

基于HTTP协议的OA期刊元数据动态收割研究

□ 吴广印 / 中国科学技术信息研究所 北京 100038
/ 北京万方数据股份有限公司 北京 100038
苏学 甘大广 / 北京万方数据股份有限公司 北京 100038

摘要: 文章针对基于HTTP协议的OA期刊元数据动态收割提出了解决方案,并以Open J-Gate元数据动态收割为例对解决方案进行了实证研究,以解决基于HTTP协议的OA期刊元数据的动态跟踪维护问题。

关键词: OA期刊, 元数据收割, 信息抽取

DOI: 10.3772/j.issn.1673-2286.2011.09.008

1 引言

自从《布达佩斯开放存取倡议》提出从自我存档(Self-Archiving)和开放存取期刊(Open-access Journals)两种途径来实现文献的开放存取共享开始^[1],OA期刊就逐渐受到社会各界的关注,尤其是各大数据厂商、高校以及图书馆。OA期刊规模较大的网络平台有DOAJ、Open J-Gate、HighWire Press、PubMed Central (PMC)、BioMed Central、PLoS等,他们均提供OA期刊的检索及期刊论文的全文链接。此外为了实现期刊元数据的共享,HighWire Press、PubMed Central (PMC)、BioMed Central等元数据符合OAI-PMH协议,可以通过OAI命令利用时间戳动态收割期刊及论文的元数据。文献[2,3]分别介绍了利用OAI-PMH协议收割的原理及实现。但是像Open J-Gate等网络平台还没有提供符合OAI-PMH协议的元数据,对于这部分元数据的收割只能通过基于HTTP协议的Web信息抽取来实现。目前,基于HTTP协议收割的介绍多数是一次性收割,而对于动态性维护很少有介绍,实证研究就更加少之。因此,本文就HTTP协议的OA期刊元数据的动态收割提出了解决方案,并且以Open J-Gate为例进行实证研究。

2 基于HTTP协议的OA期刊元数据动态收割方案

基于HTTP协议的OA期刊元数据动态收割方案是指定期对网络平台更新的Web网页数据的抽取,即周期性地对网络平台的期刊论文数据进行监测,并且把监测到的更新数据抽取为结构化的数据存储到数据库中。本文收割的数据主要是期刊论文数据,必须结合期刊及期刊论文的特性制定适合的解决方案。

2.1 期刊的特性

中华人民共和国新闻出版署在《期刊出版管理规定》中指出,期刊又称杂志,是指有固定名称,用卷、期或者年、季、月顺序编号,按照一定周期出版的成册连续出版物^[4]。根据给出的期刊定义,期刊论文具有刊名、年、卷、期等外部属性,而该期刊论文则可以根据刊名、年、卷、期逐级获取。如图1所示,一种期刊按照其出版年的不同可以划分为不同年份的期刊,不同年份的期刊又可以根据出版期的不同划分为期不同的期刊,这样一种期刊就按照出版年、期的不同具体到某一本期刊了,而这本期刊刊载了具

体的期刊论文，具体到每一篇论文就包括了诸多论文元数据，例如标题、摘要、关键词、参考文献等。像DOAJ、J-STAGE、SCOPIO、Open J-Gate、High Wire和BioMed Central等^[5,6]开放获取的网络平台都根据期刊的这一层级特性提供了期刊的导航功能。

