

网络 Web 信息资源自动采集入库的实现

陈天文

(潍坊市图书馆, 山东 潍坊 261041)

[摘要]为解决图书馆采集网络 Web 信息资源、组织专题数据库的迫切需求,通过对自动采集、入库关键技术,特别是 URL 地址转换的难点分析,指明了 Web 信息资源自动采集、入库的原理和思路,并以国家图书馆网站采集实例说明了自动采集、入库的过程。

[关键词]Web 信息资源 自动采集 自动入库

[分类号]G253

随着信息技术的迅速发展,信息的生产、存储和传递方式发生了革命性的变化,网络已成为交流和共享信息资源的重要平台,由于网络上存在大量有关对文化遗产、学术研究具有重要价值的信息,因此网络信息资源的采集和保存就显得非常重要。

图书馆作为人类信息资源的主要保存者,对开放存取的网络信息资源的开发和利用,已成为图书馆资源建设的重要组成部分,因此构建一个有效的网络信息资源采集、保存、服务平台,对网上零散的、无序的信息进行筛选、解

bLogic的应用服务器中。

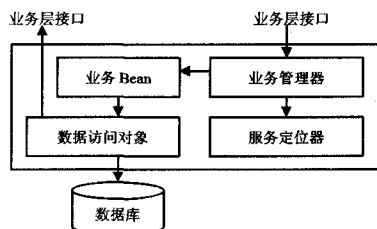


图5 基于J2EE数字参考咨询平台的业务层

业务层需要向表示层提供数据,而对业务对象每个方法的调用一般都是远程的。因此,客户端通过调用业务对象的get方法获取一个属性值,如果要获得多个属性值,这种操作的数量必然增加,也必然增加网络的负担,影响系统的性能。因此,采用数据模型来封装业务数据。当客户端向EJB请求业务数据时,客户端可以对EJB做单个远程方法调用来请求值对象,而不必启动多个远程调用来获取单个属性值。然后EJB构造一个新的值对象实例,把检索的值拷贝到该对象,并且该值对象的访问方法从该值对象中获取单个属性值。值对象是由可串行化的Java对象来实现的。EJB将处理之后产生的数据放入数据模型中返回后供JSP页面使用。图6表示了数据模型的序列图。

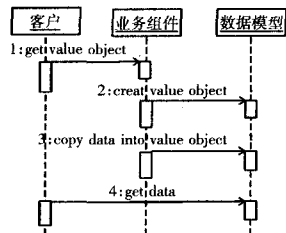


图6 数据模型序列图

4 结论

利用基于J2EE的多层体系结构构建的分布式联合数字参考咨询平台,既可保持基于Web体系结构的系统界面友好、开发方便、维护简单的特点,又通过Java把系统建立在J2EE之上使系统更加灵活、容易扩充和集成已有的第三方软件,适应数字参考咨询内容和服务多样化、个性化的要求,方便系统升级,符合数字参考咨询平台建设的特点和技术要求,是开发分布式联合数字参考咨询平台的良好解决方案。

参考文献:

- [1] 中国高等文献保障系统虚拟参考咨询子项目介绍. [2009-02-06]. http://www.calis.edu.cn/calisnew/calis_index.asp?fid=3&class=7.
- [2] 刘秋梅,郑耿忠.基于J2EE和XML的数字图书馆系统设计与实现[J].情报杂志,2006(7):3147-3148.
- [3] 刘秋梅.智能化数字参考服务系统实现初探[J].图书情报工作,2006(7):92-94.
- [4] 董慧,雷瑛.构建基于J2EE规范的数字图书馆模型的探讨[J].中国图书馆学报,2004(3):53-56.
- [5] 董慧,雷瑛.基于J2EE规范的数字图书馆示范模型的设计与实现(上)[J].情报学报,2004(2):185-190.
- [6] 董慧,张继东.基于J2EE的电子政务档案管理系统地构建与研究[J].现代图书情报技术,2006(9):73-75.

