●翟晓娟

基于 php 的电子期刊检索网站建设核心技术

摘 要 电子期刊检索网站的开发环境,包括操作系统、Web 服务器和开发工具。Windows2000 + Apache + PHP 构架了电子期刊检索网站的开发平台。参考文献 3。

关键词 因特网 网站建设 电子期刊检索
分类号 G250.76

ABSTRACT The environment for the development of websites of electronic journal retrieval includes operation system, Web server and development toools. In this paper, the author discusses a platform based on Windows2000 + Apache + PHP. 3 refs.

KEY WORDS Internet. Website development. Electronic journal retrieval.
CLASS NUMBER G250.76

期刊资源信息含量大,出版周期短,传播速度快,是广大读者首选并被广泛利用的最有价值的信息源。随着信息技术的飞速发展以及网络时代的到来,期刊资源更是日益丰富,种类不断增多。传统的图书馆期刊资源建设理念已经不能适应形势的变化。因此,基于因特网的电子期刊服务引起图书馆人的巨大关注。建设电子期刊检索网站,可以使期刊充分发挥学术价值、信息价值和参考价值,更好地

(3)检索项之间进行逻辑组配,编制检索表达式。不管是数据库还是网上的搜索引擎,几乎都具有逻辑组配检索功能。由于已确定的检索项表达的主题概念间存在逻辑关系,可以通过使用布尔逻辑算符和位置算符对检索项进行组配,编制成检索表达式。常用的布尔逻辑算符有"逻辑与"(and)、"逻辑或"(or)和"逻辑非"(not);常用的位置算符有"near"、"with"、"field"等算符。

3.2 选择检索工具

(1)科研选题和科研进行中,由于要检索比较专业的、学术性较强的文献信息,如正式发表的期刊论文、会议记录等,应选择网上数据库检索系统,以保证检索结果的全面性和权威性。

(2)网上商务信息数据库的比重逐年增加,如中文的万方数据资源系统和美国的 Dialog 公司均提供许多很有特色的商务信息数据库,用户如要检索比较准确和系统的商务信息(包括政策与法规、市场、金融、商品等),也可以选择网上数据库信息检索系统,登录其网站,进入有关数据库进行有偿信息检索。

(3)检索时效性较强的信息,如新闻报道、最新商务信息等,可以选择网上搜索引擎。搜索引擎具有信息传递速度快且免费检索的特点,但对于学术性强,比较专深的课题,用搜索引擎检索效果不理想。

(4)针对具体的检索课题,可根据实际情况选用不同类型的信息资源灵活地配合使用,取长补短,即以专业性或综合性数据库检索为主,适当辅以网页搜索的检索方法,以达到最佳检索效果。实际操作中,用户最常用的就是这种综合运用的方法。

参考文献

1 黄如花. 网上电子期刊的利用. 图书情报工作,2001(12)
2 陈光祚,夏立新. 我国网络图书现状分析与发展对策研究. 中国图书馆学报,2002(2)
3 李家清. 开发利用网络信息资源的对策研究. 图书情报知识,2001(1)
4 司莉. 因特网上的图书馆虚拟文库建设. 图书情报知识,1999(1)
5 李毅萍. 网络报纸资源及其开发利用. 图书馆论坛,2001(4)
6 王建仑. 网络科研信息资源的开发和利用. 图书情报工作,2000(2)

王云娣 浙江师范大学图书馆信息咨询部主任,副研究馆员。通讯地址:浙江金华。邮编 321004。

(来稿时间:2002-09-19)

为教学科研服务,具有重要意义。

1　电子期刊检索网站的开发环境

1.1　操作系统

当前市场上用作 Web 服务器的操作系统主要有两大类:一类是微软公司提供的 Windows 操作系统,一类是 UNIX 系列操作系统。Windows 操作系统的优点是显而易见的,它的界面美观实用,操作简便,通俗易懂,更适合于非专业人员使用和维护,能够满足中小型 Web 站点,特别是一些访问量不太大的站点的需求。UNIX 操作系统拥有 Windows 操作系统不可匹敌的网络性能,其安全稳定的特点足以使得 Web 站点能够做到全年不宕机。专业人员显然更青睐 UNIX。

