基于网格技术的数字图书馆互操作框架

周 伟

【摘 要】网格技术是近年来国际上兴起的一种重要信息技术，它的目标是实现网络虚拟环境上的高性能资源共享和协同工作，消除信息孤岛和资源孤岛。随着数字图书馆的迅速发展，互操作成为亟待解决的重要问题。目前互操作的一些技术方法在实现大规模的数字图书馆互操作方面都存在一定的局限性。笔者提出一种建立在网格技术基础上的数字图书馆互操作框架，为解决大规模数字图书馆的互操作提供了一种新的途径。

【关键词】网格 数字图书馆 互操作 OAI—PMH

Abstract: Nowadays, grid becomes an importantin formationt echnology in the world, which aims to realize high performance resource sharing or cooperation and eliminate information and resource island. With the rapid devolopment of digital libraries, the interoperability has become an important problem to be resolved. There are some limitations in solving the interoperability problem of the digital library. The author suggests a new digital library interoperability framework based on the grid technology , which makes a new way to solve this problems.

Key words: Grid digital library interoperability OAI—PMH

1 前言

Internet 已成为一个全球信息网络系统，潜藏其上的信息量达到了空前的规模，而且还在不断增长。这些信息资源是由分布在异地、异构的信息仓储构成，小到个人的信息收藏，大到一个单位的企业数据库、专业的联机检索系统、大学图书馆、搜索网站等，都是信息仓储的实例。每个仓储都是一个相对独立的信息空间，具有各自的信息组织方式、处理方式，以不同的查询方式提供利用，具有不同的权限保护和收费策略。用户信息需求的满足往往需要查询多个仓储才能完成。在网络信息环境下，了解每一个仓储的特性并掌握其检索技术的工作转移到了每一个用户的身上，这无疑是一个沉重的负担，如何屏蔽分布的各仓储间的差别，提供一致的检索界面和检索技术，由系统自动执行跨仓储的检索，在仓储间不同的信息格式、检索方式等方面进行转换，这就是数字图书馆的互操作性要解决的问题。这一问题又称为仓储的联邦。

信息社会对数字图书馆的信息检索需求已经从通过分散的网络化检索服务界面获得数字化资源转化为要求更加方便、快捷的单一入口、一次检索、统一提供的集中式信息服务方式。也就是说，读者无需为了比较全面的查找某一专题的资料而浏览多个数字图书馆网站，数字图书馆群将作为一个逻辑上统一的信息资源整体，为用户提供一站式信息服务。这种方式要求各个数字图书馆在保持内部自治的基础上，通过一定的组织方式关联起来，实现高度的资源共享和资源服务的统一管理。

2 数字图书馆互操作的实现技术方法

数字图书馆的互操作是指建立一套资源共享的有效机制，包括：数据交换与通信格式、信息交互协议、资源管理与组织方式、组织原则、用户管理与认证、访问控制等等方面，以便将独立的数字图书馆有机组织起来，协同工作，以统一界面对外提供资源服务。不同的数字图书馆由于地域分布、组织归属、经费来源等等原因，往往分别独立建设，因而采用不同的软、硬件平台和数据库系统来进行馆藏资源的数字化，无法遵循统一的标准和规范。由于体系结构和实现技术的差异，数字图书馆系统之间一般不能直接进行数据通信和资源共享，无法建立规整的、一体化的、自底向上的统一信息服务平台，只能通过有针对性的开发标准通信协议和中间层软件，来标准化数字图书馆的对外通信接口以及屏蔽其内部结构。最终将各自为政的资源服务体系“粘和”起来，成为有机的整体。

数字图书馆系统建设面临的主要互操作问题有：

(1) 屏蔽分布的各数字图书馆之间的差别，为用户提供一个一致的检索界面，在该统一界面上进行的跨仓储检索对于用户来说是透明的。

(2) 为数字图书馆系统提供一种灵活的集成机制。这种集成方法必须允许各相对独立的数字图书馆能够自由增加新的服务，或对以前提供的服务进行修改。

(3) 数字图书馆系统服务协议的制定，包括元数据协议、数字对象存储协议、信息搜索协议、付费协议等等。

(4) 开发数字图书馆系统高层协议中间件，实现分布的子系统间各项服务的互操作。

数字图书馆的互操作性不仅涉及技术层面，还包括语义互操作性、行业领域互操作性等等。目前数字图书馆主要有 3 种互操作性模型：联邦、采集、收集。它们采用相同的理念，通过一定的技术手段屏蔽各异构数据库（源）或分布的成员数字图书馆间的差别，为用户提供一致的检索界面，由系统自动执行跨平台的检索，对子系统不同的信息格式进行转换，并向用户提供最优显示，基于这 3 种互操作模型，人们提出了不同的互操作解决方案，其中具有代表性的有：分布式计算、中间件控制和元数据收集技术。

