# 借助八爪鱼采集器实现过刊网刊元数据的自动提取*

崔玉洁　　廖　坤
西南大学期刊社,400715,重庆

**摘　要**　现有的元数据提取方法提取规则烦琐、适应性差。针对这一问题,文章提出了借助八爪鱼采集器实现过刊网刊元数据提取的新方法。该方法以大型数据库的网页信息为对象,建立了提取元数据的流程图,通过该流程图设置相应的规则,并配置抓取数据模块,最后将该方法应用于网刊元数据的自动提取中。实际应用显示,该方法有效地提高了元数据的提取性能,并且具有较强的适应性。

**关键词**　采集器;网刊;元数据;自动提取

**Realization of automatic extraction of metadata in back issues of network journals by octopus collector**//CUI Yujie, LIAO Kun

**Abstract**　Existing metadata extraction methods have problems such as cumbersome rules and poor adaptability. To solve this problem, we propose a means of octopus collector to realize metadata extraction for published webzines. In this method, a large database of information on the page is regarded as an object, a flowchart of extracting metadata is established, rules are set through the flow chart, and the data capture module is configured. The method has been applied to the final webzine automatic metadata extraction. Practical application shows that the method can effectively improve the performance of metadata extraction, and has strong adaptability.

**Keywords**　collector; webzine; metadata; automatic extraction

**Authors' address**　Journal Press of Southwest University, 400715, Chongqing, China

**DOI**:10.16811/j.cnki.1001-4314.2016.05.024

随着信息共享技术的快速发展,数字资源库建设成为当前研究的重要内容,元数据是数字资源库建设中的关键问题。传统的元数据大多靠手工录用,提取效率低,制约了期刊数字出版的发展;因此,元数据的自动提取是当前期刊数字出版领域研究的热点[1-3]。

在论文的检索、分析和统计过程中,期刊论文的头部元数据(题名、作者、摘要、关键词等)起着至关重要的作用,为高效、快速地检索信息提供了保障。目前国内的期刊社或编辑部规模普遍很小,期刊网站上需要的元数据大多由人工手动输入,耗费了大量的人力、物力、财力。而万方数据、中国知网、中国期刊全文数据库、维普资源等作为国内比较有代表性的大型期刊数据库,拥有专门的人员和设备来处理并提取元数据,元数据的完整度和上传速率都较高[4]。这些大型数据库每期收到数据后首先提取元数据,然后将元数据上传至数据库中,如果期刊社或编辑部能够借助这些大型数据库网页中的元数据,将会极大地提高网刊上传的效率[5]。

本文提及的元数据提取是指从这些数据库的网页中提取所需要的信息,并将它们以一定的结构化形式保存下来,存入期刊社特定的大型数据库中方便用户使用。

## 1　网刊上传现状

在投稿系统的发行中心里,过刊数据的完善分为2个部分,即网刊元数据上传和单篇PDF全文上传。网刊元数据是指文章的基本信息,包括文章的编号、中英文题名、作者名、作者对应的单位、中英文摘要、中英文关键词、基金、起止页码、DOI、中图分类号等信息。单篇PDF全文的命名规则与元数据中的文章编号需要一一对应,这样,网刊中的元数据才能与全文对应起来。读者点击单篇文章时最先看到的是文章的基本信息,这些信息直观地显示了文章的核心内容,对提升期刊影响力起着至关重要的作用。

期刊在线投稿系统最近10年才逐渐开始使用,各期刊社或者编辑部网站上的数据大多保存得不完整,保存最完整的是纸质版的期刊数据,一篇篇扫描上传难度太大,过刊元数据提取工作量大并且容易出错。此外,大多数期刊编辑单位是采用方正排版软件排版,各期刊社或编辑部之间方正排版的命令会有一定的差异;所以,针对方正排版文件的元数据提取方法适应性较差,不易推广使用[6]。新出来的数据主要采用手工录入方式,效率低、错误率高,尤其是摘要中涉及公式时,经常会丢失符号和上下标信息,导致网刊中数据出现错误。

## 2　网刊中过刊元数据自动提取功能的实现

期刊社投稿系统中,网刊元数据的录入有多种方法。一种是逐个文件手工提取,这种方法耗时大,1期数据花费的时间基本为2~3 h,期刊社往往聘请学生来完成这项工作。另一种是借助软件公司来实现数据的提取和上传,例如北京仁和软件推出的面向学术期刊的元数据自动解析软件,获得了国内上千家优秀学术期刊的编辑部的支持与积极合作,帮助期刊编辑部