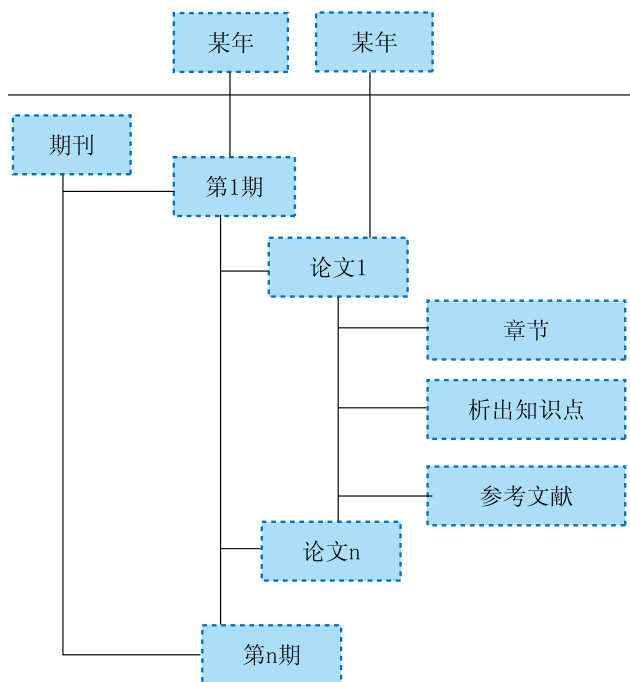


图1 期刊层级特性

2.2 具体收割方案设计

本文主要是解决不支持OAI-PMH协议的元数据的收割问题，而此类数据大部分是存储于Web页面的。考虑到当前的绝大多数Web页面是基于数据库的动态页面，动态页面的最大特征是具有相同的HTML框架，而数据则是由动态脚本语言（ASP、JSP、PHP等）从数据库中读取。所以相同HTML框架下的网页中变化的部分可能就是本文要提取的数据。因此，要实现基于HTTP协议的Web页面的元数据收割主要是监测网络平台更新的数据和更新数据的内容抽取。

本方案的设计思想是通过监测采集建立一个期刊监测库，按照期刊的特性分别获取期刊的刊名、年、卷、期信息并且到期刊监测库中进行比对，记录更新期刊的期的网址，然后抽取相应的网页的内容数据。

根据方案的设计思想，该方案主要包括以下几个步骤：

（1）构建期刊监测数据库

期刊监测数据库主要存储期刊的刊信息、期刊的年信息和期刊的期信息。期刊的刊信息主要存储期刊的刊名、期刊ID、ISSN号、出版者和该刊的URL，其中期刊的URL是下一步用来获取期刊年的信息的主要字段；期刊的年信息主要存储期刊的刊名、期刊年，以及期刊年的URL等，其中期刊年的URL是下一步用来获取期刊期信息的主要字段；期刊的期信息主要存储期刊的刊名、年、期和期的URL，其中期刊期的URL是下一步收割期刊论文元数据的主要字段。

（2）网页内容的信息抽取

网页内容的信息抽取贯穿整个期刊论文元数据收割的过程，按照获取内容的不同，可以分为期刊刊名列表的信息抽取、期刊年列表的信息抽取、期刊期列表的信息抽取和期刊论文的信息抽取。信息抽取的过程分为2个步骤，网页源代码的获取和网页内容的解析。网页源代码可以利用HTTP协议获取，网页内容的解析主要采用传统的文本信息抽取方法，利用起止字符截取其中的数据。

（3）期刊更新数据的监测

期刊更新数据的监测主要根据网络平台的特点进行监测，一般监测过程分为3个步骤：期刊刊名的监测、期刊年的监测和期刊期的监测，而且这3个步骤必须是依次执行的。首先，根据网络平台期刊列表页的网址抽取该网络平台所拥有的全部的期刊，并且逐一到期刊监测数据库的刊名列表中进行检索，插入刊名表中没有的期刊刊名及其他信息。其次，取出刊名列表中各期刊的网址，利用网页内容抽取逐个获取每一期刊的年信息，并且到监测数据库存储期刊年信息的表中进行检索，插入年表中没有的期刊年信息。最后，取出期刊年信息的期刊各年的网址，获取期刊每一期的信息，并且到监测期刊数据库中存取期的表中进行检索，插入数据库中没有的期的信息。

（4）期刊论文元数据的收割

经过期刊监测，获取到了期刊所更新的期的信息，利用期刊期的信息（主要是期的URL）和网页内容信息抽取就可以成功地实现OA期刊数据的动态收割了。

3 Open J-Gate实证研究

Open J-Gate 提供基于开放获取期刊的免费检索和全文链接。它由Informatics (India) Ltd.公司于2006年创建并开始提供服务。其主要目的是保障读者免费和

不受限制地获取学术及研究领域的期刊和相关文献。Open J-Gate系统地收集了全球约8637种期刊^[6], 包含学校、研究机构和行业期刊。其中超过1500种学术期刊经过同行评议 (Peer-Reviewed), 是目前世界最大的开放获取期刊门户^[7]。本文选取该网络平台作为收割对象进行实证研究。本文选用的开发工具是Visual Studio 2010, 开发语言是C#, 使用的数据库是SQL Server 2008。

3.1 方案的实施

根据2.2中提出的设计方案, 结合Open J-Gate期刊网络平台自身的特点, 该方案的流程如图2, 实施主要包括以下几个方面:

(1) 网页内容的信息抽取

网页内容的信息抽取是本方案实施的关键技术之一, 贯穿整个过程。无论是期刊刊名的监测, 还是期刊年、期的监测, 以及期刊论文元数据的收割都需要用到本技术, 因此首先介绍本方案中信息抽取的实现。信息抽取主要包括两方面, 即网页源代码的获取和网页内容的解析。在分析各种信息抽取方法的基础上^[8,9], 鉴于收割的对象是基于HTTP协议的Web网页数据, 本方案中网页源代码的获取主要是利用.NET Framework的基类WebRequest和WebResponse两个类封装了两个函数GetWebContentGet和GetWebContentPost。这两个函数一个是获取get提交的网页代码, 一个获

取post提交的网页代码。网页内容解析则是主要通过调用.NET Framework基类String类的成员public string Substring (int startIndex, int length) 方法封装了一个Mid函数实现的。Mid函数的输入为待抽出的网页源代码、起始字符和结束字符, 输出为待抽取的内容。

(2) 期刊刊名更新监测

利用Fiddler对Open J-Gate网站进行分析, 发现期刊列表页的参数提交方式是Post提交的, 因此期刊列表页的获取需要构造提交的post参数, 之后才能利用GetWebContentPost函数获取期刊列表页的源代码。源代码经过Mid函数的解析, 可以提取期刊的刊名、期刊类型、期刊ID和期刊网址等信息, 然后到期刊监测库中存储刊名的表中检索是否存在该种期刊, 如果在则新增记录, 并增加新增标识。期刊监测库中期刊表结构和内容如表1。其中NewAdd字段为新增字段, 如果该记录为新增, 则该记录内容为Y。

(3) 期刊年更新监测

进行期刊年更新监测之前, 首先要到期刊表中获取期刊的URL, 然后根据Open J-Gate网络平台的规律构造能够获取该期刊所有年份的网址。之后利用GetWebContentGet函数得到期刊的年份分布, 同样利用Mid函数抽取相关数据, 主要包括期刊年、期刊年的URL。与期刊刊名监测更新一样, 也需要对获取的信息到数据库中进行比对, 在数据库中新增没有的记录。此外, 对于已经监测完成期刊年的期刊, 还要修改其在期刊表中的NewAdd字段, 去除新增标识。期刊

表1 期刊表结构和内容

	JourName	JourID	JourType	JourUrl	DateTime	NewAdd	AddAll	f_id
1	Accounting & Taxation	130697		Articlelist.aspx?Journal_id=130697	2011-4-27	NULL	NULL	1
2	Acta Graphica : Journal for Printing Science and Gr...	131320		Articlelist.aspx?Journal_id=131320	2011-5-24	N	NULL	2
3	Advanced Computing : an International Journal	131315	Peer Reviewed	Articlelist.aspx?Journal_id=131315	2011-5-24	N	NULL	3

表2 期刊年表结构和内容

	JourID	JourYear	YearUrl	DateTime	NewAdd	AddAll	f_id
1	127693	2011	browse/Archive.aspx?year=2011&Journal_id=127693	2011-4-27	NULL	NULL	1
2	128671	2010	browse/Archive.aspx?year=2010&Journal_id=128671	2011-4-27	NULL	NULL	2
3	100202	2011	browse/Archive.aspx?year=2011&Journal_id=100202	2011-4-27	NULL	NULL	3
4	123751	2011	browse/Archive.aspx?year=2011&Journal_id=123751	2011-4-27	N	NULL	4

表3 期刊期表结构和内容

	JourID	Issue	JourIssueUrl	JourYearID	datetime	vol	NewAdd	f_id
1	NULL	NULL	browse/ArticleList.aspx?issue_id=15350435J...	http://www.openj-gate.com/browse/Archive.aspx?ye...	2011-4-28	Vol. 1 3 No. 1 1	NULL	1
2	NULL	NULL	browse/ArticleList.aspx?issue_id=15441635J...	http://www.openj-gate.com/browse/Archive.aspx?ye...	2011-4-28	Vol. 1 31 No. 1 144...	NULL	2
3	NULL	NULL	browse/ArticleList.aspx?issue_id=15340205J...	http://www.openj-gate.com/browse/Archive.aspx?ye...	2011-4-28	Vol. 1 61 No. 1 1	NULL	3
4	NULL	NULL	browse/ArticleList.aspx?issue_id=15401835J...	http://www.openj-gate.com/browse/Archive.aspx?ye...	2011-4-28	Vol. 1 19 No. 1 1	NULL	4

