基于网格的数字图书馆互操作技术研究*)

郑志蕴[1,2] 徐 玮[1] 牛振东[1,3] 宋瀚涛[1]

(北京理工大学计算机科学与工程系 北京 100081)[1] (郑州大学信息工程学院计算机系 郑州 450052)[2]

(中国数字图书馆有限责任公司 北京 100081)[3]

摘 要 以支持大规模资源共享和协作为核心的网格技术在解决数字图书馆(DLs)互操作问题,尤其是在异构平台兼容、集成已有系统方面有着独特的优势。本文对目前数字图书馆互操作主要解决方案进行分析和对比,包括:分布式搜索(Distributed Search)、元数据收集(Metadata Harvesting)、中间件(Middleware)技术,指出各自的优势和不足,提出一种新的数字图书馆互操作框架——数字图书馆网格 DL Grid,探讨利用先进的网格技术解决各数字图书馆的互操作性问题。

关键词 数字图书馆,互操作,网格技术,OAI

Research on the Technologies of Digital Library Interoperability Based on Grid

ZHENG Zhi-Yun[1,2] XU Wei[1] NIU Zhen-Dong[1,3] SONG Han-Tao[1]

(Department of Computer Science & Engineering, Beijing Institute of Technology, Beijing 100081)[1]

(Department of Computer ,College of Information & Engineering, Zhengzhou University, Zhengzhou 450052)[2]

(China Digital Library Carp. Ltd,Beijing 100081)[3]

Abstract Grid technology, which supports a large scale resources share and collaboration, has particular advantage in solving digital library interoperability, especially in integrating existing heterogeneous systems and platform. First this paper gives an analysis and comparison of existing conventional approaches of digital library interoperability, including Distributed Search, Metadata Harvesting , Middleware Technology, and advantages and disadvantages of each method is pointed out. in the end, a new framework of DLs interoperability, Digital Library Grid, is presented, which try to solves interoperability problem of digital library using advanced grid technology.

Keywords Digital library , Interoperability, Grid technology, OAI

1 引言

在数字图书馆领域,互操作通常用来具体描述同一数字图书馆的各个组件或不同数字图书馆之间交换、共享文档、查询和服务的能力[1]。

自从美国颁布实施"数字图书馆倡议(Digital Library Initiative)"之后,各国都掀起了一股研究数字图书馆的热潮。作为社会信息基础设施,数字图书馆将给人们的学习生活带来明显的变化。但同时,更多的组织加入到数字图书馆的建设中来,也造成了两方面的结果:一方面,不同组织建设的数字图书馆将侧重于不同的各具特色的信息内容;另一方面,由于不同组织进行数字图书馆建设的目的、方式、运行手段不同,从而在技术实现上采用的平台、协议、体系结构也各不相同。因此,现有的数字图书馆(DLs)就像 Internet 世界上一个个孤立的小岛,大量的信息还是被"锁"在各个小岛的中央数据库里,用户为了获得所需资料,往往需要访问几个 DL,同一查询请求不得不重复提交给每个 DL,且结果往往冗余度大,查全率低。用户期望今后的数字图书馆能够提供一个异构、分布信息源无缝集成的视图,实现 DLs 资源的透明访问。要达到这一目标,实现数字图书馆资源最大化的利用,需要解决数字图书馆之间的互操作问题。本文首先对已有 DLs 互操作方案进行分析和对比,指出各自的优势和不足,然后提出利用先进的网格技术解决数字图书馆的互操作性问题,并给出一种新的数字图书馆互操作框架—数字图书馆网格 DL Grid。

2 数字图书馆互操作解决方案

当前,数字图书馆主要有三种互操作模型[2]:联邦(federated)、采集(harvesting)、收集(gathering)。它们采用同样的思想,即将 DLs 的异构性用某种方式予以屏蔽或转换,使它们看起来像同构系统或标准系统,从而支持互操作。基于这三种数字图书馆互操作模型,人们提出了不同的互操作解决方案,其中具有代表性的有:分布式搜索(Distributed Search)、元数据收集(Metadata Harvesting)、中间件(Middleware)技术等。

2.1 分布式搜索(Distributed Search)

