如何建立网络信息查找检索系统

田 波　（上海财政税务局信息中心　200002）

摘要：本文对网络信息查找检索系统(NIDR)的设计和实现作了一个概括性的介绍和分析，着重讨论了网络信息资源的性质，资源的描述和分类，资源的查找过程，网络信息资源的描述信息的收集和管理等。

关键词：网络　查找检索

1. 引言

目前的网络信息查找检索(Network Information Discovery Retrieval)系统尚不能满足大量增长的用户需求。随着信息服务的产业化和商品化发展，这种矛盾更加突出。于是高效的 NIDR 系统已成为有效利用网络信息资源的关键。

网络信息资源(Network Information Resources)是指位于网络上的数字化对象，数字化对象的集合或相关的信息服务。网络上某个特定的数字化对象的定位过程称为资源查找(Resource Discovery)。通常这一过程是通过对网络信息资源的代理(Surrogates)进行操作而实现的。采用代理技术要求每个网络信息资源有一个唯一的标识符(Identification)，一般的标识包括文档名、地址等，如 URL(Uniform Resource Locators)和 URN(Uniform Resource Names)等。资源检索(Resource Retrieval)则广义地描述了网络信息环境中对网络信息资源的实际使用过程，包括从远程主机上下载文件(如 FTP)和远程交互式访问(如 Telnet)等。在此，资源查找和检索是两个独立的过程，一般的用户先查找相关资源，然后再检索得到有用的资源。用来帮助用户完成网络资源信息的查找和检索过程的系统便是网络信息查找检索(Network Information Discovery Retrieval)系统，简写为 NIDR 系统。

2. 网络信息资源的性质

网络信息资源主要有以下三种类型：

第一种是文档。网络信息资源包括大量的数字化对象，服务和信息流，其中大部分以文档的形式存储在主机上，并通过 FTP，HTTP 等传输协议在网络中传输。这类文档资源包括文本、图象、数字影像、可执行程序等。但从传输协议(如 FTP)角度来看无非是 ASCII 文本或二进制对象。这种在传输中丢失资源对象的类型语义信息是当前 NIDR 系统实现中的一个主要不足，因为这使得资源信息的互换性和可用性受到了局限。

第二种网络信息资源包括新闻组，电子邮件分发服务器和其他信息发布服务。网络信息资源主要以 ASCII 文本和二进制的方式发放，除了实际的资源信息外还包含一组变量用以说明信息的加密方法、文件格式、类型等，如 MIME 和 NNTP。

第三种是交互式访问。用户可采用标准的终端仿真协议(如 Telnet 和 X Window 系统)或特定的客户端实现协议(如 America Online)访问大量的文件，公告板或数据库。习惯上数据库只提供交互方式供用户直接访问。具体而言有两种实现方法：一种是通过网络关系数据库的 SQL 查询语言，如 RDA 协议；另一种是基于 Z39.50 协议的信息检索。后者的主要原理是在抽象语义层次上构建一个远端信息服务器，它提供了各种访问点，并能以一个通用记录传输语法向客户返回查找记录，从而对客户屏蔽了具体数据库的结构和类型。SQL 和 Z39.50 是定义在整个数据库层次上的，而非某个数据库中的记录或元组。基于 SQL 和 Z39.50 的交互式访问能提供比基于 FTP 或 HTTP 的文件传输更为丰富的网络信息资源的语义信息。

从网络信息资源的管理和使用上看，尽管 INTERNET 的信息分布于不同的网络和主机，但是用户应当可以通过通用软件和标准工具用统一的方法访问。因此可以将整个网络信息资源环境定义为一个统一的信息空间，信息空间内的信息分布对用户透明。

信息空间是概念上的重大突破。大约十年前，Gopher 系统采用菜单的方式构造了一个 Gopher 空间，而后的 WWW 以导航技术和超文本连接为基础构建了一个更灵活的信息空间。信息空间屏蔽了信息资源对象定位的复杂性，把用户的注意力集中到信息对象本身；另外，信息空间的定义使人们有可能将 INTERNET 的各类网

络信息资源作统一的管理，从而大大提高了 NIDR 系统的性能。

信息空间并不能解决所有问题。NIDR 系统尚需解决如下一些问题：

与信息服务记费相关的用户访问控制

·由信息兼容而产生的旧信息聚合对象的颗粒度问题

·由用户信息查找的随意性而产生的信息查找效率和查找信息的有效性问题

3. 资源描述和分类

信息资源的描述和分类由来已久，而网络信息资源则赋予了它新的含义，主要体现在描述和分类的层次和颗粒度的选择，以及信息资源的实时性要求上。

目前尚无一个公认的网络信息资源分类标准，这制约了 NIDR 的发展。由于网络信息经常需要更新和添加，实时性要求较高，给描述目录的编制带来了很大的困难。确定描述目录的更新周期成为保证其准确性和质量的重要因素。目前分类目录尚无一个标准方案，鉴于网络信息资源的特殊性，不宜借用现有的图书馆分类方案。

