基于 PowerBuilder 的网页数据抓取①

Web Data Extraction Based on PowerBuilder

刘书华　陈国奎　(衡水学院 数学与计算机科学系 河北 衡水 053000)

摘　要： 互联网飞速发展，WEB 已经成为一个巨大的信息资源库，各行各业的信息均可以在互联网上找到。及时准确的获得、存贮、分析、利用这些信息是非常重要的。利用 PowerBuilder 和 MicroSoft SQL Server 数据库，提出了一种对网页的数据抓取的方法。用户首先选定样本页面，其次在样本中预先定义抓取模式，然后对样本网页和其中的样本进行标记，形成信息的抓取规则，进行数据抓取，并存入数据库。最后利用数据库对信息进行分类，抽取出所需的信息，达到分析准确、抓取速度快的目的。

关键词： HTML 模式 数据抓取 抓取器 数据挖掘

1　引言

随着 Internet/Intranet 的迅速发展，WEB 已经成为一个巨大的资源库，并且数据量仍在快速的增长，成为全球传播与共享科研、教育、商业、社会信息等最重要和最具有潜力价值的信息资源。如何快速有效的利用这些信息，是一件非常有意义的事情。但是由于 WEB 数据(尤其是 HTML 网页)没有明显的信息结构[1]。在以前的数据抓取的过程中，一般采用人工的抓取方法，这样的工作效率低下。本文采用了添加模式和抽取关键信息的方法，大大提高数据抓取、分析、利用的效率。

2　抓取概述

需要抓取 WEB 站点的信息显示特点一般是：同一 WEB 站点的同类数据信息表示的 HTML 结构是相似的，尤其是对于大量的信息发布，通常都是采用相同模板或者基于某种动态网页技术生成的，通过 HTML 或者 XHTML 格式表现出来，具有一定的相似性，尤其是 XHTML，它描述信息的形式，自由、灵活、意义明确。所以，数据抓取的特点是针对相似 HTML 结构的网页。数据抓取的过程如下：选定样本页面→定义模式、生成规则→进行抓取→去除冗余 Html→数据入库并二次整理。

运行数据抓取器，程序主界面如图 1 所示。

图 1　程序主界面

3　选定样本

在同类网页中选择样本是为了生成网页抓取的规则，以黄页 www.99888.cn 的信息为例，见表 1。

表 1　黄页信息的页面示例

[铸造锻造 粉末冶金]龙玉锻压有限公司 本公司座落在美丽富饶、全国文明城市-长江金三角，江阴市东南 18 公里	
主营：	专业生产在中小各种锻件及法兰
地址：	江苏省，无锡市，周庄欧洲工业园砂山大道 168 号
网址：	点击访问
企事业单位性质：	G
员工人数：	108 人
电话：	0510-86900955

① 基金项目:衡水学院青年专项课题基金项目(2008057)
收稿时间:2008-07-27

选定的样本页必须具备所要抓取所有字段的内容，因为它决定着要抓取数据的内容，如果遗漏了一个抓取字段，比如电话，那么所有的抓取都不会得到电话这个字段。样本的重要性和意义体现在它是一个标准，是一个制定具体抓取规则的标准。

4 模式

对于同一类网页预先定义模式，是为了给抓取出的信息增加语义信息并对结果重构。对于黄页信息页面，用户可以定义出模式信息：厂家名称，地址，电话，联系人等，为后期的数据整理做准备。

4.1 规则设定

为了能够自动抓取数据的信息，必须知道抓取信息的规则，也就是从 WEB 页面的 HTML 中区分和辨别要抓取有用信息的方法。对应于预定义模式中的每个属性都需要设定相应的抓取规则。制定以下特征描述抓取规则：1)抓取信息的左右标识；2)抓取信息的文本特征；3)按照字段的顺序选定标识。

在抓取信息时，系统是按照一个个的属性进行抓取的。设第 k 个属性的左右边界为 begin_k(起始位置)，len_k(字符长度)，right_k(后续起始位置)，left_k_str(起始标识)，Right_k_str(结束标识)。

