

信息检索应用服务定义和协议规范分析与应用

沈 艺 (南京师范大学 210097)

摘要: 本文通过对信息检索应用服务定义和协议规范(即Z39.50)运行机制的分析, 给出Z39.50服务器端和客户端的实现要点, 并对目前Z39.50开发和使用时所存在的问题进行了探讨。

关键词: Z39.50 情报检索 信息共享

一、引言

Z39.50 是信息检索应用服务定义和协议规范 (Information Retrieval; Application Service Definition and Protocol Specification)的简称, 最初由美国国会图书馆等机构开发。1988年以版本1成为美国国家标准, 1992年出现版本2, 1995年出现版本3, 现在正在进行版本4的制定。国际标准化组织(ISO)已经接受Z39.50作为国际标准, 定名ISO 23950, 并且为了与原有标准ISO 10612/10613(Search and Retrieval Protocol)保持连续统一, 也定名为ISO 10162/10613 SR Version 2。Z39.50最初是针对图书馆机读目录(Machine-Readable Catalog, 简称: MARC)数据库共享而开发的标准, 随着应用范围的扩大, Z39.50已发展成为一般性的信息查询和获取标准。

二、Z39.50 运行机制

Z39.50 是一种高层次的网络信息检索标准。在开放系统互联(OSI)模型中, Z39.50 属于最高层即应用层的协

议。该层协议是向用户提供适合于应用和系统管理的信息服务。在实际的网络环境中, 它通常是运行在TCP/IP协议之上。Z39.50是面向连接的(session-oriented), 并且是有态的(stateful)。它不同于HTTP和Gopher这类无态的网络协议。Z39.50 服务系统常常要求支持有态的查询连接(session), 即先前查询的结果能被重用和细化。

Z39.50 是一个面向连接, 有关信息检索的协议, 包括书目信息、全文信息和商业信息等。它把互联的双方分别称为请求方(客户端)和服务方(服务器端)。当命令、检索方法各不相同的双方不能直接检索对方数据库时, 则需要通过Z39.50为双方提供服务。Z39.50建立抽象数据库概念, 使每个执行Z39.50标准的系统将抽象模型映射成自己专用的模型, 或者反向转换。也就是说当请求方向服务方提出服务请求时, 要把本系统的检索命令转换成Z39.50标准格式, 即使用基本编码规则1(ASN.1)中的规定, 把信息编成Z39.50的应用协议数据单元(APDU), 然后发往服务方。服务方则相反, 对APDU解码, 转成

自身系统的检索命令,再执行该命令。检索完成后按上述相反顺序把数据发回请求方,从而实现系统间的互联互通。图1给出Z39.50的一种典型的实现方式。

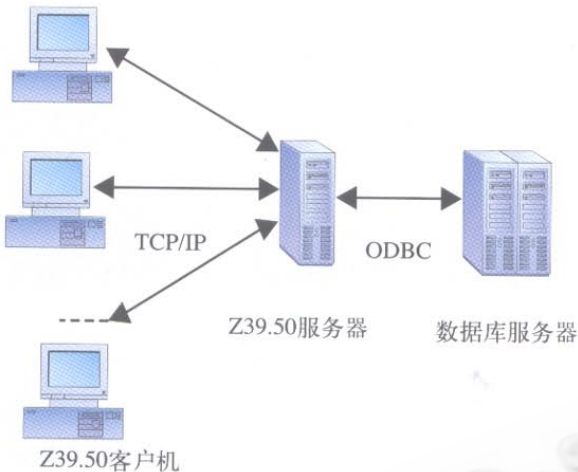


图1 Z39.50一种实现方式

三、Z39.50 服务器

Z39.50服务器的任务是侦听来自客户端的请求,根据请求类型和内容查询数据库服务器,并将结果送到客户端。具体地说,Z39.50服务器主要实现以下功能。

1. APDU 处理与事件循环

Z39.50客户端与服务器的会话是通过Z39.50消息,即Z39.50 APDU来进行的,Socket API负责完成在客户与服务之间传递APDU。APDU在客户与服务之间是以经过BER编码的字节流形式传递的。为了将APDU从ASN.1描述转换为程序设计语言表示,进而实现BER编码/解码,可用树结构表示一个APDU。