分布式搜索为跨越不同的 DLs 馆藏发现有用信息提供了一条途径。其基本思想是,实时将用户提交的查询请求,转换成每一个 DL 可接受的形式,分别送往多个 DLs 站点执行,收集每个 DL 返回的结果,综合整理后交给用户。这种方法要求在数据源端维护各自的搜索服务,由分布式搜索服务提

*)基金项目:霍英东教育基金"数字图书馆个性化服务研究"(91101)。郑志蕴 博士研究生,主要从事:网格计算,信息处理研究;徐 玮 博士研究生,主要从事网格计算研究;牛振东 教授,主要从事数字图书馆研究;宋瀚涛 教授,博士生导师,主要从事:信息处理、分布式数据库和无线网。

供使用远程 DLs 搜索服务的统一查询界面。

按是否遵循标准，分布式搜索分又为两大类：

2.1.1 *基于标准(standard)的方法* 该方法属于联邦的方法，所有参与互操作的数字图书馆，构成一个联邦。在联邦内部制定一系列的协议和规范，要求所有成员的系统都遵守协议，并按照公共的规范建造服务。在实践中，所有的组织使用相同的平台和软件，并统一调度。网上计算机科学技术报告图书馆 NCSTRL(Networked Computer Science Technical Reference Library)和 NCSTRL+是采用该方法的两个典型例子[3]。

NCSTRL 是一个拥有 100 多个机构加盟的联邦数字图书馆，它利用 Dienst 作为 DLs 的协议和体系结构，并借助于分布式搜索技术在联盟 DLs 之间实现资源共享，如图 1 所示。

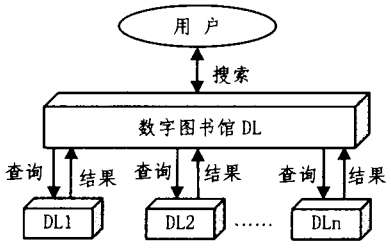

图 1 NCSTRL 方法示意图

因为严格按照达成一致的标准来建立 DLs 系统及馆藏资源，因此基于标准的方法可以提供较好的、全面的互操作性。不足之处：对于成员的要求比较高，成员之间关系紧密，所以很难形成大规模的联盟团体。因此，只有那些需求强烈的 DLs 采用该模式。

2.1.2 *基于数据驱动(data-driven)的方法* 该方法属于收集方法类，它既不要求对现有 DLs 的结构做任何修改，也不要求联盟成员的 DLs 遵从某种互操作协议，而是通过收集(Gathering) DLs 可公开访问信息的途径获得最基本的互操作。这种方法通常提供统一的用户界面，用户输入查询请求，系统执行分布式搜索，并将合并后的查询结果返回给用户。互联网上的元搜索引擎，像流行的商用元搜索引擎 Metacrawler[4]、search. com 就属于这个层次。

一个典型的代表是 Old Dominion 大学在 InterOp 项目中提出的 LFDL(Lightweight Federated Digital Libraries) [5]结构。在 LFDL 中，统一的搜索界面被定义成基本的交互中间层，只要求使用数字图书馆描述语言 DLDL，描述各自的特征、能力、交互信息，并将这些描述信息登记到注册服务器中。当用户通过联邦数字图书馆(FDL)查询时，FDL 根据注册服务器中保存的信息，选择出最合适的 DLs 执行用户的查询，并收集这些 DLs 返回的结果，合并整理后返回给用户。

基于数据驱动方法是在传统的搜索服务之上提供一个抽象层，使其利用收集方法建立联邦数字图书馆，对成员没有任何要求，在 Web 增长和变化时具有较好的适应性、可伸缩性和便携性。不足之处：所提供服务的质量比参与合作情况下的差，难以满足需要密切合作成员的要求。

2.1.3 *分布式搜索技术的局限性* 由于分布式搜索方法依赖于实时地执行查询、处理查询结果，因此，有一个规模问题。对于数字图书馆节点比较少(一般来说是不超过 20 个)的情况下，该技术是比较适用的，但在 Internet 环境中，数字图书馆节点的数量都比较大(大于 100)，在这种情况下，利用分布式搜索技术来解决数字图书馆互操作问题就变得十分困难[6]。

