●董慧 雷瑛 陈琮 杨宁

构建基于 J2EE 规范的数字图书馆模型的探讨*

摘 要 从软件工程的角度分析,数字图书馆实质是一个分布式管理信息系统。构建数字图书馆,就是构建一个分布式管理信息系统。J2EE 的多层次的分布式应用模型和一系列开发技术规范,为构建数字图书馆平台提供了可靠保证。图 1。参考文献 11。

关键词 数字图书馆 数字图书馆模型 J2EE 规范 分布式管理信息系统
分类号 G250.76

ABDTRACT From the point of view of software engineering, digital library is a distributed management information system. To construct a digital library is to construct a distributed management information system. J2EE's hierarchical distributed application models and a series of development technological specifications provide supports for the construction of platforms of digital libraries. 1 fig. 11 refs.

KEY WORDS Digital library. Model of digital library. J2EE specifications. Distributed management information system.
CLASS NUMBER G250.76

1 数字图书馆与 J2EE 规范

J2EE(Java2 Enterprise Editor)规范是一个开发分布式企业级应用的规范,它提供了一个多层次的分布式应用模型和一系列开发技术规范。这些技术规范覆盖了企业级应用方方面面,为建立安全、健壮的企业级应用提供了强大的支持。同时,这个规范为 SUN、IBM、BEA、ORACLE 等众多公司参与制定的一个开放的标准,这种开放性保证了基于这个标准而开发的企业级应用可以摆脱对具体一个公司或者一种技术的依赖。

随着数字图书馆的发展,对性能、分布式、安全、事务、健壮性的要求越来越向企业级应用的要求靠拢。J2EE 规范的提出,为数字图书馆构建提供了好的借鉴和参照。怎样在数字图书馆中应用 J2EE 规范,怎样利用 J2EE 提供的优良特性,是构建数字图书馆一个新的探索方向,对发展中国的数字图书馆事业具有特殊意义。

在 J2EE 规范中,提供了声明性安全和可编程性安全两种方式,提供了完整的认证与授权(Authentication and Authorization)机制,而且,EJB 容器所支持的 EJB 组件的安全性,可以从整个组件到具体的类乃至具体的一种方法控制用户的访问权限。这些为解决数字图书馆的安全性问题提供了方便高效的支持。

对于现在研究的热点——数字图书馆分布式应用的问题,在 J2EE 中也有众多规范支持。规范中远程方法调用(RMI),使客户机上运行的程序可以调用远程服务器上的对象,EJB 组件将这种应用更加简化,让容器实现分布式应用中远程方法调用,开发人员不必关心远程方法调用的细节。同时基于 XML 的远程方法调用(RPC)的 Java API(JAX-RPC)以及基于 XML 消息传递的 Java API(JAMX),使得分布式应用的开发选择更多。

其他数字图书馆中提出的问题,比如遗留系统的问题,J2EE 规范中提供了 J2EE 连接器(JCA),允许基于 J2EE 的系统与遗留系统相互访问。J2EE 规范中定义的 JMS 解决了未耦合的应用程序之间可靠地异步通信,解决系统之间异步的、非阻塞的消息传递。

2 基于 J2EE 平台的数字图书馆内核

在国家社会科学基金资助项目"数字图书馆相关关键技术研究"中,我们选择了 J2EE 平台作为系统的开发和运行平台。这是应用 J2EE 规范构建数字图书馆的一次有益尝试。

* 本文系国家社会科学基金项目"数字图书馆相关关键技术研究"(批准号:00BTQ004)的成果。

2.1 系统的功能

整个系统从功能上划分为两个模块:后台文献的加工、存储、索引的处理模块,面向用户的检索模块。

后台处理模块:包括将原始纸质文档扫描加工成PDF格式文档,针对PDF文档提取关键词、标题等元数据存入数据库,并将PDF文件存入数据库,最后对文档和元数据信息进行索引,建立索引文件。

检索业务模块:检索的核心业务包括为用户提供检索功能以及围绕检索而展开的相关功能,如保存检索策略、保存检索的结果文档、根据检索结果文档关键词的统计信息进行浏览。用户在浏览器中输入检索条件,然后向服务器提交检索请求。服务器接受到检索请求后,检索索引文件或者数据库,进一步构造检索结果和检索结果页面,返回给客户端。客户端根据返回的结果,可以选择浏览检索结果,保存检索策略以及根据关键词的统计信息浏览检索结果。

2.2 系统设计思想以及特色

整个系统是基于J2EE平台,采用J2EE的多层体系结构,细化各层的功能。系统核心业务比如检索业务、索引业务采用EJB组件实现,利用EJB容器提供的安全性、事务性、生命周期管理、线程安全等特性,同时提高了移植性、灵活性等。系统提供全文检索的功能,同时在检索结果的表达方式上进行了有益探索,为用户提供了可视直观的表达方式。