*中国高校科技期刊研究会2015年专项课题资助项目(CUJS2015-010);中央高校基本业务费专项资金资助项目(SWU1609165);全国理工农医院校社科学报2016年度基金资助项目(LGNY16B8)

实现了网刊元数据的提取和上传;但是该方法根据文章数量不同收费,通常 1 期是 200 元(文章数量不超过 25 篇),对于非市场性期刊来说,费用较高。还有一种方法是,各期刊社根据自身的情况研发软件或使用已有的软件,实现数据的自动提取[7]。

八爪鱼数据采集系统以完全自主研发的分布式云计算平台为核心,可以在很短的时间内轻松地从各种不同的网站或者网页获取大量的规范化数据,帮助任何需要从网页获取信息的客户实现数据自动化采集、编辑、规范化[8],摆脱对人工搜索及收集数据的依赖,从而降低获取信息的成本,提高效率。

### 2.1 八爪鱼软件的安装与登录

1)在网上下载八爪鱼采集器(Octopus_5.2_Setup.1450257345.msi)。

2)本软件需要.NET 3.5 SP1 支持,Win 7/8 已经内置支持,无须下载;但 XP 系统需要安装,软件会在安装时自动检测是否安装了.NET 3.5 SP1,如果没有安装,则会自动从微软官方在线安装。

3)注册账号并登录。

**2.2 采集网页列表信息的实现** 万方或者知网中每一期期刊的数据以列表的形式存放在网页中,人工提取的步骤为点击每一篇文章,进入相应的数据页面,复制题名、摘要、关键词等信息并保存,然后依次循环点击其余的文章。八爪鱼采集器的工作原理就是基于这一过程,对应的流程如图 1 所示。

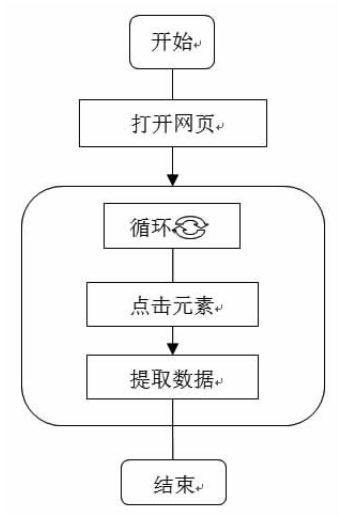


**图 1 系统执行流程**

#### 2.2.1 流程配置步骤

1)新建任务。首先打开八爪鱼采集器→点击快速开始→新建任务,进入到任务配置页面,选择任务组,自定义任务名称和备注。

2)打开网页。配置完毕之后,选择下一步,进入流程配置页面,往流程设计器中拖入一个打开网页的步骤;选中浏览器中的打开网页步骤,在页面的 URL 中输入网页 URL 并点击保存,系统会在软件下方的浏览器中自动打开对应的网页,如图 2 所示。


**图 2 网页打开视图**

3)执行循环。从图 2 的浏览器中可以看到网页是由同样的区域块组成的,我们需要抓取每一个区域块中的数据信息,而且每个区域块中的格式都是一样的。这时需要创建一个循环列表,循环抓取每一个区域块中的元素。点击图 2 中的第 1 个区域块,可以看到图 2 浏览器中的虚线框是选中了整个区域块的,如果选不中的话,可以在弹出的选择对话框里的选项上做调整。调整好之后,选择创建一个元素列表以处理一组元素;接下来在弹出的对话框中选择添加到列表。第 1 个区域块添加好之后选择继续编辑列表。接下来

以同样的方式添加第 2 个区域块。我们添加第 2 个区域块时可以看图 2,这时页面中其他元素都被添加进来了。这是因为我们添加的是具有 2 个相似特征的元素,系统会自动地将页面中其他具有相似特征的元素都添加进来。然后选择创建列表完成→点击图 2 中的循环。如上操作之后,循环采集列表就完成了。

