. 中国图书馆学报,2002(2)
10,14 原宏盛. 推动中国图书馆事业发展的强大动力. 图书馆理论与实践,2003(3)
11,15 原宏盛. 20世纪末与21世纪初图书馆事业的发展特点及其主要标志. 图书馆论坛,2005(4)
12 胡群耘. 信息时代,图书馆会走向消亡吗——访上海图书馆副馆长吴建中博士. http://www. booker. com. cn/gb/paper24/3/class002400001/hwz14113. htm(查询于2005-05-07)
13 原宏盛. 苏州市域图书馆新馆群体建设启示论. 图书馆建设,2005(5)
16 原宏盛. 世纪之交中国图书馆事业进程中的重大历史事件. 图书与情报,2004(3)

原宏盛 苏州电视大学·苏州职工大学 敬德图书馆馆长,副研究馆员。通信地址:苏州市干将西路1122号。邮编215004。

(来稿时间:2005-05-08)

(上接第53页) 或"搜索资源",整合、调用、控制和优化利用源搜索引擎的技术称为"元搜索技术"。因此,在学科信息门户可采用元搜索技术来实现WWW信息集成检索。

智能代理又称智能体,是人工智能研究的新成果,它是在用户没有明确具体要求的情况下,根据用户需要,代替用户进行各种复杂的工作,如信息查询、筛选及管理,并能推测用户的意图,自主制定、调整和执行工作计划。具有智能性,是可进行高级、复杂的自动处理的代理软件。智能代理技术可应用于学科信息门户中实现智能化、个性化信息检索服务。

5 结语

以信息与应用高度集成、个性化和智能化为显著特征的新一代学科信息门户系统框架正在成为构建基于WWW的专业数字图书馆的理想模式。网络信息技术如门户构件、跨系统集成检索、元搜索引擎和智能代理等技术的迅速发展,为构建学科信息门户提供了关键技术支撑,成为其实现的基础。然而理想的一站式信息检索在实践应用中存在诸多问题,在高效和可用性方面还有待完善。对跨系统集成检索技术的研究在未来一段时间内仍然是计算机界和图书情报界关注的焦点。门户构件在互操作性上还有很多路要走。尽管学科信息门户在实践应用上存在与理论技术滞后的问题,关键技术还有待提高,但随着学科信息门户概念的普及推广,学科信息门户关键技术的深入研究和广泛应用,新一代学科信息门户将成为专业数字图书馆共建共享的首选模式。

参考文献

1 http://www. rdn. ac. uk/publications/terminology/
2 Mary Jackson. The advent of portals. Library Journal,2002-09-15
3 Lesly Huxley. Renardus: Building an Academic Subject Gateway Service in Europe. In CUC 2002-2nd CARNet Users Conference. Zagreb, Croatia,2000
4 张诚. Web服务之路越走越亮. 计算机世界报,2002(7)
5 于学锋,单启成. 下一代Z39.50技术探讨. 现代图书情报技术,2003(2)
6 张晓林. 分布式数字图书馆机制. 情报学报,2002(2)
7 张晓林. 元数据研究与应用. 北京:北京图书馆出版社,2002
8 张晓林. 数字化信息组织的结构与技术(1). 大学图书馆学报,2001(4)
9 张晓林. 数字化信息组织的结构与技术(2). 大学图书馆学报,2001(5)

孔 敬 中国科学院文献情报中心博士研究生。通信地址:北京北四环西路33号。邮编100080。

李广建 中国科学院文献情报中心信息技术部主任,博士生导师。北京师范大学管理学院教授。通信地址同上。

(来稿时间:2004-12-23)