2.2.1 采用J2EE的多层体系结构,分化各层的功能

系统分为客户端层、Web层、业务层、资源层四层,客户端层提交检索请求,在客户端显示检索结果,以及构造交互界面,即浏览器中的HTML页面。Web层响应客户端请求,构造检索输入界面和检索结果界面,返回给用户端,即服务器端的JSP和Servlet,负责与业务层的业务组件交互。业务层实现数字图书馆的核心业务,包括检索和索引,访问数据库和索引文件,以及提供访问这些核心业务的接口,即服务器端的EJB组件。资源层,包括数据库中数据和索引文件。

这种多层的体系结构,具有如下优点。

低耦合性。在J2EE多层体系结构中,分离了检索的业务逻辑代码和表示逻辑代码。业务逻辑层中的EJB组件提供访问的接口,表示逻辑层Web组件通过访问这些接口,使用业务逻辑层所提供的功能。当业务逻辑层的部分功能变更时,如果提供给表示逻辑层的接口不变,表示层表示逻辑则不需更改。例如:检索组件为Web组件提供了检索的入口。当检索的算法需要改进时,如果检索算法的改进不影响到组件所提供的接口,表示层业务逻辑则不需要更改。这使得业务层与Web服务器的表示层分离,从而两者之间是低耦合度的。同样,如果Web服务器的表示层根据需求进行变动,不会影响到业务逻辑层的检索组件。这种低耦合性可以使得程序实现上分工明确,在需求设计阶段定义好接口之后,业务逻辑实现者可以专注业务部分,表示逻辑实现者可以专注表示部分。

安全性。这种体系结构的安全性体现在多个方面。首先,在这个多层体系结构中,业务逻辑实现为EJB组件,表示逻辑实现为Web组件,可以在Web组件和EJB组件之间增加防火墙。这样暴露给客户端的只是表示逻辑的Web组件,业务逻辑的EJB组件很好地隐藏起来了。其次,这种体系结构可以在细颗粒度上实现安全策略。EJB规范要求EJB容器实现可以在任何的EJB组件甚至组件中具体方法上实施安全策略,系统中的用户和用户组可以被授权和禁止任何EJB组件或方法的操作权。例如,在数字图书馆应用中,可以指定检索用户组只能使用检索组件,不能使用索引组件。后台加工用户在使用索引组件时,可以在方法级别上指定可以使用索引组件哪些方法。第三,EJB服务器提供的声明方式授权和编程方式授权的方式,使得这种体系结构的安全性得到进一步加强。

2.2.2 采用组件的思想

将核心业务逻辑编写成组件,这些组件包括检索组件、索引组件、保存检索策略组件、保存检索结果组件等。

采用组件实现核心业务逻辑具有以下优点。

重用性。EJB组件是基于标准分布式的对象技术、CORBA和RMI的服务器端的Java组件。我们将检索逻辑分离出来,实现为一个检索组件(EJB组件)。用户和管理员检索时,使用不同的检索界面,提交不同的检索请求到同一个检索组件。这样,可以重用检索逻辑模块。当有更复杂的检索需求时,这个检索组件可以和其他组件构建更复杂的组件。所以,这里的可重用性包括两个层面的含义:横向可重用性和纵向可重用性。横向可重用性是指同一个组件可以提供给不同类型的应用使用,而纵向可重用性是指一个组件可以和其他组件集成为更复杂组件。

灵活性。这种灵活性同样也体现在多种方法。首先,客户端的灵活性。通常一种应用会有多种类型、需要访问相同信息客户端。例如,检索应用会提供外部检索用户的基于Web组件的HTML前端,以及供内部人员使用的基于Application的更完整的应用前端。通常这个问题是通过为同一应用编写两个共享相同数据源的版本来解决的。但是,此方式开

销很大。EJB组件技术可以将共享数据和业务逻辑集成到一套EJB组件中,以供不同类型的客户端(Servlet/Jsp/HTML和Application)访问。其次,移植的灵活性。J2EE规范是一个众多厂商参与制定的规范,所有编写的EJB组件以及EJB组件所运行的服务器都必须符合这个规范。这样,标准的EJB组件可以在标准的EJB容器中运行。可以将EJB组件从一个EJB容器移植到另一个EJB容器,而不需要重新修改程序或者编译程序。第三,应用部署的灵活性。EJB组件在具体应用时,可以通过XML文件类型的部署描述文件来进行部署。这样,针对不同的安全需要,针对不同的环境,修改部署描述文件,从而使得应用更加灵活。

