文章编号：1005-8451（2013）03-0041-04

基于Web-Harvest的Web铁路信息采集系统的设计与应用

汤　立，李雪山
（中国铁道科学研究院　科学技术信息研究所，北京　100081）

摘　要：基于Web-Harvest开源软件，并对其功能进行了扩展，设计并实现了具有较强通用性的Web铁路信息采集系统，阐释了系统构架和相关的技术，并通过实例介绍了该系统的应用。

关键词：Web-Harvest；Web信息采集；开源
中图分类号：U285：TP39　**文献标识码：**A

Design and application of Web Railway Information Harvest System based on Web-Harvest

TANG Li, LI Xueshan
(Scientific And Technical Information Research Institute, China Academy Of Railway Sciences, Beijing 100081,China)

Abstract: Based on Web-Harvest OSS and its function extension, Web Railway Information Harvest System with universality was designed and applied, system framework and related technologies are expounded and examples were introduced to explain the application of the System.

Key words: Web-Harvest; Web information harvest; open source

随着互联网技术的快速发展，Web 信息呈爆炸性增长，万维网目前已成为重要的信息资源共享平台。基于某领域研究或应用的需要，人们经常需要从网上获取相关信息。例如，世界铁路发展动态追踪；某一事件的社会网络舆情分析；企业竞争情报研究等。因此，如何快速、准确地从海量数据中获取所需信息，已成为信息工作者研究的重点。设计开发基于 Web-Harvest 的 Web 铁路信息采集系统必将在提高信息资源建设质量和效率上起到积极作用。

1　基本功能描述

目前，国内外涉及铁路信息的网站门户较多，信息分散于各个网站，且更新速度不一，不便于信息获取利用。铁路信息采集系统的主要目的就是将分散、零碎的数据进行分别采集，统一存储，统一展现，为铁路信息工作者提供统一的信息检索门户，使其准确、快捷地获得其关心的内容。

基于以上分析，铁路信息采集系统应满足以下要求：(1) 能完成对多站点、多门户采集任务，及时发现网络中的新消息、新内容；(2) 能实现对站点内容的批量和增量采集，并能过滤掉信息噪声，精确采集到所关心内容；(3) 能自动识别并去除重复信息，减少数据冗余，保证良好的用户体验。

针对以上需求，本文以开源软件 Web-Harvest 为基础，通过网页特征分析，利用正则表达式实现了对网页信息的噪声过滤、XML 转化、数据采集和存储。实现了周期性、多站点、多任务的采集方式，提高了工作效率，保证了情报数据库建设的及时性和准确性要求。

2　网络数据采集系统分析

网络数据采集系统，也称网络爬虫。目前，相对成熟的爬行软件较多，分为商用和开源两

收稿日期：2012-11-17
基金项目：中国铁道科学研究院基金项目（2010YJ44）。
作者简介：汤　立，助理研究员；李雪山，副研究员。

种。基于今后方便系统集成、定制及二次开发等方面的因素考虑，本文仅对具有一定影响力的开源采集项目进行了调研，主要为 Apache Nutch、Heritrix 和 Web-Harvest 等。

2.1 Apache Nutch

Apache Nutch[1] 采用 Java 语言开发，源于 Apache Lucene 文件索引框架，后逐渐发展为一独立体系。它由网络爬虫（Crawler）和搜索引擎（Searcher）两部分组成。Crawler 主要用于从网络上抓取网页并为这些网页建立索引。Searcher 主要利用这些索引检索用户的查找关键词来产生查找结果。它提供了一种插件框架，降低了各功能模块之间的耦合度，使数据采集、查询、过滤等功能模块易扩展和易定制，极大地增强了 Nutch 的适用性。

2.2 Heritrix

Heritrix[2] 是一个由 Java 开发的开源 Web 爬虫系统，用来获取完整、精确的站点内容的深度复制，可通过 Web 用户界面启动、监控和调整，允许弹性地定义要获取的 URL。其最出色之处在于强大的可扩展性，允许开发者任意选择或扩展各个组件，实现特定的抓取逻辑。

Heritrix 由 CrawlOrder、CrawlController、Frontier、ToeThread、Processor 等主要组件组成，通过这些组件的配置及相互调用实现抓取工作配置开始、调度、URI 链接、网页内容处理及多线程控制。

2.3 Web-Harvest

Web-Harvest 也是由 Java 开发的开源 Web 信息提取工具，它提供了一种基于网页的快速定位、提取数据方法，主要利用 XSLT、XQuery 和正则表达式等技术实现了对 Text/XML 的操作[3]。

Web-Harvest 的主要目的是加强现有数据提取技术的应用。它的目标不是创造一种新方法，而是提供一种更好地使用和组合现有方法的方式。如图 1 所示，它提供了一个处理器集用来处理数据和控制流程，每个处理器可以被看作一个函数，它可以有输入参数和返回结果。并且每个处理过程是被封装成一个管道的模式，这样它可以将多个不同的功能的过程以链式的形式连接起来，分步完成某些复杂的数据处理和流程控制。此外，为了更易于操作和数据重用，它还提供了上下文存储已经声明的变量。