2.2 Harvesting 方法

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概念(Harvesting)的出现。Harvesting 方法对于联盟成员不要求遵守许多复杂的协议，只需少量的工作就可以实现与其它成员的互操作，因此是加入联盟的一种低门槛(Low-barrier)的方法。其基本思想是：从每个 DL 中采集并提取元数据，经过处理、合并后集中保存在一个元数据仓储中，用户对保存在元数据仓储中的元数据进行查询。如图 2 所示。

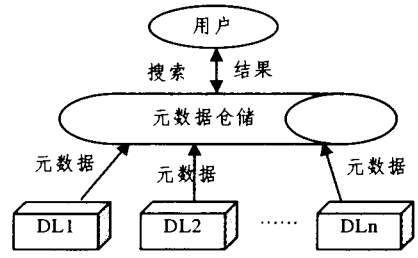

图 2 Harvesting 方法示意图

对 Harvesting 方法的研究源于 20 世纪 90 年代 Colorado 大学开发的 Harvest。该方法后来被 OAI(Open Archives Initiative)利用，建立了典型的元数据采集框架 OAI-PMH[7]，为 DLs 的互操作问题提出了一种简单、可行的解决方案。该框架区分两种不同类型的参与者：数据提供者 DP(data providers)和服务提供者 SP(service providers)。前者要求按照标准的元数据格式(Dublin core)建立馆藏元数据，后者利用 OAI-PMH 协议从数据提供者处获取元数据以实现增值服务(如搜索、浏览等)。这种结构的代表是 Arc，Arc 是第一个采用 OAI 互操作框架实现的联邦搜索服务，它能够从遵守 OAI 协议标准的 DLs 馆藏中提取元数据，经过处理后集中保存在一个关系型数据库中。

目前，基于 Harvesting 的联邦搜索依然是 DLs 界研究与开发的热点，一些著名的 DLs 项目，如 NDLTD 和 NSDL，也都采用 Harvesting 方法作为互操作的解决方案。

由于 Harvesting 方法采用集中处理方式，所以能够保证有较好的查询响应时间。另外，Harvesting 方法不要求严格遵守一组完整的技术协定，只要求做少许支持基本共享服务(如数据访问)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。不足之处：各 DLs 馆藏元数据的变化不能够及时地得到反映。

2.3 中间件 (Middleware)技术

20 世纪 90 年代，基于网络计算平台的分布计算(Distributed Computing)技术迅猛发展。其中以面向对象技术为主要特征的分布式构件技术，即中间件技术，经过几年的蓬勃发展，进入成熟时期，为解决数字图书馆的互操作问题提供了很好的参考。数字图书馆的应用环境包含异构的硬件平台、OS、通信协议和数据库管理系统。在这种异构的环境中，需要通过中间件所提供的具有标准编程接口和协议的服务，建立独立的软件层，隐藏数字图书馆的底层信息源和服务的异构性，从而实现数字图书馆的互操作。目前，实现 DLs 互操作常用的中间件技术有：CORBA 技术、中介层结构。

2.3.1 CORBA 技术 CORBA 是 OMG(对象管理委员

会)提出的服务于系统间互操作的一种体系结构。它能让计算机应用程序在分布式网络中相互协作，较好地解决了封装对象在分布式计算环境中的资源共享，软件重用以及功能扩展等问题。使得各种各样的对象系统能够进行集成。斯坦福大学数字图书馆的 InfoBus 系统和康奈尔数字图书馆研究小组的数字图书馆存储系统 FEDORA(灵活可扩展数字对象和存储体系结构)，均采用 CORBA 技术来实现。

2.3.2 中介(Mediation)结构[8] 该结构利用一个中介层为每种数据源提供一个通用的数据模型和查询界面，使用包装层(wrapper)屏蔽各种数据源之间的异构性。中介层负责接受用户的查询，并将其转换成通用模型。包装层将中介层提供的通用模型转换成针对具体数据源的查询并执行。中介层收集来自包装层转换后的查询结果，将其归并后返回给用户。Goncalves 等人利用面向对象的数字图书馆系统 MARIAN 作为 NDLTD 的中介层中间件(mediation middleware)，提供一个公共的查询界面和集成平台。