建立一个电子期刊检索网站,必须两方面因素权衡考虑。检索网站大多都是对数据库的读取操作,而删、增、改的情况较少,所以服务器的压力并不是很大。考虑到网管人员操作的简便易行,Windows 2000 不失为很好很实用的选择。

1.2　Web 服务器

现在流行的 Web 服务器有:Microsoft IIS, Netscape Enterprise, IBM Dom ino Server + Webspare 等。微软的 IIS 是 Windows 自带的,非常容易获取。IIS 使用多线程方式,基本上还是值得信赖的。但是系统管理员必须根据经验定期重新启动 NT 服务器,来预防不可预料的 Web 服务停止现象,要求更高的服务器只好采用昂贵的多服务器热备份系统。

免费 Web 服务器 Apache 使一切变得完美。Apache 充分考虑到进程带来的稳定性特征,以及线程带来高效率的特点。它会预生成多个进程,而每个进程中使用多个线程提供 Web 服务。由于存在多个进程,即使一个进程死了,也不会影响到整个 Web 服务。Apache 带来了稳定性和高负载能力。

Apache 属于 GNU 软件,有专门的 Apache 组织提供,可以免费下载,站点地址是 http://www.apache.org。

1.3　开发工具

当前,服务器端脚本是开发动态网页的常用方式,比较流行的有:Active Server Pages (ASP), JavaServlets, Personal Home Page (PHP)等。这些技术中,既包括有大公司支持的软件,如 ASP,也包括通过因特网进行合作开发的开放源代码软件,如 PHP。

PHP 除了向浏览器发送动态网页,还能发送不同 HTTP 头标识,使其能提供网页重定位、与 Web 服务器的安全认证结合的功能,以及设置 Cookie 的能力。PHP 能提供与多种数据库直接互联的能力,包括 MySQL、Sybase、Informix、Oracle、MsSQL 等,也能支持 ODBC。并通过额外的库,能够支持会话管理和 XML 处理(这些库都是基本的库,因此也是 PHP 用户的基本配置)。

现在的 PHP 功能相当丰富,除对 Web 数据库的支持,还可以进行 Web 安全认证,图像的动态生产,Socket 的链接,文件的读写,系统调用等功能。丰富的函数使它变得相当灵活易用,并能很好保证 Web 站点的安全可靠。至此,Windows2000 + Apache + PHP 构架了电子期刊检索网站的开发平台。

1.4　后台数据库

PHP 能兼容各种数据库,所以后台数据库的选择面就相当的大。一般来说 Oracle 稳定成熟,适用于大型数据库。而 MySQL 与 PHP 结合紧密,在运行小型数据库时性能优异,对于电子期刊网站是个不错的选择。但是笔者在建立网站的时候,更多考虑到 PHP 源代码的兼容性,所以采用了 ODBC 方式与数据库相连,这样 PHP 就无须过问后台是何种数据库,都一样可以运行。当后台数据库变化时,源代码基本可以照搬套用,不会有很大改动。而且 PHP 以缺省方式支持 ODBC,所以不需要改变 php.ini 的任何配置,就可以和数据库畅通无阻地建立联系。

2　安装文件的配置

安装电子期刊检索网站的开发平台,需要配置的主要文件是 Apache 的 http.conf 和 PHP 的 php.ini。

2.1　Apache 的 http.conf

首先需要指示 Apache 安装的主目录,只要把 documentroot 的属性改为主目录的路径名即可。例如:documentroot = D:/web.

其次,为了使 Apache 能够解释 PHP,必须在 http.conf 中加入 3 句话:

```
ScriptAlias/php/"c:/php/"
AddType application/x-httpd-php.php
Action application/x-httpd-php"/php/php.exe"
```

再次,如果缺省的语言不是中文简体,就必须将 AddDefaultCharset 的属性改为 ISO-2022-CN。

2.2　PHP 的 php.ini

首先将 extension_dir 的属性改为 php 安装的实际路径名。例如:C:\php。

其次,配置与 php 相连的数据库。例如:使用 sql server 数据库,只要将 extension = php _ mssql. dll 前的分号去掉即可。

