●焦玉英 雷 雪

异构分布式信息检索系统整合研究*

摘 要 实现异构处理器、异构数据源之间的共建共享,必须对分布式检索系统的各个子系统进行整合。现有的技术对于实现具体的、个别的分布式检索系统的互操作虽然没有问题,但要建立普遍意义的互操作规范,尚有一定难度。数据/语义层的互操作仍是异构分布式信息检索系统整合的重点,相关研究仍需深入。图 1。参考文献 10。

关键词 分布式检索 异构系统 互操作 系统整合
分类号 G354

ABSTRACT To realize the cooperative development of information resources in heterogeneous processors and heterogeneous data sources, we should make integration of individual subsystems in a distributed retrieval system. Although present technologies can be applied to realize the interoperability of specific distributed retrieval systems, it is still difficult to establish universal interoperability specifications. The interoperability at the data/semantic level is the key for the integration of heterogeneous distributed systems. 1 fig. 10 refs.

KEY WORDS Distributed search. Heterogeneous systems. Interoperability. Integration of systems.
CLASS NUMBER G354

1 异构分布式信息检索系统及整合问题的提出

分布式信息检索系统是由地域上分散、相对独立但又相互联系和制约的各部分(子系统),通过网络互联构成的完成特定功能的整体。各子系统的自治性较强,相互之间的耦合较为松散,可以对自己的功能和信息加以控制,制定自己的检索策略和方法。分布式检索系统使处于分布环境下的多个用户能够实现并行检索。

单一检索系统实现的是面向人的界面,检索过程是一个"人—机"对话的过程;分布式检索系统实现的是面向机器的界面,检索过程是一个"机—机"对话的过程。在分布式信息检索系统中,检索代理接收用户的查询请求后,首先判断哪些服务器可能含有与查询式相关的文档,并依据其相关性的大小将服务器进行排序;然后将查询式发送给排序靠前的 n 个服务器,并收集从各检索服务器返回的中间结果;最后将中间结果进行合并,形成最终结果列表返回给用户。

从体系构成上来看,分布式信息检索系统中的每个处理节点都可以是一台并行计算机,各节点依对应的子系统,或处于同等地位,或有主从之分,即分布式检索系统的数据源及检索处理器都可能是异构的。因此,要实现异构处理器、异构数据源之间的共建共享,必须对分布式检索系统的各个子系统进行整合。

要实现分布式信息检索系统的整合,最重要的就是能够使各子系统之间实现互操作。分布式检索系统互操作的主要障碍有:

(1)底层硬件异质:即分布式检索系统的检索处理器可能异构。比如各个子系统采用不同的 CPU,不同的通讯部件,或者不相容的系统架构等都会造成彼此间难以交互。

(2)操作系统及通讯模块异质:不同的操作系统下,数据存储模式不同;即使是同一类存储模式,其模式结构和属性空间也存在着差异,无法直接交互。此外,不同操作系统下通讯模块间的数据交换机制也不同,从而影响分布式检索系统各子系统间信息的互联互通。

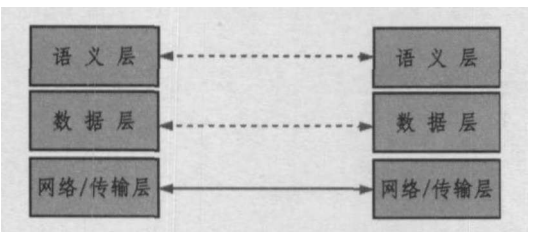
(3)应用系统功能逻辑异质:分布式检索系统各子系统可能采用不同的检索策略,不同的检索算法;支持不同的提问句法,不同的结果返回格式,即查询请求和查询结果的表达不同;彼此的语种不同,标引的词汇有差异等。

因此,如何将 Web 上分散的、异构的检索子系统联合起来,向用户提供统一的服务,即实现各个子系统之间的互操作,是异构分布式信息检索系统研究与开发的关键问题。

* 本文系教育部人文社会科学重点研究基地重大项目(06JJD870006)的研究论文之一。

2 异构分布式信息检索系统整合策略

信息系统的异构是有层次的，因而互操作也是有层次的。互联网设计的初衷是为了解决机器的互联互通，因而只要在网络/传输层面达到互操作就可以了；随着各类标准规范的建立，达成网络操作系统、分布式数据库等数据层面的互操作也逐步成为可能；当信息资源的增加和网络规模的扩张积累到一定程度时，仅仅数据层面的互操作往往不能尽如人意，直接表达和处理"语义"的需求就提了出来。具体来讲，异构分布式检索系统的互操作可以细分为网络/传输层、数据层、语义层三个层次，如下图所示：