服务器系统启动后,在Z39.50协议默认端口210上侦听客户请求,进入事件循环状态。对于侦听到新的连接请求,即产生一个新的线程,所有与该客户的会话将完全由该线程负责完成,直到连接被某一方关闭,终止线程并释放其所占用资源。在Z39.50服务系统实现中,要确保系统能对多个客户同时请求的处理。

2. 查询请求分析

在查询请求中,查询属性是通过注册属性集中的注册数字来表达,属性集Bib-1中USE类型包含了大约100种书目数据的可检属性及其代码值。为了将接收到的Z39.50查询属性映射成MARC的一个或多个字段/子字段,或者本地检索属性,需要在服务端建立一张“属性-字段映射表”。将查询请求转换成查询数据库服务器的SQL语句,并将查询结果反馈给客户端。服务端处理Bib-

1/USE属性,可使用如下的语法:

〈本地属性〉|{〈MARC类型〉:〈MARC字段〉〈MARC子字段〉}{,〈MARC字段〉〈MARC子字段〉};

该语法表示,对于一个Bib-1/USE属性,或者有一个本地属性和它对应,或者被映射到一个具体的MARC类型中的某些字段或子字段。

映射表以文件的形式存放于服务端。当Z39.50服务启动时,映射表被读入内存,一旦接收到来自客户端的查询请求,服务端将依照该映射表和数据库模型快速地生成SQL语句。

3. SQL 语句构造

Z39.50服务接收到的查询请求通常是一个经过BER编码了的逆波兰(Reverse Polish Notation, RPN)查询表达式。通过对APDU解码,构造一个与查询请求等价的完整的SQL语句,并向数据库服务器提交查询。

在构造SQL语句过程中,首先要过滤掉查询请求中有可能包含的、服务器暂时不支持的查询属性,或者将其转换成系统支持的近似属性;其次对Bib-1/Use以外的属性,如“截词检索”(truncation)和“关系检索”(relation)等,可在构造的SQL语句中反映出来;此外对于不指明属性的查询项,应该增加Bib-1/Use属性值1016(即any),作为关键词查询看待。

4. 命名结果集

对于支持命名结果集Z39.50服务器,客户端应该可以随意指定结果集的名字。在多个请求同时使用相同的名字时,相同命名会发生冲突。解决冲突的办法是,在实现结果集命名时,服务器所使用结果集的真正名字,由客户查询请求中指定的名字和系统为该客户分配的唯一标识合并组成。如果系统不长期保存客户的查询结果,那么在中断与这些客户的连接时,Z39.50服务应该自动删除相应的数据库视图。

四、Z39.50 客户端

Z39.50客户端可以有三种实现方式,一是专门的Z39.50客户端软件,二是与图书馆应用软件集成在一起,三是采用通用的浏览器软件。这三种方式都需要有灵活方便的操作界面,让用户方便地构造和编辑检索式,完成复杂的检索任务,并可选择多个查询目标。查询目标由Z39.50服务器所在的单位名、数据库名、每个数据库的IP地址和服务的侦听端口号(缺省值为210)组成。客户可在这些目标中漫游查询。

1. 发送查询请求

在客户端与服务器连接成功后,客户端必须向服务器发送初始化请求,以实现客户端和服务器的初始会话,然后才能发送查询请求。

一旦用户确认执行查询,系统首先要做的就是分析和处理检索式,即将检索式由二叉树表示转换成 Z39.50 标准支持的一种查询格式,并进行 BER 编码。根据 Z39.50 标准, RPN 查询式是所有 Z39.50 服务器必须支持的查询请求格式,客户端应使用 RPN 格式提出查询请求。