再次,php4. 0 以上版本需要把 register _ globals 属性改成 On,使得页与页之间可以传输变量。否则,所有的表单变量都将在提交到下一页的时候被丢失,失去这一功能,即使最简单的 PHP 程序都无法正常运行。需要特别注意的是,如果服务器中存在多个 php. ini,一定要把每个文件的这一属性都改变过来才能生效,而不能只改变 c:\ winnt 下的 php. ini。

3 电子期刊检索网站的主要功能模块

3.1 数据导入模块

该模块的主要功能是把大量的数据导入到数据库表中,并建立相应的索引。

数据更新及时与否,是一个检索网站至关重要的方面,甚至可以决定网站的价值和生命力。数据的导入并不只是在网站的初始安装中进行,而是一个长期频繁的维护工作,几乎每隔一段很短的时间就需要系统管理员导入一次数据。数据导入模块的功能是评估检索网站质量的一个重要指标。

第一,页面外观。数据导入必须简单易行,界面友好,操作直观。系统管理员是检索网站面对的第 1 个来访者,数据导入时需要管理员键入数据文件的路径和名称,并且确认,最后返回一个导入成功与否的结果。导入之前的网页必须给管理员足够的提示,使他们能够比较容易地把文件名填写准确。数据导入的过程,需要很长时间,也许几十分钟,也许几个小时,甚至是 1 天。导入的进度需要程序体现出来,让程序员掌握情况,不至于误认为系统处于死机状态。数据导入完毕后提醒管理员重新启动服务器,这样可以关闭某些可能残留的进程,以免影响数据检索的速度。

第二,容错纠错。数据导入时,必须具有强大的容错纠错能力,而不能因为一些小小的意外情况就产生导入错误,甚至不能完成数据的导入。与此有直接关系的是数据库结构的设计。通常需要导入的数据都有一个固定的格式,字段的数量已经是个定值。人们需要注意的是字段的类型和长短怎样规定。笔者从实践中得出结论:只有当字段为字符型并且长度无限大时,容错能力最强。也就是说,此时无论数据为怎样的形式都可以被插入到数据表中。但是,实际操作中,不同的数据库对于字段长度都有所限制,只能在有限的范围内最大程度扩大字段的长度。当然字段长度过长会直接影响检索(select)的速度,所以就有必要建立原始数据表以外的表作为索引,来改善查询速度。至于字段的类型,建议原始表中全部定义为字符型,除非某一字段可以确定是数字而绝不会出现字符。可是面对如此庞大的源数据,有谁能够担保一定没有意外出现呢!

此外,在 PHP 程序中用单引号来标识 SQL 语句,当记录值中包含单引号时,就会被程序误认为表示 SQL 已经结束,这自然会出现错误,导致这一条记录无法成功写入数据表。所以每一条记录必须预先检查一遍,过滤掉所有的单引号,或者用别的符号代替。

第三,建立索引。索引是检索效率的保障,正是因为有索引的存在,才能让网站在几秒、几十秒的时间内把检索结果返回给用户。索引是在数据导入的同时建立的。这里把索引分为两类。一类是主表自带的索引,用 create index 语句建立,这类索引是帮助主表中的源数据排序,当指针遍历主表的时候,可以根据此索引很快找到需要的记录。其实,这就是人们通常所说的主键。第二类索引实质上就是一个数据表,它通常只有两个字段,其中一个是关键字字段,另一个是主表的外键字段。外键字段中的数值与主表的主键值多对一对应,当检索某关键字时,根据外键字段中取得的值到主表中寻找,就能找出与此关键字对应的记录(可能找到多条记录),从而完成检索过程。值得注意的是,建立索引的目的就是提高检索效率,因此,索引中关键字就应该尽量缩短长度,这样才能有效保证检索速度。

第四,源数据分年度存储。由于源数据通常都非常庞大,而且源源不断,所以如果把所有的数据都放在一张表中,主表将变得无限冗长,影响检索效率。解决这个问题的最好方法就是把源数据分年度存储。分年度建表和插入数据是数据导入编程的难点之一。首先需要做的就是在每插入一条数据之前检测该年度的表是否存在,如果存在就插入数据,否则就建表,建索引,再直接插入该条数据,然后进行下一次循环,检测下一条记录的年代以及此年度是否有数据表,如此循环往复。这样每次循环都做一次判断,根据判断结果决定是否新建数据表(同时建立索引),但无论判断结果如何都要插入数据,所以程序会显得过于冗长。笔者建议把巨大的 SQL 语句写成函数,在每次循环中间调用,从而有效简化程序。但是函数不能编得太多,否则会影响程序的可读性。