刘秋梅 女,1976年生。硕士,副研究馆员。发表论文10多篇。

郑耿忠 男,1975年生。副教授,博士研究生。研究方向:复杂系统建模及应用、网络计算与优化。

(收稿日期:2010-03-29;责编:张欣。)

构、整合,使之有序化,实现知识增值已成为很多图书馆系统建设所关注的问题。传统的网络信息资源采集,主要以人工采集为主,辅之以相应的计算机技术与网络技术。近年来,信息技术的发展使自动采集成为网络信息资源采集和利用的主要手段。自动采集可以减少重复性工作,大大缩短采集时间,节约人力物力成本,提高工作效率。

Web信息自动采集是利用爬行技术在网页中自动采集,先指定采集的起始页面,然后根据起始页面中的超链接采集延伸页面的信息。

1 图书馆Web信息资源采集方法

图书馆网络服务的快速发展使图书馆意识到采集Web信息资源、组织专题数据库的重要性和实用性,图书馆对Web信息资源采集曾进行过不同层次的探索和实验。目前主要采用两种方法实现:

一是手工采集、组织。手工采集是指图书馆工作人员在浏览过程中,发现所需要的信息后,手工下载保存。该方式检准率高,但效率低下,不能满足信息时效的要求,面对互联网中变化无常的海量信息,图书馆信息内容收集员不得不频繁地登录各大网站利用搜索引擎进行信息资料的发现、跟踪与下载等人为操作。

二是利用成熟的商业化软件。如TRS网络信息雷达系统、清华同方KSpider网络信息资源采集系统等,这些商业化软件的优点是技术成熟、操作方便、功能完善,大型图书馆采用的较多,但这类软件价格昂贵,中、小型图书馆由于经费限制难以承受,同时中、小型图书馆对自动采集的功能要求比较单一,所以如何开发一套适合自己的Web信息采集系统,对于图书馆组织网络信息资源意义重大。

笔者对Web数据自动采集的基本思路是通过分析特定网页源文件信息结构并根据HTML标识构造正则表达式,从而分离出相关字段及其内容,存储在本地数据库中构造专题网络信息资源库。其基本流程为:首先确定信息列表的URL地址,对信息列表页面的源文件进行分析,通过正则表达式提取出信息列表的超级链接集合,根据提取的超级链接集合自动转到正文部分进行自动采集。

2 Web数据自动采集的应用范围

采集、整合的Web信息资源大多来源于公开的网页资源,即半结构化数据,半结构化数据比较容易提取,因为数据中会有一些标识,如html文档,可以利用文档中的配对标识符来识别、抽取信息。这种方法的缺点是稳定性不够,比如html网页中的标识改变,就不能正常工作了。

Web数据自动采集与信息提取是面向不断增长和变化的某个具体领域的查询,并且这种查询是长期的、持续的。与传统搜索引擎基于关键字查询不同,信息提取基于查询,不仅要包含关键字,还要匹配各实体之间的关系,如标题、发布时间、正文等。信息提取属于技术上的概念,Web数据自动采集很大程度要依赖于信息提取的技术,以实现长期的、动态的追踪。同时,Web数据自动采集不是将Web查询结果直接输出给用户,而且通过进一步的分析处理,查重去噪,整合数据等,将半结构化的数据变为结构化的数据,然后以统一的格式呈现给用户,本示例是将网页格式的信息转化、存储到ACCESS数据库中。

3 开发环境:vb.net+asp.net+access2000

.NET技术是微软Web架构主流应用,目前大多Windows应用及Web应用的开发都以.NET技术为架构;access主要应用于中、小型系统的开发,按照.NET的三层架构开发的应用程序可以平滑地提升到SQL、ORACLE等大型数据库的应用。

4 自动采集的关键技术

自动采集系统通过对网页源文件进行分析,分离出信息的标题、发布时间、发布人、正文(包含图片)等内容。笔者主要通过构造正则表达式的方法进行提取信息,正则表达式由于难以读写,容易出错,所以找一种工具对正则表达式进行测试是很有必要的,它能够根据构造的正则表达式快速检索出网页源代码中的信息块,我们可以根据检索结果不断修正表达式,直到满足要求为止。Web信息自动采集过程中用到的主要函数有:

GetHttpPage:主要功能是根据信息列表,URL自动提取网页源代码返回HTML文档,该函数采用.net2.0的WebRequest和WebResponse两个类实现。

Get_url_Array:按照正规则匹配相应的数据,该函数主要通过MatchCollection类型收集网页源代码中与正则表达式相匹配的数据集合,然后把数据集合分类存储于本地ACCESS数据库中。

ReplaceSaveRemoteFile:获取源码文件中正文部分的图片并保存到本地根据日期形成的指定目录。

GetDateDir:根据日期创建目录,如20100320,根据日期建立目录或文件增加其灵活性,防止出现重名现象。

GetDateFile:根据日期创建文件,主要用来保存图片文件,如20100320001.jpg等。

DefiniteUrl:格式化连接地址。用于相对地址和绝对地址的转换,该函数自动判断获取的源文件中的URL是绝对地址还是相对地址,根据具体URL自动组配为绝对地址。