2. 提取查询结果

结果的提取是通过发送一个请求消息来实现的。由于此时已经知道了服务器端查询得到的记录总数,所以可以采用多种结果提取策略,如逐个顺序提取、连续分段提取、不连续分段提取等。

采用一次提取一个记录的方法可能效率较低,可采用连续分段或不连续分段提取结果。如果一个查询得到的查询结果太多,往往是没有意义的查询,常会造成查询结果溢出,系统应对这类查询的结果提取加以限制或警告。客户向服务端请求的记录格式通常是符合 ISO 2709 标准的 MARC 格式。但 MARC 有很多类型,根据不同的 MARC 类型,记录被解释的含义也不同。记录的 MARC 类型由服务端返回记录时一起返回。客户端对每条记录不但要保存 MARC 类型,还要保存其出处和记录选中状态。

3. 查询结果的呈现

查询结果返回客户端后,必须以方便阅读和处理的方式呈现给用户。被处理的记录范围可能有 4 种类型:所有记录、选中记录、限定记录和当前记录。其中,选中记录指用户在已有结果中随机选择的零散记录;限定记录指用户按某种条件或规则对已有结果进行进一步细化的记录;当前记录指用户鼠标点击的那条记录。

为使用户的处理更为方便,可以将结果集以列表视图、限定视图和细节视图同时呈现给用户。列表视图是以若干字段为标题将查询结果的这些字段值依次列出,形成的一张二维表。用户在该视图选择当前记录和感兴趣记录;限定视图为用户提供用户构造、浏览限定条件。它可以让用户以若干字段的若干取值为条件,对已有结果进行进一步细化、筛选;细节视图提供对当前记录的详细解释,如全 MARC 字段/子字段分解、自定义格式解释等。

五、问题与对策

目前, Z39.50 多采用 C/S 结构,需要有专门的客户

端软件。如果采用 B/S 方式, Z39.50 与 WWW 结合起来,用户界面更加统一、简单、友好,可以极大地拓宽 Z39.50 的应用范围。为此要提供 WWW-Z39.50 网关服务,使上网用户可以通过浏览器检索 Z39.50 资源。

图书馆对信息资源共享的需求非常迫切, Z39.50 是实现资源共享的有效工具。而国内研制的图书馆应用软件,真正实现 Z39.50 极少。是否具有 Z39.50 服务功能,已成为图书馆选择信息管理软件的一个重要因素。虽然国外一些大型的图书馆应用软件实现 Z39.50,但其价格昂贵,几乎没有独立的 Z39.50 商用软件。Z39.50 软件大都与价格昂贵的国外图书馆自动化系统捆绑在一起,并且这些软件的内核和界面都是西文的,很少有支持中文图书情报检索服务和我国的书目数据机读格式标准 CNMARC (或 UNIMARC)。为此要开发具有中国特色的 Z39.50 应用软件,建立 Z39.50 服务,为信息资源的共享提供条件。

从国内现有的 Z39.50 应用情况来看,仅实现了 Z39.50 的一些基本功能,其服务功能,尤其服务器端的功能还需不断完善。

六、结束语

Z39.50 是目前众多的网络信息查询和获取的标准之一,也是目前所有这些协议中最繁杂、最难以理解和实现的一种。Z39.50 提供了从简单到复杂的查询规则,允许用户通过适当的查询请求来获取远程数据库记录。Z39.50 协议并不只限于查询文献书目数据,它原则上可以用于任何类型的数据查询。对用户来说, Z39.50 的优势在于它能使各种各样的信息资源看起来就象出自一个系统。■

参考文献

- 1 北京图书馆自动化发展部. 中国机读目录通讯格式. 北京:书目文献出版社, 1991
- 2 Marty Hall 著. Web 编程指南. 清华大学出版社, 1999
- 3 <http://icweb.loc.gov/z3950>