2.2.3 全文检索功能

实现中文的全文检索功能,主要就是解决好对中文的分析和索引问题。对于中文的索引分词主要有两种方式:一种是基于单汉字的自动切分方式,一种是基于词表的切分方式。自动分词方式,实现简单,索引效率低;基于词表的切分方式,实现维护难,效率较高。APACHE基金会Jakarta下的开源项目Lucene很好地支持了自动切分词以及维护索引等功能。Lucene现在的版本并不支持中文,但是,国内有人已经在Lucene的基础上进行了有益探索(http://www.chedong.com/),并且取得了一定成果。本系统的全文检索功能,利用了上述成果,采用了分词方式为单汉字的自动切分方式。

考虑可扩展性和规范性,数据文档内容依照都柏林核心元数据的类进行索引。根据都柏林核心元数据不同,其索引的方式也不相同。例如:对著者、出版日期等进行检索时,要完全匹配,所以需要对其内容进行索引,但是不进行分词。对标题、文档内容、简述、出版者进行全文检索时,既要索引也要分词。Lucene中对是否索引和分词提供了多样选择,满足上述索引的需求。中文分词,是通过JavaCC生成实现的,对中文部分按一个字符一个TOKEN进行索引。

Lucene中查询包括:and、or、+、-、&&、‖等,而且还有"短语查询"和针对西文的前缀/模糊查询。而且,是以类似于数据库表的行提供查询结果,使用更简单和实现更容易。

正是Lucene优良的特性,为全文检索功能的实现提供了便捷方式。

2.2.4 检索结果的表达方式

项目在设计时,对检索结果的表达方式进行了有益探索:一是关键词词频排序树的方法,二是检索结果可视化的方法。

帮助用户发现信息是数字图书馆提供的核心服务之一。一般来说,用户查找信息的时候,不可能一次检索就得到准确有用的结果,往往需要多次检索。为此,本检索系统提供了一种有效的途径,帮助用户挖掘自己的需求。由于用户与图书馆系统标引和检索系统的透明性,因而对检索出的大量结果,用户无所适从,为了从其中得到更接近需要的知识点,往往需要一篇一篇查找。一种好的信息组织方式不仅能辅助用户查找到准确有用的文献,而且可以更有效地揭示检索结果中所有文献的总体知识结构。因此,我们采用了关键词词频排序树的方法,来解决这个难题。所谓关键词词频排序树,是一种通过文献中关键词的出现频率,揭示大量文献中所含知识整体结构的一种方法。

检索结果表达可视化。过去我们所见到的许多数字图书馆的检索系统所能提供的检索结果往往只能达到一篇文献的层次,用户只能根据文献的标题、著者、关键词等外在信息了解该篇文献的主题。一般情况下,用户有时并不是对整篇文献感兴趣,而是对自己最关心的某些段或某些页感兴趣。为了满足用户的这种需求,系统对关键词或者用户输入全文检索词在每一页的分布频率上进行了统计,然后用颜色的深浅表示词频在这一页中出现频率的高低。这样以一种直观的方式,引导用户选择自己可能最感兴趣的部分,而不是整篇文献。

在整个项目中,还提供了保存检索结果、保存检索策略以及组合保存检索策略进行联合查询等服务。这些功能也是围绕用户检索功能而展开,完善了系统的功能。

3 分布环境框架

从软件工程的角度分析,数字图书馆实质是一个分布式管理信息系统。在这个系统中,异种数据资源和不同的应用系统分布在不同的网络节点。怎样在这种分布式环境下,对数据资源和应用系统进行整合,是研究数字图书馆的一个巨大挑战。对于数据资源的分布式,现在比较流行的元数据收集协议(Protocol for Metadata Harvesting)解决分布在不同地点的数据资源问题。应用系统之间的分布式,更主要是依靠技术解决不同系统之间的协同合作问题。

数字资源的分布式:按照PMH协议,数字图书馆分成两个部分——数据提供者和服务提供者。服务提供者同步集成不同数据提供者的数字资源的元数据描述。服务提供者直接面对用户,用户检索服务提供者的元数据,当用户需要具体数字资源时,再从数据提供者得到具体的数字资源。