年表结构和内容如表2。

(4) 期刊期更新监测

期刊期的更新监测的方法和期刊年的更新监测相同，只是获取URL和插入期信息的表发生了改变。期刊期的表结构和部分内容如表3。

(5) 期刊论文的元数据的收割

经过期刊的刊名、年、期的监测，期刊更新的数据已经在数据库中存在了。利用这些更新的期的URL、GetWebContentGet函数和Mid函数，就可以实现OA期刊论文元数据的动态收割。收割的字

表4 期刊论文元数据结构

	title	author	authoradd	keyword	email	abst...	pdfurl	htmlurl	doiurl
79	The Physician betwe...	Urlika Zelaznik	Univerzita Primorsk...	medicine; health; medical e...			http://hrcak.srce.hr/file/35254		
80	War Health Care Syst...	Hivoje Cvitanovic	Department of Dema...	War in Croatia; warfare hea...			http://hrcak.srce.hr/file/33057		
81	Correlation between D...	Freyr G Sigmun...	Department of Ortho...	MRI Findings:Lumbar Spina...	freyr.gauli.si...		http://informahealthcare.com/...	http://informahealthc...	http://dx.doi.o...
82	Results of the Scandi...	Sigbjorn Smell...	Division of Surger...	Sarcoma Group,Osteosarc...	sigbjorn.smell...		http://informahealthcare.com/...	http://informahealthc...	http://dx.doi.o...

表4 期刊论文元数据结构 (续)

	issuevol/year	Year	journalName	publisher	peerReviewed	url
79	Vol. 8, No. 2, 1nbp;1nbp;1nbp;1nbp;2010	NULL	Acta Medica Historica Adriatica	Croatian Scientific Society for the History of...	Peer Review...	http://www.open-j-gate.com/browse/ArticleList.asp...
80	Vol. 8, No. 1, 1nbp;1nbp;1nbp;1nbp;2010	NULL	Acta Medica Historica Adriatica	Croatian Scientific Society for the History of...	Peer Review...	http://www.open-j-gate.com/browse/ArticleList.asp...
81	Vol. 82, No. 2, 1nbp;1nbp;1nbp;1nbp;1nbp;1n...	NULL	Acta Orthopaedica	Informa Healthcare	Peer Review...	http://www.open-j-gate.com/browse/ArticleList.asp...
82	Vol. 82, No. 2, 1nbp;1nbp;1nbp;1nbp;1nbp;1n...	NULL	Acta Orthopaedica	Informa Healthcare	Peer Review...	http://www.open-j-gate.com/browse/ArticleList.asp...