2.1 通过网络/传输层的互操作进行整合

网络/传输层是异构分布式信息检索系统互操作的底层。为了利用其他子系统的数据资源或计算能力资源，需要平台提供相应的网络传输能力，即要求提供资源的服务器和请求资源的客户端之间能够协作解决问题。目前网络/传输层用到的互操作技术主要有：TCP/IP，CORBA，DCOM(.Net)，Java，中间件和Web Services。

(1)网络通讯与互联支持。异构环境下系统互操作的基本问题是实现无障碍的通讯与互联。以TCP/IP协议为核心的Internet是当前运用最为广泛、技术最为成熟的互联环境。TCP/IP既是一个协议集，又是一种通讯的技术实现，之后的许多高级通讯协议(如CORBA、DCOM等)都是基于TCP/IP协议发展的。在异构分布式信息检索系统中，既可以利用这些高级通讯协议，也可以直接在模块中调用TCP/IP协议的API函数库来实现网络间的通讯互联。比较而言，TCP/IP提供了高效、灵活而强大的网络互联通讯支持，适用于存在多种高层协议标准的复杂环境下的异构分布式检索系统的通讯实现。

(2)跨平台支持。异构环境下系统互操作的核心问题是实现不同计算平台和编程语言平台下的数据交换。CORBA和DCOM协议便是适应这一要求而产生的高级协议。CORBA(Common Object Request Broker Architecture，公共对象请求代理体系结构)是由OMG组织制订的一种标准的面向对象应用程序体系规范，在面向对象编程中用于实现分布式、异构系统、不同代码、不同计算环境下的对象实例间的通讯问题。DCOM(Distributed Component Object Model，分布式组件对象模型)是微软公司定义的一整套计算规范和程序接口。利用这个接口，客户端程序对象能够请求来自网络中另一台计算机的服务器程序对象，即对异构网络间的通讯可实现对象间的参数传递。与TCP/IP协议相比，CORBA和DCOM提供了更广泛的编程语言支持而不仅仅是传统的C语言，同时也提高了计算环境的扩展性。此外，CORBA的独特优点在于跨语言的支持，不同检索处理器、不同语言平台都可以利用CORBA框架来实现互联互通；DCOM的特点在于它对微软平台的无缝连接(支持VC.Net、VB.Net、VJ.Net和C#.Net等语言)，与CORBA相比，DCOM具有微软平台的高效与健壮性。

(3)统一互操作实现环境。丰富的互联协议为异构分布式检索系统的构建提供了多种选择，然而单一的协议选择往往会给未来的系统维护和升级带来技术风险，因此协议的灵活性和可动态选择性是构建新系统的重要方面。Java技术和中间件技术便是解决这一问题的优秀方案。建立于Java虚拟机(JVM)之上的Java技术，一方面通过使用统一的数据类型、调用统一的程序库，屏蔽了不同平台间的异构性；另一方面，内置的网络通讯能力使得Java可以灵活、动态地以多种网络通讯技术解决异构计算平台间的互操作问题。中间件是介于操作系统(包括底层通信协议)和各种分布式应用程序之间的一个软件层，具有强大的通信能力和良好的可扩展性，旨在屏蔽底层分布式环境的复杂性和异构性。因此，建立于Java技术或中间件技术之上的异构分布式信息检索系统可以在保持高层系统稳定性的前提下，灵活有效地改变底层的互联通讯协议。

(4)Web整合支持。现今传统的C/S结构的应用程序架构已逐渐被B/S结构的Web应用程序架构所取代，因而后者必将是未来异构分布式信息检索系统的主要实现方式。Web Services就是以Web环境为基础，在各种现有异构平台的基础上，构筑一个通用的、与平台无关、与语言无关的技术层，依靠这个技术层来实现各种不同平台的连接和集成。为了实现其目标，Web Services以XML作为数据描述和交换的标准，以WSDL作为服务的描述语言，以UDDI作为服务的注册和发现机制，以SOAP作为交换信息的协议。比较而言，Web Services是在各种网络新技术涌

现的背景下产生的集众优点于一身的并将广泛应用的互操作技术,该技术日益成为未来分布式环境较为理想的实现机制。

2.2　通过数据/语义层的互操作进行整合

对于计算机之间的信息交换来说,语法与语义缺一不可。数据层互操作是指通过建立统一的资源描述标准,解决信息交换中语法层面的问题;语义层互操作是指通过知识体系(概念术语、约束、关系、公理的表达)的参照、映射或其他方法,理解多个领域的知识表达,使信息系统具有语义交互的能力[1]。因此,数据层和语义层是异构分布式信息检索系统互操作的核心层次,涉及到异构数据源从形式到内容的整合。现有的研究通常将数据层、语义层的互操作结合起来进行考虑,未加严格区分。