NoHtml:清除所有html格式,如果在应用程序过程中只保留纯文本,则调用该函数,如用于提取文章标题。

ScriptHtml:过滤部分HTML,用于获取正文内容的处理,过滤掉包括字体、大小、颜色、表格的元素,只保留表示图片的标识符。

在以上函数中,格式化连接地址DefiniteUrl是整个采集系统的核心,在网页信息列表的超级链接中有绝对地址和相对地址,其中相对地址有多种类型。所以在设计DefiniteUrl函数时必须把相对地址的类型全部包含进去,在采集开始前全部转换为绝对地址,通过绝对地址直接访问采集所需要的数据。组配绝对地址需要用到首页地址、列表地址。绝对地址的组配主要有以下几种类型:

首先确定要采集网页的首页地址(域名)和信息列表地址,如下所示:

http://www.nlc.gov.cn和http://www.nlc.gov.cn/book1/book2(例)

类型1:http://www.nlc.gov.cn/syzt/2010/0309/article_463.htm

该类型URL属于绝对地址,通过GetHttpPage函数获取源文件后直接提取所包含的相关信息。

类型2:/syzt/2010/0309/article_463.htm

该类型URL属于绝对地址,“/”表示根目录,与类型1描述方法不同,需要在该地址前面加上域名,组配结果为http://www.nlc.gov.cn/syzt/2010/0309/article_463.htm

类型3:/syzt/2010/0309/article_463.htm

该类型URL属于相对地址,“.”表示当前目录,需在该地址的前面加上当前列表地址,组配结果为http://www.nlc.gov.cn/book1/book2/syzt/2010/0309/article_463.htm

类型4:../syzt/2010/0309/article_463.htm

该类型URL属于相对地址,“../”表示上一级目录,所以必须把列表页地址按照../的个数循环去除子目录,然后链接该地址。本例组配结果为http://www.nlc.gov.cn/book1/syzt/2010/0309/article_463.htm

类型5:syzt/2010/0309/article_463.htm

该类型URL属于相对地址,组配方法同类型3。

信息正文采集图片,图片URL同样按以上规则组配。

5 国家图书馆新闻频道自动采集实例解析(以采集标题和正文内容为例)

5.1 确定采集的网站首页地址http://www.nlc.gov.cn

5.2 信息列表页面地址:http://www.nlc.gov.cn/syzt/bokelindex.htm

5.3 对信息列表网页的源文件进行分析,确定以下项目及正则表达式

5.3.1 网站编码:常见的编码方式有GB2312/UTF-8,如果选

择网站编码不正确,则会出现乱码现象。国家图书馆网站使用的是GB2312编码,在网站源文件头中<meta http-equiv="Content-Type" content="text/html; charset=GB2312"/>这一行说明。

5.3.2 获取信息列表超级链接集合正则表达式:/syzt/2010.*?htm,该表达式属于绝对地址,需按类型2进行组配。

5.3.3 获取正文标题正则表达式:<div align="center"><div class="big">.*?</div>

在获取正文标题后,用NoHtml函数过滤掉HTML格式。

5.3.4 获取正文内容正则表达式:[\wW]*?

获取正文内容后用ScriptHtml函数过滤掉除图片标识外所有HTML标识。

确定以上正则表达式后,就可以自动采集新闻频道的标题和正文了。

表1说明了本例用到的正则表达式符号。

表1

.	匹配除换行符以外的任意字符
*	重复零次或更多次
?	重复零次或一次
\W	匹配任意不是字母、数字、下划线、汉字的字符
\w	匹配字母或数字或下划线或汉字
[a-z]	匹配 a-z 任意一个字母
*?	重复任意次,但尽可能少重复,即匹配到最靠前的位置

5.4 自动入库

自动入库的实现是在处理正文数据时,通过循环每解析出一篇正文信息后当即把标题及正文内容通过SQL语句插入ACCESS数据库中,同时在入库过程中需要根据标题等关键字段对采集的数据进行过滤、去重,保证数据的唯一性和有效性。

网络Web信息资源的自动采集,为图书馆整合专题信息资源提供了方便,大大提高了图书馆服务效率和服务水平,通过不断积累,可以形成一个良好的本地信息资源使用环境,进一步拓展图书馆的服务范围。

(注:实验源码下载地址:http://www.wflib.com/cj.rar)

参考文献:

- [1] 蔡焰.网络信息资源自动采集探讨.江西图书馆学刊,2009(2).
 - [2] 韩群鑫.网络信息资源采集研究.农业网络信息,2007(4).
- 陈天文 男,1980年生。馆员,研究方向:图书馆自动化、网络化建设。

(收稿日期:2010-03-24;责编:杨新宽。)