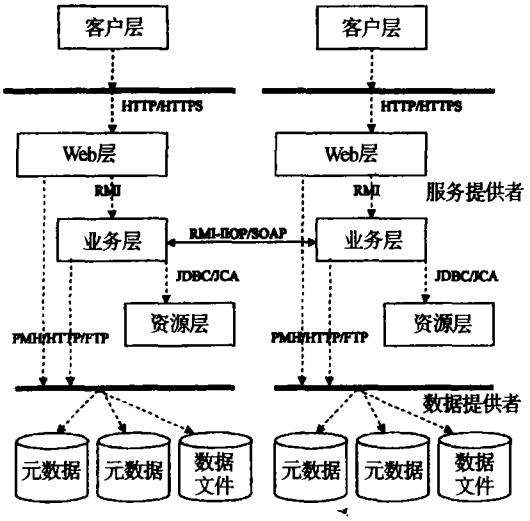

图1 分布式体系结构图

应用系统的分布式：基于J2EE的数字图书馆应用系统，从多层体系结构来看，就是一个典型的分布式环境的应用。系统的数据层、业务层、Web层以及客户层都可分布在不同的网络地点，通过网络访问(HTTP/RMI/IIOP/JDBC)将四层集成在一起。同时针对具体的业务层，又可以通过集群将这一层的应用分布在不同的运行网络地点。随着Web Service技术的发展，J2EE规范中SOAP等一系列的技术支持，分布式应用技术解决方式更加丰富。

利用J2EE平台提供的分布式应用的优势，结合PMH协议，一个数字图书馆分布式体系结构如图1所示。其中服务提供者主要利用J2EE平台提供的分布式技术，在整合不同资源时，利用PMH协议访问不同的数据提供者。

4 结语

目前，数字图书馆的研究进展很快，理论研究成果和实施成果较多。而实施项目工程的理论与方法相结合的研究较少。实施数字图书馆工程，应该用什么规范、采取什么方法，本模型从设计到开发，作了有益探索和尝试。实践证明，构建数字图书馆，从软件工程的角度分析，实质是构建数字化图书馆的分布式管理信息系统。J2EE的分布式企业级应用的规范、完整的认证与授权机制、运程方法调用和对遗留系统的处理等优良特性，为数字图书馆的设计和开发提供了十分完善的规范。借用此规范，规划我国数字图书馆建设具有现实意义和使用价值。

参考文献

1 Shengru Tu, Gongqin Li, Paul Augustin. Strategies for Integration of a Non-00 EIS and the J2EE Framework. Proceedings of the 26th International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopment. 2002(8)
2 Changtao Qu, Thomas Engel, Christoph Meinel. Implementation of an Enterprise-Level Groupware System Based on J2EE Platform and WebDAV protocol. Proceedings of the 4th International Enterprise Distributed Object Computing Conference. 2002(9)
3 Paola Inverardi, Fabio Mancinelli, Henry Muccini, Patrizio Pelliccione. An Experience in Architectural Extensions: Active Objects in J2EE. Revised Papers from the International Workshop on Scientific Engineering for Distributed Java Applications. 2002(11)
4 Janet L. Balas. Online treasures: and what of special collections in the digital library? Computers in Libraries. 2002(4). Volume:22, Issue:4
5 Nabil R. Adam, Vijayalakshmi Atluri, Igg Adiwijaya, Sujata Banerjee, Richard Holowczak. A Dynamic Manifestation Approach for Providing Universal Access to Digital Library Objects. IEEE Transactions on Knowledge and Data Engineering 2001 (7). Volume: 13, Issue: 4
6 Andreas Paepcke, Michelle Q. Wang Baldonado, Chen-Chuan K. Chang, Steve Cousins, Hector Garcia-Molina. Using Distributed Objects to Build the Stanford Digital Library Infobus. Computer 1999(4). Volume:32, Issue:2
7 José H. Canós, Javier Jaén, Juan C. Lorente, Jennifer Pérez. Building Safety Systems with Dynamic Disseminations of Multimedia Digital Objects. D-Lib Magazine 2003 (1). Volume 9, Number 1
8 Michael P. D'Alessandro, Richard S. Bakalar, Donna M. D'Alessandro, Denis E. Ashley, Mary J. C. Hendrix. The virtual naval hospital: the digital library as knowledge management tool for nomadic patrons. Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries 2001(1)
9 Bill N. Schilit, Morgan N. Price, Gene Golovchinsky. Digital library information appliances. Proceedings of the third ACM conference on Digital libraries 1998(5)
10 Lourdes Fernández, J. Alfredo Sánchez, Alberto Garcia. MiBiblio: Personal spaces in a digital library universe. Proceedings of the fifth ACM conference on Digital libraries 2000(6)
11 Constantinos Phanouriou, Neill A. Kipp, Ohm Sornil, Paul Mather, Edward A. Fox. A digital library for authors: recent progress of the networked digital library of theses and dissertations. Proceedings of the fourth ACM conference on Digital libraries 1999(8)

董 慧 武汉大学信息管理学院教授、博士生导师、软件工程国家重点实验室客座研究员。通讯地址：武汉大学信息管理学院。邮编430072。

雷 瑛 陈 琼 武汉大学信息管理学院硕士生。通讯地址同上。

杨 宁 武汉大学信息管理学院博士生。通讯地址同上。

(来稿时间：2003-10-15)