(1)异构数据源描述标准。在异构分布式检索系统中,不同的检索子系统通常采用不同的数据表示方式和存储格式,使得各个子系统中的数据形成孤岛难以直接交互。因此,如何识别以及处理来自不同子系统的数据是异构系统互操作的前提。

为了使分布异构数据源具有互操作性,首先需要建立一个统一的信息资源描述标准。元数据即是描述信息资源或数据的一种结构化数据,是网络信息资源组织的重要工具。从构成上看,元数据是一个三层结构体,它包括语义、句法和内容标准。目前围绕着SGML,HTML 和 XML 等环境,已建立了各种元数据标准,其中较有影响的有 Dublin Core,PICS,CDWA,CDF 和 MCF 等。标准化的元数据集为特定领域的信息资源描述提供了一个基准方法,使得构建在此基础上的各检索子系统的异构数据源之间能够进行交互,从而为信息资源数据的有效整合及共享奠定了基础。

(2)不同描述标准的转换。在分布式信息检索系统中,尽管特定领域的异构数据源可以采用元数据进行统一描述,实现一定程度的交互,但不同的领域(甚至同一领域)往往存在多个元数据格式,对于整个分布式检索系统的资源整合来说仍然存在障碍。要在不同元数据格式表示的资源体系之间进行检索、资源描述和利用,就必须解决元数据的互操作问题,即需要完成多个不同元数据格式的释读和转换。目前元数据互操作的主要途径有:

① 格式映射方式:可利用特定转换程序对不同元数据格式进行转换,如 Dublin Core 与 USMARC,FGDC 与 MARC 等;或者利用 metadata crosswalks 技术以解决不同标准元数据间的集成访问[2];也可利用一种中介格式对同一格式框架下的多种元数据进行转换,如 UNiverse 项目利用 GRS 格式进行 MARC 格式和其它记录格式的转换。利用映射实现元数据互操作的准确率较高,但这种方法的应用效率在目前多种元数据格式并存的开放式环境中明显受限。

② 标准描述框架方式:即建立一个标准的资源描述框架,用这个框架来描述所有元数据格式。XML 和 RDF 从不同角度起着类似的作用。XML 有标准的 DTD 定义方式,只要能够解读 XML 语句的系统就能辨识统一定义的元数据格式,从而解决了对不同格式的释读问题。XML Schema 标准可以说是 DTD 的发展,不仅包括了 DTD 能实现的所有功能,而且本身就是规范的 XML 文档,并规范了文档中的标签和文本可能的组合形式。RDF 定义了一个由资源、属性和陈述等三种对象组成的基本模型,通过这个抽象的数据模型为定义和使用元数据建立了一个框架,元数据的元素可看成元数据所描述的资源的属性。

③ 数字对象方式:一种基于数字对象体系结构的元数据互操作框架,即通过建立包含元数据和元数据模式的数字元数据对象来解决元数据互操作问题,数字对象提供元数据之间的转换机制[3]。如 Cornell/FEDORA 项目提出由内核(Structural Kernel)和功能传播层(Disseminator Layer)组成的复合数字对象,可有效实现元数据之间的互操作。

④ 应用协议方式:即通过遵守相同的协议,提供应用层面上的数据发布和检索,如 Z39.50 协议和 OAI 协议。OAI 采用了中间层次的互操作策略,与 Z39.50 相比,实现的成本和对成员的要求均较低,易形成开放自由的大规模团体,但功能上相对而言要稍差一些。

(3)异构数据源的语义互操作。异构分布式信息检索系统的建设目标是屏蔽底层数据源的异构性,整合各个检索子系统,向用户提供围绕信息资源的统一的检索服务。从分布式检索的整个过程(文档集合划分、信息集选择、单文档集合检索、查询结果合并)来看,语义的表达贯穿其中。语义层的互操作是分布式检索系统互操作的高层,也是其资源整合的关键所在。元数据和本体均可在一定程度上解决数据源语义的异构性。

本体作为一种能在语义和知识层次上描述信息系统的概念模型建模工具,近年来引起了国内外科研人员的广泛关注。在 AI 领域,本体被定义为"共享概念模型的明确的形式化规范说明"。与元数据相比,本体可以从某种程度上弥补元数据难以对不同"粒度"资源进行描述、难以实现元数据方案本身的进