2.1 分布式搜索 (distributed search)

该方法实时将用户提交的查询请求转换成每一个数字图书馆可接受的形式，分别送往多个数字图书馆站点执行，收集每个数字图书馆返回的结果，综合整理后交给用户。这种方法要求数据源端维护各自的搜索服务，由分布式搜索服务提供统一的查询界面。NCSTRL 和 OldDominion 大学提出的 LFDL 结构是采用该方法的两个典型例子。

但是由于分布式搜索方法依赖于实时地执行查询操作，因此，一般来说对于数字图书馆节点不超过 20 个的情况下，该技术比较适用，但在 Internet 环境中，数字图书馆节点的数量比较大（大于 100），利用该方法解决数字图书馆互操作问题就变得十分困难。

2.2 中间件技术 (middleware)

目前的数字图书馆互操作，比较多的采用中间件技术来实现，如：微软的分布式组件对象模型 DCOM (Distributed Component Object Model)、对象管理组织的公用对象请求代理结构 CORBA (Common Object Request Broker Architecture) 等。中间件是由软件实现的功能层，位于应用程序与包含异构操作系统、硬件平台以及通信协议的底层结构之间，用来隐藏底层平台的特殊性，在复杂的后台服务与相对较简单的前端应用之间起着承上启下的作用。中间件技术在实现分布式异构系统互操作方面起到了重要的作用。但是，这些系统有个共同的缺陷，就是它们无法扩展到 Internet 上，它们要求服务客户端与系统提供的服务本身之间必须紧密结合，即要求一个同类基本结构。另外，这样的系统往往十分脆弱，如果一端的执行机制发生变化，另一端就会崩溃。

采用中间件技术实现的数字图书馆互操作系统的局限性主要表现为：

(1) 当数字图书馆内部服务发生变化时，中间件需要修改或重写，否则将导致客户端服务调用的失败。

(2) 由于中间件缺乏严格一致的概念模型和形式语义，因而软件重用性比较差。

(3) 没有对通信安全进行控制，无法保证广域网环境下信息交互的安全性。

(4) 中间件主要用于实现组织内部的资源共享，采用不同中间件技术实现互操作的数字图书馆组织之间的互连互通仍然存在问题。

2.3 元数据采集 (harvesting)

建立大规模分布式搜索服务的困难，导致了基于元数据采集方法的出现。该方法后来被 OAI (openarchivesinitiative) 利用，建立了典型的元数据采集框架 OAI—PMH，为 DLs 的互操作问题提出了一种简单、可行的解决方案。

该解决方案的基本思想是：数据提供者将其元数据使用公共元数据格式表达，服务提供者利用开放协议从每个 DL 中采集元数据，经过处理、合并后集中保存在一个元数据仓储中，利用元数据仓储为用户提供增值服务。该方案有效地解决了各资源库在元数据格式上可能存在的异构性问题，实现了跨资源库检索。OAI 协议的模型如图 1 所示。

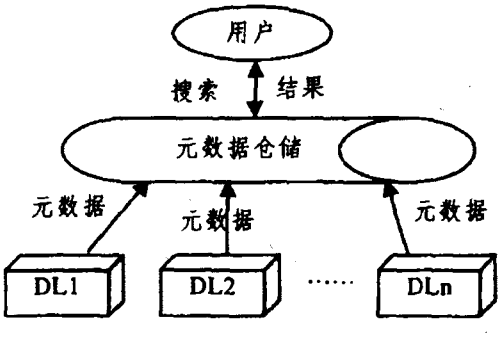

图 1 OAI 协议模型

由于 harvesting 方法采用集中处理方式，所以能够保证较好的查询响应时间，而且它不要求严格遵守一组完整的技术协定，只要求做少量支持基本共享服务的工作，对联盟成员的要求很少，因此许多组织可能会加入这种松散的 DLs 联邦。但是由于 OAI 比较新，

有些元数据采集的重要问题尚未涉及，有一定的局限性：

（1）没有规定如何选择数据源，随着数据提供者数量的急剧增加，必须提供选择数据资源的机制。（2）没有强调如何实现服务提供者，对于增值服务，OAI并没有说明如何具体地实现。（3）OAI互操作框架也存在元数据的同步更新问题。目前主要利用基于日期戳的pull模型来解决数据同步问题。

当然以上几种技术实现方法在单一应用时只能解决互操作性上的某些问题，而在实际系统中，则需要将它们有机地结合起来使用，才能更为有效地对不同系统中的数据源的格式进行转换，提供一个统一的数据介质层面，从而使用户对分布式数字环境下的异构信息资源进行透明化的检索。

