网络信息采集技术在教育领域的应用研究

李奇涛[1] 管佳[2]
(1.对外经济贸易大学 信息化管理处 北京 100029;2.中央电化教育馆 资源综合部 北京 100031)

摘 要:为实现网络信息采集技术在教育领域信息采集过程中的应用,对网络信息采集技术进行了研究,在详细阐述技术架构及其核心技术基础上,完成了信息采集系统的构建。同时简介了其他两种信息采集技术,通过对比,分析三种方法的优缺点,方便了用户和研究者的选择与应用。
关键词:教育 信息采集 应用
中图分类号:G642 文献标识码:A 文章编号:1674-098X(2014)08(a)-0114-02

信息技术的迅速发展,使得网络上的信息日益增多,从日常生活到科学研究,人们越来越习惯于从网络上获取知识、信息,网络成为人们获取信息、知识的首要途径。但是,人们在面对如此繁杂巨量、形式不一的信息时往往感到无所适从。笔者在从事相关研究中就遇到这样问题,需要从某些教育技术资源网站中采集满足特定条件的信息。该文就针对这一问题进行了相关的研究。

在本研究中,需要从教育技术相关网站(中央电大开放教育教学资源查询系统、教育技术资源网、中国教育技术研究网等)中采集特定主题的,符合一定规律和格式的信息,因此信息采集的过程中,需要对网站进行分门别类,针对这些不同类型的信息,编写相应的代码和公式。同时,由于要采集信息量比较大,要实现对网络站点的自动填充和自动点击功能,在此基础上完成对网页的解析和信息抽取工作,其中要对采集的页面是否重复采集进行判断,最后实现对数据的精加工。本研究提出并实现满足上述需求的网络信息采集系统,并实现了在教育领域的应用。

1 网络信息采集技术系统结构

1.1 系统设计思路

该研究中设计的网络信息采集系统基于Windows平台开发,以Microsoft visio studio 2008作为开发工具,采用C#语言编写,数据采用XML存储格式,并实现与Oracle 10 g数据库连接。系统主要实现对相关目标网站信息的采集,采用单线程、固定模式、制定框架采集,针对不同网站制定不同框架模式,采集方式灵活。

1.2 系统基本架构

根据上述系统设计思路的简单介绍,网络信息采集系统的基本框架如下所示[1]:

(1)保存种子URL和待抓取URL的数据结构。

(2)保存已经抓取过的URL的数据结构,防止重复抓取。

(3)页面获取模块。

(4)对已经获取的页面内容的各个部分进行抽取。

(5)对抽取内容进行精加工处理。

(6)数据的存储。

系统所对应的机构图如图1所示。

系统运行的流程如下:

(1)确定要采集主题信息所在网站,并制定所要采集信息主题。

(2)将要采集信息主题导入系统中,由系统模拟点击搜索按钮,搜索本网站所包含与采集信息主题相关的信息。这里起始页面的URL为网站首页,将其放入采集器Web Spider中,通过相应设置,如:页面采集深度等,让采集器Web Spider对其进行爬取,搜索其中包含的URL信息,然后通过URL地址查新,分析其中是否含有新的、符合要求的URL,如有则将未抓取的URL加入到采集器Web Spider,继续循环采集信息页面,直至再无新的URL。

(3)采集器按照相应规则采集信息,调整页面结构,对页面实施规范化,并按照规则自动实现聚集,生成初步采集信息。

(4)采集过的信息经过信息提取,主要通过Xpath表达式提取,经过相应处理、格式转换等生成处理完毕的信息,并生成相应的索引,到此,信息采集就已完毕。

(5)将采集完的信息存储到XML文件格式中,按照需要,决定是否要存储到关系数据库中。

(6)信息展示。

2 核心技术

在本系统中,用到的支撑技术主要有URL地址查新技术、基于HtmlAgilityPack和Xpath的数据提取技术、模拟填充和自动点击功能,数据精加工技术。

2.1 URL地址查新技术

URL的地址查新是通过布隆过滤器来判断一个经过Hash函数散列的URL是否已经被访问过,从而避免重复采集同一URL数据以及程序陷入死循环。

2.2 基于HtmlAgilityPack+Xpath的数据提取技术

HtmlAgilityPack是一个开源的项目,为网页提供了标准的DOM API和Xpath导航。在整个系统中,HTML页面解析,文本抽取,遍历等都要用到这个包,而Xpath作为一种路径表达式工具,可以很好的“深入”WEB页面代码中的最小单位,精准定位到目标数据所在的代码行。通过将两者结合,可以有效地对经过解析的页面进行目标数据采集。