4)点击元素。点击图 2 中第 1 个区域块,浏览器中打开的网页随之跳转到相应的页面。

5)配置抓取数据模块。点击图 2 流程设计器中的提取数据,再选择浏览器中需要提取的字段,然后在弹出的选择对话框中选择抓取这个元素的文本。上述操作之后,系统会在页面的右上方显示我们将要抓取的字段;接下来配置页面中其他需要抓取的字段,配置完成之后修改字段名称;修改完成之后点击图 2 中的保存按钮,再点开图中的数据字段可以看到,系统将会显示最终的采集列表。如图 3 所示。

配置抓取模板 (请点击你要抓取的数据)

| 字段名称 | 提取到的数据 |
|---|---|
| TT 标题 | 分数阶泛函微分方程边值问题正解的存在性 |
| TT 英文标题 | On Existence of Positive Solutions for Boundary Value Problem of Fractional Fun... |
| TT DOI | 10.13718/j.cnki.xsxb.2014.07.001 |
| TT 摘要 | 考虑一类分数阶泛函微分方程边值问题，利用锥上的不动点定理，得到了其正解及多个... |
| TT 英文摘要 | The boundary value problem of fractional functional differential equations have ... |
| TT 作者 | 宋利梅 |
| TT 作者英文名 | SONG Li-mei |
| TT 单位 | 嘉应学院数学学院,广东梅州,514015 |
| TT 中图分类号 | O175.8 |
| TT 关键词 | 分数阶泛函微分方程 ；边值问题 ；正解 ；存在性 ... |
| TT 英文关键词 | fractional functional differential equation , boundary value problem , positive so... |

**图 3　配置抓取模块视图**

6)元数据自动提取。点击图 2 中的下一步操作启动单机采集模式,进入任务检查页面,以确保任务的正确性;点击开始单机采集,系统将会在本地执行采集流程并显示最终采集的结果。执行结果如图 4 所示。本文借助八爪鱼采集器提取到的过刊信息为“中英文题名、中英文作者名、作者对应的中英文单位、中英文摘要、中英文关键词、基金、DOI、中图分类号等信息”,但是八爪鱼采集器只能采集网页上存在的数据,“栏目”“起止页码”这 2 项网页中没有的信息则无法提取,需要手动添加。


| | 标题 | 英文标题 | DOI | 摘要 | 英文摘要 | 作者 | 作者英文名 | 单位 | 中 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 一个高精度... | On Recons... | 10.13718/j.... | 给出一个高... | A kind of n... | 管林挺；郑... | GUAN Lin-... | 南昌航空大... | O24 |
| 3 | Euler-Poiss... | On the Blo... | 10.13718/j.... | 研究了宇宙... | Euler-Poiss... | 尹永刚；刘... | YIN Yong-... | 嘉兴学院数... | O17 |
| 4 | 李型单群 L5... | On a New ... | 10.13718/j.... | 众所周知，... | It is a well-... | 张伟[1]；... | ZHANG W... | 重庆第二师... | O15 |
| 5 | Keggin型Si... | On Synthes... | 10.13718/j.... | 采用改良的... | In this pap... | 尹彦冰；金... | YIN Yan-bi... | 齐齐哈尔大... | O61 |
| 6 | 耦合KdV型... | On Existen... | 10.13718/j.... | 利用平面动... | Applying t... | 何彩霞；刘... | HE Cai-xia ... | 贵州民族大... | O17 |
| 7 | 基于流星熵... | On Vulnera... | 10.13718/j.... | 道路交通系... | Road traffi... | 曹娟[1]；... | CAO Juan[... | 重庆交通大... | U49 |
| 8 | 具有全局优... | On K-Mea... | 10.13718/j.... | 传统的K均... | The traditio... | 李彬 | LI Bin | 乐山师范学... | TP3 |
| 9 | 哈萨克语NP... | On NP and... | 10.13718/j.... | 以现代哈萨... | This paper ... | 木合亚提·尼... | Muheyat N... | 新疆大学信... | 哈萨 |
| 10 | 恐龙化石的... | On Artificia... | 10.13718/j.... | 为了深入探... | To underst... | 邓建国[1]；... | DENG Jian-... | 四川理工学... | Q91 |
| 11 | 高于滚球半... | On Indepe... | 10.13718/j.... | 根据滚球法... | In lightnin... | 华国立[1]；... | HUA Guo-... | 雅安市防雷... | TM |

**图 4　系统执行结果**