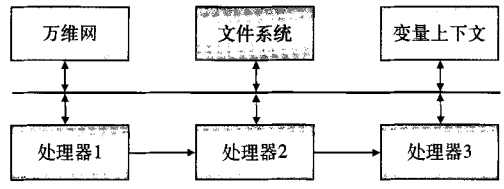

图1 Web-Harvest数据处理流程

本文对以上 3 个典型的 Web 抽取工具进行了比较，结果如表 1 所示。

表1 3个典型Web抽取工具比较情况[4-6]

比较项目	Web-Harvest	Heritrix	Nutch
运行方式	界面/命令行/程序调用	Web界面	命令行
定制能力	XML配置，灵活	配置参数较多	定制能力不够强
编码形式	XML配置、拓展性好	硬编码、拓展困难	硬编码、拓展困难
抽取过程管理	无	有	有
处理效率	处理步骤多、效率低	深度复制、较快	仅抓取和索引内容、快

以上分析可见，Web-Harvest 具有以下优点：(1) 基于 XML 配置文件实现信息抽取，可因网页格式改变，重新配置 XML 文件，无需硬编码，灵活、易用。(2) 可对每个特定的采集任务，分别编制 XML 文件，具有较强的通用性。(3) 支持 Beanshell、Javascript 以及 Groovy 多种脚本语言。(4) 开源软件，小巧，结构清晰，易于被集成到其它具体应用中。

与其它工具相比，Web-Harvest 尽管效率方面有所不足，而铁路信息采集相对于其它专业搜索引擎来说是基于领域的采集，采集量、采集范围相对较小，经过笔者试验完全可以满足要求。

3 系统框架设计

如图 2 所示，根据项目实际，基于 Web-Harvest 的信息采集系统主要由制定抓取规则、抓取控制器、信息去重、信息分类标引等 4 部分。

3.1 制定抓取规则

此步主要目的是对网页进行降噪处理，将要采集的信息实现从文本到结构化转换。

(1) 根据采集主题确定要采集的网页；(2) 对网页源代码（html 文件）的结构进行分析，确定要采集的具体信息及其所在的位置；(3) 利用 Web-Harvest 的 html-to-xml 命令将 HTML 文件转

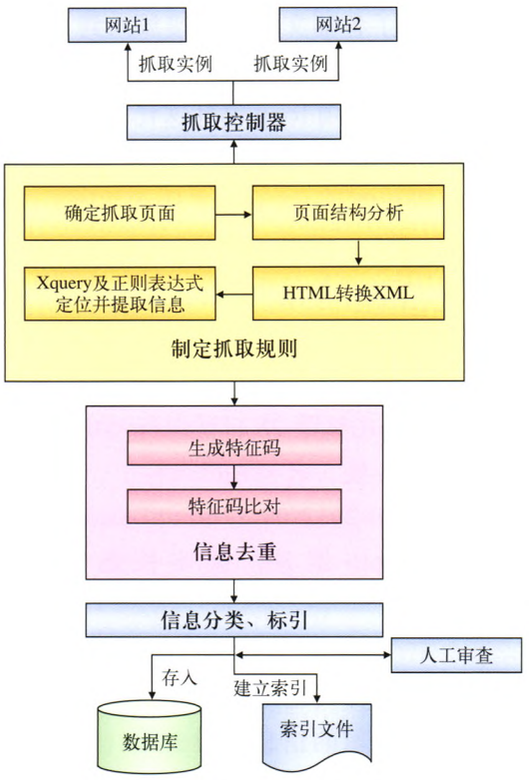

图2　Web-Harvest信息采集流程

换成 XML 文件；(4) 利用 XQuery 结合正则表达式将要获取的信息准确提取出来，完成数据结构化存储，同时完成 XML 文件的编制工作。需注意的是，在此过程中，应利用 Web-Harvest 本身所带的客户端工具进行调试和测试，以保证配置文件及采集信息的正确性。

3.2　抓取控制器

抓取控制器主要读取 3.1 中的抓取规则配置文件，实现对采集任务的管理，可完成对各任务采集策略、采集频率、存放目录的控制。主要通过 Web-Harvest 中的 org.webharvest.definition.ScraperConfiguration 及 org.webharvest.runtime.Scraper 类实现。

关键代码如：

```
ScraperConfiguration config = new ScraperConfiguration("c:/rails.xml"); // 读取配置文件
Scraper scraper = new Scraper(config, "c:/temp/scrapertest/"); // 设定工作目录
    scraper.execute(); // 执行采集任务
```

3.3　信息去重

在网络信息抓取过程中经常会出现内容相同的页面，不但浪费了存储资源，也加重了用户浏览的负担，给用户今后的检索带来诸多不便。本文研究采用了基于特征码的方法[6]，利用标点符号多数出现在网页文本中的特点，以句号两边各 5 个汉字作为特征码来唯一标识网页。其过程是：(1) 基于新采集信息的正文字段生成特征码；(2) 将特征码与数据库中已有信息的特征码进行比对，若有相同的特征码，说明数据重复，放弃存储，否则将新采集信息连同特征码一起存入数据库。