数据库中还应设计另一张表,专门记录源数据表的数量,每增加一个源数据表,就在该表中增加一条记录,它的值是年代。只要在检索页面上显示出该表的每一条记录,就可以动态地显示出目前共有哪几个年度的数据。

3.2 检索模块

这个模块是电子期刊检索网站的主体。它直接面向用户,为用户提供多种检索途径,并返回用户需要的检索结果。

检索途径有很多种,比较常用的包括作者、ISSN 号、期刊名、文章名、关键词、年代等入口。

笔者建立的网站提供简单检索和复杂检索两种方式。

简单检索,顾名思义,只须读者键入一个关键词,立即返回结果。在结果中,此关键词可能出现在题名、刊名和作者名中。

复杂检索(advanced search)中,编程的难点主要有三:全名检索、关键词检索和二次检索。

第一,全名检索。要实现全名检索的功能就必须在索引中加入全名作为关键词。但是全名通常都非常长,而索引为了提高检索效率,把关键词字段定义得很短,所以就有必要对全名做一些处理。首先应过滤所有的标点符号,其次需要过滤的是一些常用并且没有意义的词,比如英语中的介词、冠词、副词以及期刊名中常出现的 journal、review 等。这些词都存储在另外建立的非用词表中,把全名中每个词和非用词表中的值做比较,如果有相同的就在全名中去掉。最后压缩掉所有的空格,并且取出全名的前 50 个字符就生成了全名关键词。用户在检索界面键入全名后,也要做同样的处理,才能在索引中找到相应的记录,取得主表的主键值。这些工作相当烦琐,需要用到 PHP 的很多函数。

第二,关键词检索。在高级检索中可以分出刊名关键词、篇名关键词及作者关键词。建立索引时,需要把刊名、篇名、作者名中的所有词(过滤掉非用词后)作为关键词插入索引,并且用特殊的标识区分出篇名、刊名和作者名。当然,同一个全名中取出的关键词指向的都是主表中的同一条记录。索引和主表是多对一的关系,用户键入关键词时,可能会键入很多词和符号,只能取其第一个空格之前的字符串。如果有必要,可以提供多个填写关键词的文本框,便于读者输入多个关键词,以缩小结果集。

值得注意的是当存在多个关键词时,一定要先用每个关键词独立地在索引中检索,单一的关键词得到单一的数组,数组的值是多个主表主键值的序列。然后比较这几个数组,取出其中相同的值,作为最后的结果。笔者做过实验,如果不用这种方法,而先用一个关键词检索出结果数组,然后在此结果集范围内找出有其他关键词的值,那么检索速度反而大大降低。原因说明:假设有两个关键词,用第一种方法只需和数据库联系两次,然后比较一下数组就可以得到结果,但用第二种方法时,假设第一个关键词检索出 n 个主键值,那么第二个关键词至少要跟数据库联系 n 次做判断操作,才能得到结果集。和数据库打交道是程序运行速度的瓶颈,应该尽量减少和数据库联系的次数,才能有效提高检索效率。

第三,二次检索。所谓二次检索就是在上一次检索的结果中再次检索,提取出更贴近用户需求的记录。二次检索可以和一次检索在同一个页面上进行。二次检索时,用户输入新的关键词,然后提交表单,那么第一次的结果集必须自动保留到二次检索的阶段。笔者用了表单中的隐含(hidden)控件。当提交表单时,结果集就被自动隐含传送到指定的数组内。此时,用新的关键词得出新的结果集,最后再与原结果集比较,就得出了二次检索的最终结果。

参考文献

1　EBSCO 联合西文期刊篇名目次数据库. http://202.119.47.85//html/first.php(暂定)
2　高建忠,邵晶. 利用 ASP 技术实现图书馆电子期刊导航系统. 图书馆杂志,2001(8)
3　陈万米. 用 Linux + Apache + Oracle + Php 构架电子商务系统. 微型电脑应用,2000,16(9)

翟晓娟　南京大学图书馆技术部工作。通讯地址:江苏南京。邮编 210093。　(来稿时间:2002-10-21)