2.3 模拟填充和自动点击功能

模拟填充和自动点击主要针对例如百度这样具有搜索功能的网页。而大多数网站都具有站内搜索功能,WEB信息数据挖掘系统就可以利用这一功能实现信息抓取。对于我们要采集的目标网站而言,站内检索页面往往提供了普通搜索和高级搜索两种不同的搜索方式。普通搜索往往只提供了一个可供用户输入的文本框,而高级搜索则提供了除文本框外其他辅助选项(包括下拉列表框、互斥选项集等)。本系统采用Web Broswer控件来模拟用户的一次检索行为,包括填充文本框、选择下拉列表项和点击按钮等操作[2]。

同时对于JSP和ASPX、PHP等动态网页,如果没有明确的URL指向爬虫运行的下一页,则需要模拟点击页面中的“下一页”按钮或者点击下一页页标对应的超链接来实现(一般诸如[1],[2],[3]…等形式)[3]。

2.4 数据精加工

以上几步之后,得到的数据只是比较粗糙的“原始数据”,我们需要进一步进行精加工才能得到我们想要的数据。数据精加工分以下几种情况[4]:

(1)“原始数据”中经常出现诸如“ ”、“&amp”等HTML文本,我们需要将这些占位符去除。

(2)对于零散的原始信息,需要将其加工成规范格式,(例如新闻等信息,就要把标题,作者,发布日期等信息统一为诸如:某单位.关于召开XXX技术应用区域推进研讨会的通知[图].2009-5-5. http://jyjs.e21.cn/e21web/content.php?acticle_id=489)

(3)某些信息(比如作者信息,发布日期等)存在于一大段文字中的括号引号之内,或者在某些标点符号(逗号,冒号)之后,需要用正则表达式定位目标信息并将其进

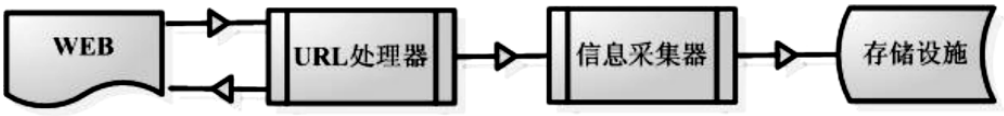

图1 系统结构图

一步抽取出来。此项涉及到自然语言处理等[5]。

(4)对于图片，PDF文档，RAR压缩包等文件，需要得到下载URL，然后导入下载程序进行下载。

3 系统实现

为了验证上述所提方法的有效性，这里通过实现一个简单案例来证明。数据提取内容为教育技术资源网（http://www.chinaret.com）下教育资讯栏目的信息。获取的信息内容主要是信息标题和信息URL链接地址。

首先加载WEB页面，通过XX Encoding.GetBytes("gbk")设置编码信息，然后定位目标数据所在位置，这里用到了Xpath表达式XX.GetElementbyId("content")，实际获得的值为http://www.chinaret.com/column.aspx?id=241/*[@id="content"]，意思为获取这个页面下所有ID为"content"中的信息，接着通过SelectNodes()来判断相应代码下是否包含要提取的信息，如本例中要提取的是链接信息，相应的代码就应该表示为SelectNodes("//a")，最后将Xpath表达是定位在要提取的数据节点上，提取节点信息，代码为GetAttributeValue()。