除了描述目录外，描述信息还包括“评价信息”。用户负责选择评价准则，NIDR 系统负责选用相应的评价信息（如信息的出处是否为某权威机构），以帮助用户优化查找结果。

确定资源信息描述的聚合层次和颗粒度十分困难。一方面，信息空间的定义打破了原来各独立信息源之间的界限，使得各同类信息对象高度聚合，用户往往无法通过进一步设定查找标准来缩小信息的查找范围；另一方面，诸如数据库之类的交互式访问，其信息对象的颗粒度又太细，由于数据库有自身的索引机制，使人很难从外部改变数据库信息的颗粒度。

在实际应用中，用户希望网络信息资源的聚合和颗粒度可以在一个很大的范围内任意选取，小到数据库的一条记录或一个 FTP 文件，大到整个数据库或整套 FTP 文档。可惜，目前的技术距离这一要求相差尚远。

4. 查找过程和资源描述

同传统的图书馆系统相比，信息的查找和检索在网络信息系统中显得愈发重要。信息查找需要用户提供查找标准，一般包括查找条目或主题词。基于条目的查找比较简单和直接，如查找“毛泽东选集第一卷”。基于主题词的查找要困难得多，如查找“毛泽东关于文艺工作的指示”。这需要系统能为主题词建立主题词库，通过上下文查找与主题词相关的术语，短句或文章，并建立映射，还要考虑交叉引用和多重索引等问题。

由于目前的网络信息资源环境同时存在着免费服务和记费服务，因而经常需要用户提供一些辅助的查找标准，如最大访问费用等。辅助查找标准还包括其他一些变量，如文件格式、压缩方法等。在查找过程中，用户可以用不同方法设置这些变量：可以采用一个通用的设置或根据用户所用工作站的软硬件条件，可用网络带宽等给出一个特定的设置，设置项包括文件格式、大小、压缩方法和图像的分辨率等。

描述网络信息资源的数据元素可分为两组：内部描述数据元素和外部描述数据元素。前者基于信息资源本身，典型的例子有：文章的标题、作者、摘要等。后者基于信息资源的管理和使用，典型的例子有每个数据库或 FTP 档案的所有信息资源的全局描述，包括文件名、信息的类型和数量等。实际的网络信息资源查找一般分两步进行，先找到相应的数据库，再找到数据库内相应的文档。第一步主要依赖于信息资源的外部描述数据元素，第二步主要依赖于内部描述数据元素。网络信息资源是不断更新变化的，为了保证信息资源查找的稳定性，要求内、外部描述数据元素作周期性的刷新。

5. 支持资源查找的信息收集和管理

支持资源查找的信息主要指上一节所提到的网络信息资源的内部和外部描述数据元素。这类信息是由描述数据库（descriptive database）统一收集和管理的。索引服务（indexing service）描述了描述数据库的生成和管理过程。我们把保存被某个索引服务索引的网络信息资源的主机称为资源服务器（resource server）。查找服务（discovery service）则描述了使用一个或多个描述数据库以帮助用户查找和选择相关的网络信息资源的过程。整个网络信息资源的查找便依赖于描述数据库。通常这是一种分布式数据库管理系统，每个节点数据库可构成一个自治服务体系，可以独立的维护和经营。

描述数据库的生成和维护采用两种基本的体系结构：拉模型（Pull Model）和推模型（Push Model）。

拉模型是一种轮询模型，其过程是：描述数据库的管理软件同网络上的资源服务器建立连接，查看其内容并生成相应的索引条目。由于索引服务的申请是由描述数据库发起的，因而不能及时的反映网络信息资源的变化，系统扩展性不好，再者索引条目的建立是一个索引服务与被索引的资源服务器的网络间的协同操作过程，网络

信息传输量很大。

推模型同拉模型正相反，是一种事件驱动模型。其过程是：当某个资源服务器的信息内容被更新后，它便发出请求同相应的描述数据库建立连接，申请索引服务，并将更新的内容传输至描述数据库以便更新索引条目。由于推模型中更新操作是由被索引的资源服务器发起的，因而能及时反映网络信息资源的最新情况，系统扩展性好。又由于只传输更新信息从而大大减少了更新过程中的网络传输量。但由于资源服务器和描述数据库通常存在一对多的关系，而且描述数据库的描述范围也会变化，因而推模型存在一个致命的问题：即资源服务器如何找到与之对应的所有的描述数据库。

鉴于两类模型各自的优缺点，有人提出了一种综合两种模型的混合模型：用拉模型建立资源服务器和描述数据库的初始连接，采用推模型实现描述数据库的更新和维护。这种混合模型尚无实例。

描述数据库的生成和维护是整个 NIDR 系统的核心，因而无论是拉模型还是推模型，其协议的健壮性要求都非常严格。

主要参考资料

[1] Clifford A. Lynch "Networks Information Resource Discovery: An Overview of Current Issues" IEEE Journal on selected areas in communications. Vol. 13. No. 8. Oct 1995

[2] 马鸿飞《INTERNET 资源与使用》西安电子科技大学出版社 1995

（来稿时间：1997 年 4 月）