3 基于网格技术的数字图书馆互操作框架

网格是高性能计算机、数据源、Internet三种技术的有机组合，它具有高性能、一体化、知识生产、资源共享、异地协同工作、支持开放标准、功能动态变化等优点，为数字图书馆建设提供了有利的条件。网格计算是指在动态的、异构的、广域的虚拟组织中进行的协同资源共享和问题求解，是把整个国际互联网集成为一台巨大的超级计算机，实现全球范围的计算资源、存储资源、数据资源、信息资源、知识资源、专家资源、设备资源等的全面共享。它的根本特征是资源共享、消除资源孤岛，而其规模不是主要因素。

3.1 网格技术在数字图书馆建设领域中的应用

近年出现的网格技术，使我们对资源的组织和利用方式有了一个全新的认识。网格的核心思想是：将网络上的各种设备，包括计算机、大型仪器等统一包装为网格设备，对外提供标准访问接口，在全局管理机制和安全控制机制下，形成一个大规模的资源基础设施。这种设施具备像电力网一样的服务方式：插上插头就能获得源源不断的计算能力、信息资源和应用服务。网格的目的是共享，这种共享已经不再是简单的资源互连和单一使用，而是通过资源互连、资源组合和资源协作来解决用户需要解决的问题，产生具有附加值的新的服务、数据、知识等资源，满足用户新的需求。总之，网格具有分布、异构、动态和自主的特性，它不仅强调资源的互连互通，更强调资源的互操作能力，最终目的是支持虚拟组织的协同工作。网格技术在数字图书馆建设领域中的应用有：

（1）网格为数字图书馆构造统一的平台，为数字图书馆提供各种一体化信息服务的信息基础设施。网格利用现有的网络基础设施为用户提供一体化的智能信息平台，创建一种基于因特网的新一代信息平台和软件基础设施。

（2）网格有利于实现数字图书馆的资源共享，网格支持对异构数据资源的访问，为用户提供统一的访问接口，选择适当的访问协议来实现用户提出的数据访问请求。

（3）网格有利于数字图书馆的海量数据处理，用户只需通过客户端发出要求计算的指令，网格就把这些任务调配给各个计算机执行，然后将各个计算机计算出来的结果汇总反馈给用户，省去了多次访问不同数据库的麻烦，并能直接调用网格中的算法和程序等资源，避免许多重复性的工作。

（4）网格有利于数字图书馆的信息集成，在分布式的异构环境中，网格技术能够精确定位所需的数据集，并为后续处理提供支持。人们利用这些资源就像用电源一样，不必计较这些资源的来源和负载情况。网格计算可以合理而有效地将远程资源高效地组织起来，形成网络虚拟计算机，形成超强的能力。

（5）网格有利于数字图书馆进行知识管理，网格能根据用户的要求自动地生产知识。它把从数据源得到的各种原始数据，运行特定的程序加工成信息知识。

3.2 一种基于网格技术的数字图书馆互操作框架

OAI—PMH框架通过元数据的互操作实现数字图书馆的互操作，克服了分布式搜索无法解决的规模问题，而网格计算技术与传统的分布式计算不同之处在于，关注多机构之间大规模的资源共享和合作使用，提供了资源共享的基本方法，在解决异构平台兼容及集成已有系统方面有着独特的优势。为此，作者将网格技术与元数据采集方法相结合，提出一种如图2所示的数字图书馆互操作的新框架，在原有框架的基础上引入网格的概念，通过OGSA网格构架和OAI—PMH来解决数字图书馆资源发现、收集、跨仓储检索等问题，克服传统互操作方案的局限性，实现整个社会范围内的联邦数字图书馆。

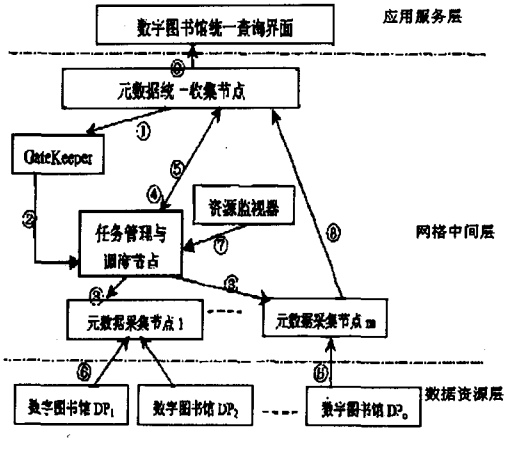

图2 数字图书馆互操作框架模型

数字图书馆网格架构由三层组成，位于底层的是数据资源层，由广域分布的多个数字图书馆组成，作为互操作数字图书馆的数据提供者，提供 OAI－PMH 协议规定格式的元数据，形成遵循 OAI－PMH 协议的元数据仓储。然后是网格层，它利用开放的网格技术和 OAI－PMH 协议屏蔽了底层数字图书馆的分布性、异构性，通过进行元数据的发现、收集和全局索引工作，向应用层提供统一的服务接口。位于最上层的是数字图书馆的应用服务层，它在集成数字图书馆元数据的基础上，为用户提供统一的服务接口，提供检索、参考咨询等服务。