2.3.3 中间件技术的局限性 中间件技术是一种传统的分布式计算技术。而分布式计算技术强调的是分布系统的集成能力，以两层或多层 Client/Server 为主要计算模式，关心的是简化用户端的工作，强化多层服务器的功能，典型的共享是基于静态的，并且通常关注一个组织内的资源共享。交互的基本形式不是对等使用多种资源，没有提供多组织之间的资源共享通用框架，不能实现大规模的信息共享。

3 网格技术与数字图书馆结合

3.1 网格概念的引入

网格(Grid)技术[9]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。著名的网格计算项目 Globus 的主持人之 Ian Foster 认为："网格是构筑在互联网上的一组新兴技术，它将高速互联网、高性能计算机、大型数据库、传感器、远程设备等融为一体，为科技人员和普通老百姓提供更多的资源、功能和交互性。互联网主要为人们提供电子邮件、网页浏览等通信功能，而网格功能则更多更强，能让人们透明地使用计算、存储等其它资源。"[10]

互操作性问题是数字图书馆的一个关键性问题，第 2 节论述了解决该问题的主要方法，指出了各自的不足和局限性。为了更好解决 DLs 互操作中出现的问题，需要建立新的框架体系结构，采用新方法来建设整个社会范围内的联邦数字图书馆，以支持大规模资源共享和协作为核心的网格计算技术在解决异构平台兼容、集成已有系统方面有着独特的优势。本文提出将网格技术与数字图书馆互操作技术相结合，在已有的信息基础设施之上，架设网格层，采用网格本身的信息服务协议，屏蔽各 DLs 系统的异构性，解决数字图书馆互操作问题。

如前所述，在数字图书馆互操作方案中，元数据采集 harvesting 的方法可以克服分布式搜索无法解决的一个规模问题。而网格与传统的分布式计算相比，显著之处在于，关注大规模的资源共享，革新的应用，以及在某些事例上高性能的需求，它强调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[11]。因此，本文将网格技术与数字图书馆互操作技术之一——元数据采集 harvesting 方法相结合，不但可以解决 DL 互操作中的动态性和异构性问题，还可以更好地解决资源发现、整合、安全等问题，从而克服传统 DLs 互操作方案的局限性，实现全球 DLs 信息资源共享和跨仓储无缝查找。

3.2 解决方案

OAI-PMH[7]是利用 Harvesting 概念建立的典型的元数据采集框架，该框架为 DLs 的互操作问题提出了一种简单、可行的解决方案。通过对网格计算技术和 OAI-MHP 协议的分析，提出一种增强数字图书馆互操作的新框架，即在原有 OAI-PMH 框架的基础上，引入 Grid 的概念，建立数字图书馆网格 DL Grid。为了增强收集和索引的动态性能，加快元数据的更新速度，在 DL Grid 体系结构中引入三类 Grid 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。过去 OAI-PMH 直接连接 DP(data providers)和 SP(service providers)，简单地在 http 上实现元数据的采集，现在采集(harvester)节点必须通过 Grid 收集 DP 的元数据。通过使用 Grid 计算节点，可以增加服务提供者的质量，支持高性能的 OAI 联邦搜索服务。DL Grid 主要体系结构图见图 3。

DL Grid 的每一个组成和功能描述如下：

(1)采集调度服务(Harvest Scheduler Service)

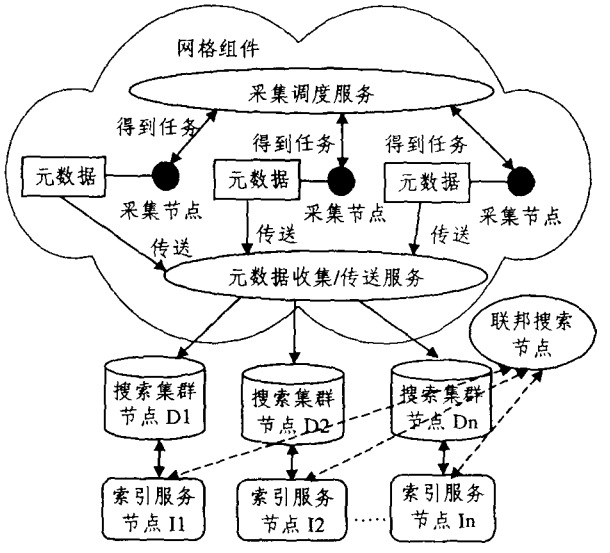

图 3 DL Grid 主要体系结构图

该服务(软件)可以运行在一个单独的节点上，或是其中的一个元数据采集节点上。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 DP 的列表；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，决定合适的参数。