化、不具有普适性等不足，同时也部分解决了诸如灵活性和可扩展性等问题[4]。因此，与元数据相比，利用本体可较好地解决语义异构问题。

当前本体是语义互操作领域的研究热点，众多学者提出基于本体的语义互操作解决方案。Fuhr 和 Klas[5]认为可采用基于 agent 的三阶段（Matchmaking, Planning, Contract networks）选择过程来处理异构系统间语义上的互操作问题。Yi Shanzhen[6]等通过将领域本体和 XML 相结合，发挥各自优势来解决语义网 GIS 的信息语义异构问题。Leo Obrst[7]认为本体在语义互操作中起着重要作用，可以通过本体建立基本的语义表达，可以在本体中定义语义映射和转换规则，可以利用本体定义能够确定语义相似性的算法等，并指出使用基于本体和语义映射的软件能够减少异构应用系统信息交换中语义的损失。由上可见，目前大多数研究（特别是 Semantic Web 的研究）都把语义互操作归结为实现本体之间联系的研究。

本体虽然为异构分布式检索系统的语义互操作提供了解决方案，但其本身也存在着异构性，即需要解决不同本体间映射的问题。信息系统中常用的本体映射方法有定义映射、词汇关联、建立顶级本体、语义对应等[8]。通过本体及本体映射，可建立相关术语及服务的语境联系，以更好的实现分布式检索系统的整合。

2.3 通过异构分布式资源检索协议进行整合

为保证异构的信息资源和服务之间能够实现共享和协同工作，必须提供标准的检索协议。典型的适合异构分布式检索系统的协议有：①Z39.50 是严格基于 ISO 的 OSI 参考模型的应用层协议，其目的是规范查询格式，简化检索过程，实现异构机型、异种操作平台和不同图书馆系统之间的通信，实现分布式检索。②ZING（Z39.50—International; Next Generation）是 Z39.50 在 Web 时代的发展，既是 Z39.50 各种功能在新的网络协议和应用模式下的拆解，又是对 Z39.50 的一种简化。③STARTS（Stanford Agreement for Internet Retrieval and Search）协议没有规定底层传输协议的具体实现方式，仅作为一个元搜索引擎的架构，考虑了对于一组资源库，如果发出一个查询请求，该元搜索应能够找到合适的资源库实施检索；能够将用户通过统一界面以统一形式输入的全局检索指令转换为各个成员搜索引擎的局部指令语言；能够对结果集进行合并排序。④SDLIP（Simple Digital Library Interoperability Protocol）是一种基于 HTTP 或者 CORBA 的互操作架构，规定了查询接口、资源元数据接口、结果存取接口三类基本接口，灵活性和可扩展性较好。⑤SDARTS 协议[9]是由 STARTS 和 SDLIP 两个协议相互补充合并而成的一个互操作协议，充分利用了 STARTS 和 SDLIP 在互操作方面的优势和可取之处，同时对于元搜索支持的元数据形式进行了定义。⑥Dienst 协议是一个面向文档检索服务的大型复杂协议，支持对自治馆藏的分布式搜索。其互操作性通过采用相同的协议或软件结构实现。⑦Emerge 协议提供一套用户可定制的工具，用于在不同元数据视图之间建立映射，实现对多种数据源的搜索和查询以及进行查询前、后的处理，使跨越科学领域的搜索服务具有互操作性[10]。

上述异构分布式资源检索的互操作协议规定了客户机和服务器之间信息传输的报文语义和语法，建立连接和实施检索的逻辑顺序以及底层传输机制；同时，检索协议能够使用户可以获得检索服务器的相关信息，可以使用规范的查询语言向一个或多个数据库提交查询请求，可以以规范的格式接受检索结果。因此，检索协议既在内容上包含了网络/传输层、数据层、语义层三个层次，又在处理逻辑上把这三层有机的结合起来执行资源检索功能，从而实现异构分布式检索系统的有效整合。

3 总结与展望

分布式信息检索系统面临数据源异构和检索处理器异构的问题，要对其进行整合，即实现各个子系统之间的互操作，需要涉及多个层次、多种复杂因素，是一件非常困难的工作。现有的技术对于实现具体的、个别的分布式检索系统的互操作应该说是没有问题的，但是要建立普遍意义的互操作规范，尚存在一定难度：

（1）目前虽然有一些元数据描述标准和互操作协议，但是仍然难以做到通用。

（2）每个检索服务器一般都采用自己专用的查询语言，这使得每个查询在转换成后端查询语言时初始查询内容遗失的可能性会加大。因此如何进行查询转换，以尽量减少原查询内容的失真，是一个值得研究的问题。