在数字图书馆网格体系结构中主要通过网格来收集元数据，因此在其体系结构中引入三类网格节点：采集调度服务节点、元数据采集节点和元数据收集结点，各节点的组成和功能描述如下：

(1) 采集调度服务节点。它的主要功能是存储一个配置文件，该文件包含所有可以被收集的数据提供者列表，将收集元数据的任务分配给 harvesting 节点，并对最近的收集工作进行跟踪，一旦历史数据收集完毕，在预定新的收集之前，确定合适的参数。

(2) 元数据采集节点。这类节点接入网格后，通过一个 Web 方法与计划节点联系接受收集任务，一旦被分配一个数据提供者，该节点就在其上执行收集任务，任务完成后，再次与计划节点联系得到新的收集任务。

(3) 元数据收集节点。该节点收集所有采集节点采集到的元数据，并按内容主题进行整理，并把它们分送到不同的搜索集群节点，这类节点的引入有两种功能，一是引入某种形式的负载均衡，二是简化灾难处理过程。

3.3 数字图书馆网格工作原理

数字图书馆网格通过开放的网格平台，屏蔽了数字图书馆的分布性和异构性，完成地理分布的数字图书馆间的元数据的发现、收集和全局索引，为用户提供一致的信息服务。网格的具体工作步骤如下：

(1) 元数据收集节点启动元数据收集任务，网格启动一个负责任务分配工作的进程，根据任务请求进行任务创建，由任务管理与调度节点负责任务与资源的分配。(2) 任务管理与调度节点选择相应节点进行元数据的采集工作，并向元数据统一收集节点提供任务的状态信息。(3) 元数据统一收集节点向任务管理与调度节点提交任务取消等请求。(4) 广域分布的各数字图书馆的元数据提供这项元数据收集节点提供自己的元数据信息。(5) 元数据采集节点收集元数据后，将其发送给元数据统一收集节点，后者将所有采集节点采集到的元数据进行重新组织与整理，并分送到相应的元数据分类仓储中存储起来。(6) 元数据统一收集工作完毕，存储在元数据仓储中，便能够通过统一查询界面为用户提供一致的搜索界面，将用户的查询请求发送到相应的元数据仓储中，收集查询结果，并提交给最终用户。

根据该原理我们用 GT3.2 搭建了实验原型的网格环境，采用 2 个节点，从 2 个数据提供者执行延迟的元数据采集任务，同时利用网格将收集到的元数据记录解析后并存储在 SQL－SERVER 数据库中，采用 1 个索引节点对元数据库进行索引，搜索引擎在索引上执行搜索。

4 结论

当前全球正在兴起的有关网格的研究，使我们感受到一种信息社会的新的基础设施正在出现。根据数字图书馆互操作的要求和网格在资源利用与共享方面的优势，本文提出基于网格技术的数字图书馆互操作架构，通过将网格与 OAI－PMH 协议结合起来，利用网格进行元数据的高性能搜索、收集和索引工作。这将大大降低数字图书馆互操作的实现成本，并能提高服务的可靠性。

但在很多方面还有待于改进和完善，尤其是在系统具体实施方面涉及到的具体设计和技术问题，还需要通过进一步深入的研究，主要有元数据问题、数字图书馆系统高层服务协议中间件的实现以及异构数据源集成的具体实施。

参考文献

1 ShiR, MalyK, ZubairM. Dynamic interoperation of non－cooperating digital libraries [A]. Proceedings of Digital Library IT Opportunities and Challenges in the New Millennium C]. Beijing : Beijing Library Press, 2002

2 ShiR, MalyK, ZubairM. Interoperable federated digital librar yusing XML and LDAP [J]. GI

3 郑志蕴等．基于网格技术的数字图书馆互操作关键技术．北京理工大学学报，2005 (12)

4 郑志蕴等．基于网格的数字图书馆互操作技术研究．计算机科学，2005 (8)

5 都志辉，陈渝，刘鹏．网格计算．北京：清华大学出版社，2002

周 伟 盐城工学院。