(2)元数据采集节点(Metadata Harvesting note)这类节点接入 grid 后，通过一个 Web 方法与 Scheduler 节点联系，接受收集任务。一旦被分配一个 DP，该节点就在其上执行收集任务。任务完成后，再次与 Scheduler 节点联系，得到新的收集任务。

(3)索引服务节点 Ii (Index service nodes)

当新的收集完成，新增加的元数据被送均衡到各个不同 Di，在预先规定的时间(例如，每天的午夜)之后，Ii 对 Di 中的元数据或重新索引，或作增量索引，并把更新的索引集送回到相应 Di 。

(4)搜集群节点 Di(Search cluster nodes)

这类节点提供搜索服务，它既存储元数据，也存储最新的

索引。联邦搜索服务收到的请求被分配到这些节点，这些节点在它们各自元数据的基础上，利用索引执行搜索任务，并返回结果。

(5)元数据收集节点(Metadata Collection Node)

该节点收集所有 harvesting 节点收集到的元数据，并把它们分配到不同的搜索集群节点，即 D1，D2，…，Dn。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。

(6)联邦搜索节点(Federated Search node)

该节点负责为用户提供统一搜索界面，将搜索请求分配给 cluster 上的所有 Search nodes(D1…Dn)收集搜索结果，并提交给最终用户。

作为探索性研究的一部分，我们正在采用 GT3.2，建立一个实验系统，使用 2 个 grid 节点从 2 个 DP 中执行高延迟的收集和索引元数据的工作，同时也将利用 grid 将收集到的元数据送到一个搜索引擎的小集群(SP)，每一个搜索引擎将再从 harvesting 节点获得的索引上执行搜索。

总结 互操作性是数字图书馆所面临的重大问题和关键挑战，它几乎渗透到数字图书馆作为一个分布式计算系统的每个方面。至今数字图书馆界已经开发出了许多方法，取得了一些成果，但这些方法在实现 Internet 上大规模的数字图书馆互操作方面有一定的局限性。当前，全球正在兴起的有关网格的研究，使人们感受到一种信息社会的新的基础设施正在出现，这种新的 infrastructure 可能带来信息资源的获取、分布、传输和有效利用的革命性的、结构性的巨大变化。本文从资源共享的角度，提出利用先进的网格技术，在原有 OAI-PMH 框架的基础上，构建数字图书馆网格 DL Grid，利用网格建立和存储用于信息发现的元数据，实现 DLs 信息资源共享和跨仓储无缝查找，从而解决数字图书馆的互操作性问题。

参 考 文 献

1 Paepcke A, Chang C-C K, Garcia-Molina H, et al. Interoperability for Digital Libraries: Problems and Directions-Stanford University. 1998
2 NSDL. National SMETE Digital Library. http: // www. smete. org/nsdl/
3 Shi R, Maly K, Zubair M. Dynamic interoperation of non-cooperating digital libraries. In: Proc. of Digital Library IT Opportunities and Challenges in the New Millennium, Beijing: Beijing Library Press, July 2002. 350～361
4 Metacrawler. http: // www. metacrawler. com
5 Shi R, Maly K, Zubair M. Interoperable Federated Digital Library using XML and LDAP. Global Digital Library Development in the New Millennium, 2001(5): 277～286
6 Enhancing Infrastructure for OAI. http: /dlib. cs. odu. edu/ # dtic, 2004
7 The Open Archives Initiative Protocol for Metadata Harvesting. http: //www. openarchives. org/OAI/2. 0/openarchivesprotocol. htm
8 张付志. 异构分布式数字图书馆互操作技术研究:[博士论文]. 北京:北京理工大学, 2003
9 都志辉，李三立，刘鹏. 网格计算. 清华大学出版社. 北京，2002
10 Foster I, Kesselman C. The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, San Fransisco, CA, 1999. http: // mkp. com/grids, http: // www. gridforum. org/, http: // www. ccgrid. org/
11 Foster I, Kesselman C, Tuecke S. The anatomy of the grid: enabling scalable virtual organizations [J]. CCGRID2001, First IEEE/ACM International Symposium on Cluster Computing and the Grid: 6～7