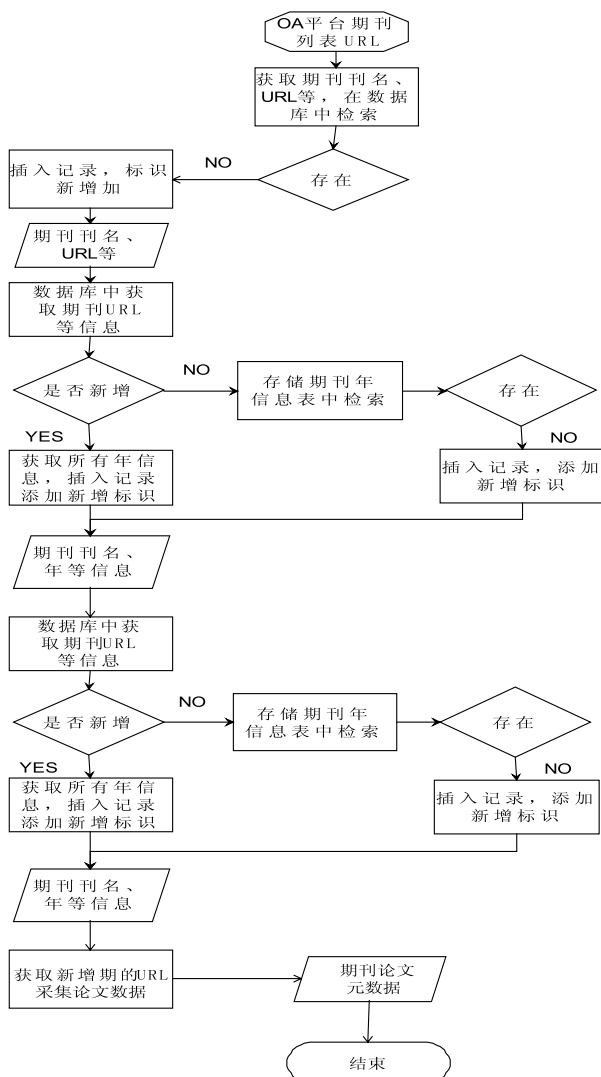


图2 Open J-Gate元数据收割流程

段主要包括标题、作者、作者地址、关键词、作者E-mail、摘要及摘要URL、DOI连接、全文连接等字段，如表4。

3.2 采集结果分析

(1) 数量分析

利用该期刊元数据收割工具进行期刊监测时，Open J-Gate网络平台给出的OA期刊数量是8637种^[10]，而期刊监测数据中刊名表中的数量是8414种期刊。期刊刊名监测完毕之后，监测到新增期刊的数量223种，而这与刊名表中的数量之和正是Open J-Gate网络平台给出的期刊数量，说明期刊刊名监测正确。由于Open J-Gate没有给出年期的统计数据，故无法利用数量验证监测结果。但是，在进行年监测时，监测出更新的年有1743条记录，由于其他原因无法监测的数据44条，出错率2.5%，基本可以接受。

(2) 质量分析

利用该期刊元数据收割工具收割的数据，能够全部收割网络平台给出的各个字段的数据，而且数据内容与网站内容一致。此外，也没有发现有乱码现象。

收割论文元数据总共938374条，对收割数据进行去重整理，未发现有重复数据。与之前收割的元数据进行查重，发现有46条重复记录，误差基本可以接受。

4 结论及进一步研究

本方案提出了不支持OAI-PMH协议的元数据动态收割方案,并以Open J-Gate为实例进行了实证分析,基本实现了基于HTTP协议的OA期刊论文元数据的动态收割。此外,利用OAI命令获取不到的字段也可以

利用该方案进行获取,满足了元数据收割的要求,在一定程度上促进了OA期刊论文的开放共享。

本方案的不足之处在于,网页内容的解析是基于特定模板的。当网页模板变化时,网页内容解析部分也需要进行相应的调整。因此,本方案还缺少对网页模板变化的判断,这需要在下一步研究中实现。

参考文献

- [1] Budapest Open Access Initiative [OL]. [2011-05-26]. <http://www.soros.org/openaccess/read.shtml>.
- [2] 王宇芳,等. OAI-PMH协议及应用新趋势[J]. 现代情报,2006(5):81-83.
- [3] 贾丽. 基于OAI-PMH架构高校学位论文元数据的远程收割[J]. 现代情报,2008(6):196-198.
- [4] 期刊出版管理规定[OL]. [2011-05-30]. http://www.chinacourt.org/flwk/show1.php?file_id=105838.
- [5] 罗爱静,胡德华,刘双阳. OA期刊的分布规律研究[J]. 数字图书馆论坛,2009(6):1-12.
- [6] 程维红,等. 国外科技期刊开放存取网络平台[J]. 中国科技期刊研究,2009,20(1):36-43.
- [7] 刘进军,李玲. 印度开放获取期刊集成平台分析[J]. 中国科技资源导刊,2009:34-36.
- [8] 陈要武. 网页信息抽取方法的研究[J]. 活力,2010(4):145.
- [9] 朱华. 浅谈网络信息资源采集技术[J]. 国家图书馆学报,2004(2):38-40.
- [10] Open J-Gate [OL]. [2011-05-01]. <http://www.openj-gate.com/Browse/ByJournal.aspx?alpha=ALL>.

作者简介

吴广印(1965-),男,中国科学技术信息研究所研究员,北京万方数据股份有限公司总工程师,北京万方软件有限公司总经理,研究方向:情报检索与知识组织管理。E-mail: gywu@wanfangdata.com.cn
苏学(1986-),男,研究方向:情报检索与知识组织管理。
甘大广(1984-),男,研究方向:情报检索与知识组织管理。

Dynamic Harvesting of the Metadata of OA Journals Based on HTTP Protocol

Wu Guangyin / Institute of Scientific and Information of China, Beijing, 100038
/ Wanfang Data Co., Ltd., Beijing, 100038
Su Xue, Gan Daguang / Wanfang Data Co., Ltd., Beijing, 100038

Abstract: Aiming to dynamically harvest the metadata of OA journals, this paper puts forward a scheme to dynamically harvest the metadata of OA journals based on HTTP protocol, and then conducts an empirical research over the Open J-Gate.

Keywords: OA journals, Metadata harvesting, Information extraction

(收稿日期: 2011-06-10)