通过上述实验证明，采用本文所提出的技术能够很好的来实现对WEB页面信息的采集，可以应用到教育技术领域，为教育信息采集服务。

4 其他信息采集方法

在本研究中，除了上述介绍的C#语言编写的，采用HtmlAgilityPack+Xpath的采集方式外。还尝试了其他两种采用JAVA语音编写的网络信息采集方法。

其中VietSpider HtmlParser是一个纯JAVA的HTML DOM解析器，是一种开源的网络数据采集器。它提供一个图形化界面方便用户使用，可以用于特定主题、目的的网络信息搜索、采集和分类。其最大特色在于提供的图形化界面，使得数据采集简单化，正如其口号所说：Getting Web Data={Clicks}[6]。其主要特色如下：采用web3.0爬虫技术，提出网站模板解析概念，网络爬虫可以为每一个站点提供代理和多线程配置；VietSpider服务器可以在Linux/Windows系统下运行，管理员可以通过VietSpider的远程客户端进行管理；支持多种数据库系统，如：MySQL、MS SQL、ORACLE、Postgres、H2等；VietSpider提供了内置浏览器功能，支持JavaScript解析；支持多种数据输出格式，如MS Excel、CSV、XML等，支持数据除杂和改造。VietSpider的应用非常简单，所需专业知识较少，方便使用。

另外一种方法是采用Heritrix + HtmlParser组合系统方法。Heritrix是一个纯由JAVA开发的、开源的Web网络爬虫，用户可以使用它从网络上抓取想要的资源。Heritrix出色之处在于它的扩展性，使用者可以扩展它的各个组件，来实现自己的抓取逻辑。HtmlParser是一个用来解析HTML文件的JAVA包，主要用于转化、抽取两个方面。利用HtmlParser，可以实现文本抽取、链接抽取、资源抽取、链接检查、站点检查、URL重写、广告清除和将HTML页面转化为XML页面[7]。

从作者运行效果来看，三种方法各有优势。总的而言，从便捷性和提取速率来看，VietSpider较HtmlAgilityPack+Xpath和Heritrix + HtmlParser有较大优势；从存储格式上看，HtmlAgilityPack+Xpath的存储类型多样，并更容易与数据库结合；从灵活性而言，HtmlAgilityPack+Xpath和Heritrix+HtmlParser又较VietSpider简单，扩展性较强；从采集方式而言，Heritrix+HtmlParser需要分为两步，而VietSpider和HtmlAgilityPack+Xpath采用的是在线采集方式，一步到位。因此，结合以上分析，作者最后采用HtmlAgilityPack+Xpath方式来实现WEB数据的在线采集。

5 结语

网络信息采集技术属于数据挖掘领域，是WEB数据挖掘研究的热点。本研究中通过对网络信息采集过程中URL地址查新技术、基于HtmlAgilityPack和Xpath的数据提取技术、模拟填充和自动点击功能，数据精加工等关键技术的介绍，为读者提供了一种实用工具和研究思路。通过在教育技术资源网信息采集中的应用，实现了在教育领域对信息采集技术的尝试。同时通过对笔者在研究过程中尝试的几种方法的介绍和对比，方便读者在以后的研究和工作，研究者可以采用适合自己的工具进行相应研究。

参考文献

[1] 罗刚.使用C#开发搜索引擎[M].北京:清华大学出版社,2012:22-114.
[2] 孟宪军.互联网文本聚类与检索技术研究[D].哈尔滨:哈尔滨工业大学,2009:89-108.
[3] 于满全.面向人物追踪的知识挖掘研究[D].北京:中国科学院计算技术研究所,2006:15-35.
[4] webBrower控件实现winform和webpage交互[EB/EL].(2008-03-28)[2012-10-23].http://www.cnblogs.com/AganCN/archive/2008/03/28/1090737.html.
[5] (美)Jeffrey E.F.Friedl.Mastering regular expressions[M].O'Reilly,2007:14-37.
[6] VietSpider网站[EB/EL].(2012-03-13)[2012-10-19].http://binhgiang.sourceforge.net/webextractor.
[7] 罗刚,王振东.自己动手写网络爬虫[M].北京:清华大学出版社,2010:24-36.