(上接第 229 页)

DACBTM 系统对于基本信任关系的表达能力与它们基本相同，支持授权、委托等复杂的信任关系，同时引入了信任度的概念以表达实体间的相对的信任关系。DACBTM 系统在实现了凭证搜索算法的基础上，还增加了系统对于安全信息不足即无法查找到足够的信任凭证的情况的处理步骤。现存的信任管理系统 dRBAC[8] 也做了类似的扩充工作。与 dRBAC 系统相比，DACBTM 在算法的搜索过程中，通过将搜索分解为若干个子搜索，使得每个子搜索运行与不同的结点之上，从而降低了单个节点的负荷，同时利用路径信息来缩短搜索路径，提高搜索效率。DACBTM 系统与 dRBAC 系统相比一个重大的改进和扩充就是提供了对于安全信息不足的处理机制。DACBTM 系统利用软件实体的历史信息，采用基于经验的信任评估模型[5]，在一定程度上解决了安全信息不足的问题，DACBTM 系统同时提供相应的历史信息管理和更新机制。

DACBTM 系统提供了凭证管理的功能。DACBTM 系统分析凭证的有效性，在凭证的 Property 域条件不被满足时，系统将把该凭证标记为无效凭证。系统可以定期检查无效凭证，若凭证超过有效期时，系统通过凭证库存取接口将该凭证删除。同时，为了支持信任评估，DACBTM 还提供了经验信息的收集与管理机制。

结束语 在开放协同软件环境下，应用系统的结构发生了根本性的转变，实体的应用需求也变得复杂多变。传统的安全手段和安全凭证已经无法满足系统的应用需求。在开放环境下，系统的动态性和安全信息的不完整性，都导致了用户对于应用系统中所存在的安全问题提出了更高的要求。本文在传统信任管理系统的基础上，设计并实现了一个新型的分布式访问控制系统，该系统适用于开放协同软件环境下的安全授权。它将传统的信任管理系统中所使用的授权方法与基于经验的信任评估方法做了有机的结合，使得系统能够在安全信息不足的情况下进行授权，同时也保证了应用系统的安全性。

本文的进一步工作，是使信任搜索过程和信任评估过程结合得更加紧密，即在信任链的搜索过程中使用信任评估的方法，例如在对于信任链进行双向搜索时，当信任链中断，使用信任评估方法将信任链补充完整。同时，如何对于经验信息采用更为有效和准确的描述方式，也很值得探讨和研究。

参 考 文 献

1 Blaze M, Feigenbaum J, Lacy J. Decentralized Trust Management. In: Proc. of the IEEE Symposium on Research in Security and Privacy, Research in Security and Privacy, Oakland, CA, May 1996. IEEE Computer Society, Technical Committee on Security and Privacy, IEEE Computer Society Press.
2 Blaze M, Feigenbaum J, Ioannidis J, et al. The KeyNote trust-management system, version 2. IETF RFC 2704, Sept. 1999
3 Li N, Mitchell J C, Winsborough W H. Design of a role-based trust-management framework. In: Proc. of the 2002 IEEE Symposium on Security and Privacy. IEEE Computer Society, 2002
4 Li Ninghui, Winsborough W H, Mitchell J C. Distributed credential chain discovery in trust management(extended abstract). In: Proc. of the Eighth ACM Conference on Computer and Communication Security(CCS-8), ACM Press, Nov. 2001. 156～165
5 徐锋，吕建，郑玮，曹春. 一个软件服务协同中信任度计算模型的设计. 软件学报，2003, 14(6): 1043～1051
6 徐锋，吕建. Web 安全中的信任管理研究与进展. 软件学报，2002, 13(11): 2057～2064
7 马晓星. Internet 软件协同技术研究:[南京大学博士论文]. 2003
8 Freudenthal E, Pesin T, Port L, et al. dRBAC: Distributed role-based access control for dynamic coalition environments. In: Proc. of the 22nd Intl. Conf. on Distributed Computing Systems (ICDCS02), 2002