（3）互操作是异构分布式信息检索系统需要解决的一个关键问题，而语义互操作是互操作的目的和重点。目前的互操作解决方案中虽然考虑到语义互操作，但大多数系统都是将语义互操作隐式地、内含地包含在语法和其它结构中，并没有当做独立的目标来考虑。因此，定义和设计独立的语义互操作层，使

分布式检索所包含的信息资源的语义"显性"化，将大大地促进分布式检索系统语义互操作问题的解决。

综上所述，数据/语义层的互操作仍是异构分布式信息检索系统整合的重点、难点所在，相关研究仍需深入。

参考文献

1 数图研究. [2007-04-26]. http://blog. donews. com/kevenlw/archive/2005/09/06/543143. aspx
2 J. Nogueras-Iso, F. J. Zarazaga-Soria, J. Lacasta, R. Bejar, P. R. Muro-Medrano. Metadata standard interoperability: application in the geographic information domain. Computers, environment and urban systems, 2004(28): 611 ~ 634
3 Christophe Blanchi, Jason Petrone. Distributed Interoperable Metadata Registry. D-Lib Magazine, 2001, 7(12)
4 石翌轶，宋自林，乔可春，艾未华. 基于本体的 Web 数据集成研究及实现方法. 情报科学，2006，24(4)
5 Norbert Fuhr, Claus-Peter Klas. Combining RDF and Agent-Based Architectures for Semantic Interoperability in Digital Libraries. Proceedings of the DELOS-Workshop on Interoperability in Digital Libraries, 2001
6 Yi Shanzhen, Zhou Lizhu, Xing Chunxiao, Liu Qilun, Zhang Yong. Semantic and interoperable WebGIS. The Second International Conference on Web Information Systems Engineering, 2001(2): 42 ~ 47
7 Leo Obrst. Ontologies for semantically interoPerable systems. Proceedings of the twelfth international conference on Information and knowledge management. 2003: 366 ~ 369
8 刘海滨，李冠宇，刘发军. 基于 Ontology 的信息集成研究综述. 计算机工程与应用，2005(25)
9 Noah Green, Panagiotis G. Ipeirotis, Luis Gravano. SDLIP + STARTS = SDARTS A Protocol and Toolkit for Metasearching. ACM/IEEE Joint Conference on Digital Libraries, 2001: 207 ~ 214
10 张付志，孔令富，刘明业. 几种典型的数字图书馆互操作协议分析比较. 情报学报，2003，22(4)

焦玉英　武汉大学信息管理学院教授，博士生导师。通讯地址：武汉。邮编 430072。

雷　雪　武汉大学信息管理学院博士研究生。通讯地址同上。

（来稿时间：2007-05-14）

（上接第 50 页）对管理效果进行评测，也可调查用户对新的质量目标、新改进的服务项目的满意度，找出用户不满意之处或对数字馆藏服务质量的要求。如果未达到预期目标或未满足用户对馆藏质量的需求，可以制定短期质量目标，重新利用系统图分析，对馆藏质量进行短期的管理和改进，直到达到管理效果，使用户满意，然后可以对用户新的需求和改进过的馆藏质量进行调查分析，进入新一轮循环管理。

据推测，到 2043 年，美国国会图书馆将有 50% 的馆藏数字化，2016 年将出现第一座大型虚拟图书馆[9]，这意味着数字资源在图书馆中所占比例将会进一步扩大。而数字馆藏作为现代图书馆的一个重要组成部分，服务效果对整个图书馆服务效果影响很大，其质量管理自然不容忽视。

参考文献

1 ［日］水野滋. 新 QC 七种工具. 北京：机械工业出版社，1991
2 伍爱. 质量管理学. 广州：暨南大学出版社，2003
3 秦现生等. 质量管理学. 北京：科学出版社，2002
4 黄清芬. 用户信息需求探析. 情报杂志，2004(7)
5 counter 项目. [2007-03-23]. http://www. projectcounter. org/
6 尹红，张宇. 数字资源评价指标研究. 四川图书馆学报，2006(1)
7 Database Evaluation. [2007-03-23]. http://www. westga. edu/library/depts/govdoc/dala-eval. shlml
8 索传军. 论数字馆藏管理政策. 中国图书馆学报，2005(5)
9 Cleveland, Galy. Digital libraries: definitions. issues and challenges. [2007-03-23]. http://www. ifla. org/VI/5/op/udtop8/udtop8. htm

索传军　国家图书馆研究院院长，研究馆员。通讯地址：北京中关村南大街 33 号。邮编 100081。

陈良金　郑州大学信息管理系 2005 级研究生。通讯地址：郑州。邮编 450002。

（来稿时间：2007-04-16）