3.4　信息分类标引

为方便采集到的信息数据再利用，可结合铁路实际，对《中国铁路叙词表》进行精简、补充，形成铁路热点受控词表。基于受控词表，利用计算机文本处理技术，从采集信息的标题、摘要、正文等字段中提取受控词，并针对词频和预设的权重实现自动化标引。

同时可根据需要，对采集的数据进行进一步的人工审查干预，以保证信息的完整性和分类标引的正确性。

4　系统应用

系统采用程序调用的方式，基于 XML 配置模版对目标网页进行抽取信息配置，实现信息抓取、去噪去重、分类和存储功能。用户可根据数据采集的需求，定制抽取规则，进行信息的采集。

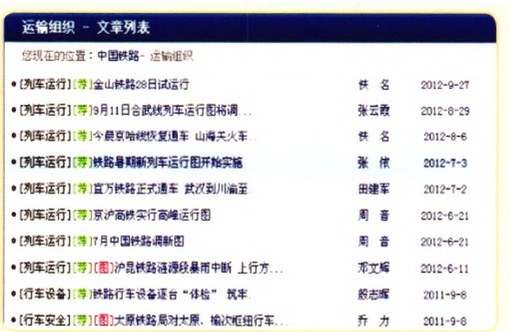

图3　铁路网运输组织专栏网页部分信息

在此以中国铁路网（http://news.chineserailways.com/）中的“运输组织”专栏如图 3 所示，具体阐明如何编制配置文件进行数据采集。其配置文件脚本如下：

```
<?xml version="1.0" encoding="UTF-8"?>
<config charset="ISO-8859-1">
  <include path="functions.xml"/>
  <var-def name="products">    <!-- 第 1 阶段
```

```
-->
        <call name="download-multipage-list">
          <call-param name="pageUrl">http://
news.chineserailways.com/HTML/NewsList.
aspx?NCID=13</call-param>
          <call-param name="nextXPath">//a[starts-
with(., '下一页')]/@href</call-param>
          <call-param name="itemXPath">//table[@
id="Dg_UserList"]/tbody/tr/td[1]/a[2]/@href</call-
param>
          <call-param name="maxloops">10</call-
param>
        </call>
      </var-def>
       <loop item="railItem" index="i" filter="
unique"> <!-- 第 2 阶段 -->
        <list>
              <var name="products"></var>
        </list>
        <body>
        <var-def name="oneItem">
          <html-to-xml>
            <http url="${railItem}" charset=
"GB2312"/>
          </html-to-xml>
        </var-def>
         <file action="append" path="D://rail.xml"
charset="GB2312">
            <xquery>
            ……
                </Rail>
            ]]></xq-expression>
          </xquery>
         </file>
        </body>
       </loop>
      </config>
```

在第 1 阶段，配置文件利用了 Web-harvest 自带的 download-multipage-list 函数，通过 <call-param name="pageUrl"> 配置了初始网页（种子网页），<call-param name="nextXPath"> 配置下一页的链接，<call-param name="itemXPath"> 定位信息块，并取出了各篇文章的链接放入到 products 列表中。

在第 2 阶段，用 loop 循环迭代 products 中每个链接页面 railItem，并将每个页面 railItem 进行 xml 转换，然后利用 xquery 表达式获取每篇文章 oneItem 中的标题、作者、发布时间及内容等信息，并保存到 rail.xml 文件中。经过以上步骤，可得到的 rail.xml 信息文件（片段）如下：

```
<Rail>
  <title> 金山铁路 28 日试运行 </title>
  <Author> 佚　名 </Author>
  <PubDate>2012-09-27</PubDate>
  <Content> 据悉，金山铁路 28 日 ...</
Content>
  ... ...
```

以上抽取任务可通过抓取控制器设置为周期性运行，并将获取的信息进行去重、分类标引并存入指定的数据库中。

5 结束语

本文基于 Web- Harvest 开源软件，根据实际需要，扩展了去重及自动化分类、标引功能，设计并实现了一个通用性强的 Web 信息抽取系统。用户可以根据自身需要，定制相关规则，实现信息抽取。但抽取规则需要掌握 HTML 结构、Xquery 及正则表达式等方面的知识，如何简化此部分的工作还需进一步研究。

参考文献：

[1] Nutch[EB/OL]. http://nutch.apache.org/about. html, 2012-10-25.

[2] 邱　哲，符滔滔，王学松．开发自己的搜索引擎 Lucene +Heritrix[M]. 2 版．北京：人民邮电出版社，2010，1.

[3] Web-Harvest [EB /OL] . http://web - harvest. sourceforge. net, 2009-12- 25.

[4] Heritrix Introduction [EB /OL]. http://crawler. archive. org, 2009-12-25.

[5] Nutch Tutorial[EB /OL]. http: //lucene.apache. org /nutch/tutorial. pdf, 2009-2-25.

[6] 王　哲．基于特征码的网页去重算法研究 [J]. 山东广播电视大学学报，2009（1）：14-15.

责任编辑　杨利明