4.2 规则设定实例

下面以黄页的 HTML 属性 compnay_name(厂家名称),business(厂家主营) 信息为例做规则设定[2]

```
//厂家名称的规则设定及代码实现
Long begin_name; //声明起始位置变量
Int len_name,end_name;//声明字符长度和后续起始位置变量
String Left_name_str,Right_name_str ,company_name;//声明起始、结束标识和字段变量
Left_name_strt=”<h3 id=~"title~">”;
Right_Name_str=”</h3>”;
Begin_name=pos(html,left_name_str);// 得到 left_name_str 在 HTML 中的起始位置
If begin_name=0 then //未找到匹配字符,下一条 url
Continue
End if
len_name=len(name_left_str);// 得到 name_left_str 的字符长度
Html=mid(html,begin_name+len_name);//得到 html 的前 begin_name+len_name 个字符后的内容
end_name=pos(html,right_name_str);// 得到属性 right_name_str 字符串在后续 html 中的起始位置
Company_Name=left(html,end_name-1);//得到 html 中的公司名称
//厂家主营( business )规则的设定同厂家规则。略……
```

抓取到的属性 company_name business 的 html 其中还是存在一些冗余的 html 的，在后面将给出去除这些冗余的函数。

5 抓取器设计

抓取器是集抓取 HTML 及抓取规则的一个综合。在 PowerBuilder 中，抓取对方网页的 HTML 要依靠 Inet 对象。以下为 Inet 对象的定义及实例化[3]。

```
/*实例 inet 对象得到 url 的 html*/
n_internet html_data //定义 n_internet 对象
inet inet_object //定义 inet 对象
inet_object = Create inet //实例 Inet 对象
html_data = Create n_internet// 实例 n_internet
inet_object.GetURL(url, html_data)// 取网页的 url
string html;//定义字符串 html 用来存储的 url 源代码
html=html_data.is_data; //把源代码赋值给 html
destroy html_data;//销毁对象
destroy inet_object;//销毁对象
```

以上的代码实现了对目标 URL 的 html 的抓取。

5.1 URL 的 html 抓取

以上的实现只是一个单一 URL 的 html 的抓取，

海量的对数据的抓取之前，要把需要抓取的 URL 放入一个文件中，然后再结合上面的抓取程序循环的进行数据的抓取。

循环读取的程序及说明。

```
li_FileNum = FileOpen( file_path, LineMode!, Read!, LockRead!, Append! ) //打开目标 txt 文件
li_rtn = FileRead(li_FileNum, url)//读取 txt 文件内容
Do While li_rtn <> -100 And li_rtn <> 0 And li_rtn <> -1 //逐行循环读取 url
If interrupt = True Then //停止抓取
    MessageBox('提示',停止数据抓取')
    Return
End If
//实例对象
String html
n_internet html_data
inet inet_object
inet_object = Create inet
html_data = Create  n_internet
//选择抓取网页的编码
code_type = String(ddlb_1.Text)
Int i
i = inet_object.GetURL(url,  html_data) //抓取 url
If i = -1 or i=-2 or i=-4 Then //目标 url 不合法则跳过
    li_rtn = FileRead(li_FileNum, url)
    Continue
End If
html = html_data.is_data
Destroy  inet_object
Destroy html_data
//以下为抓取规则，省略……
.....
li_rtn = FileRead(li_FileNum, url) //循环读取 url
Loop
FileClose( li_FileNum ) //关闭文件
```

以上的程序实现了从 TXT 文件中逐行读取 URL，并根据规则进行数据的抓取。然后筛选出有用的 URL。还有一种方法就是把 URL 读入到数据库中，通过 SQL 来筛选数据，然后利用 inet 对象抓取 URL 的 HTML。在抓取中遇到的一些非法的 URL(不合法或者不正确)，inet 对象会有相应的返回值，可根据这些返回值程序做出的处理跳过这个 URL，执行下一个 URL。

5.2 除冗余数据

抓取到的内容中会存在一些冗余的 HTML、过多的空格、回车符号、制表符号、换行符号，所以针对于冗余数据的整理，主要针对这两方面[3]。下面给出程序的完整实现。

```
/*功能：去除冗余 HTML
*函数名字:f_striphtml
*形式参数:html(需要处理的字符串)
*/
Long  Left,Right
Left = Pos(html,"<");//左尖括号的起始位置
Right = Pos(html,">",Left) //在左尖括号后续右尖括号的位置
Do While Left > 0 And Right > 0
html = Replace(html,Left,Right - Left + 1,'') //把左右尖括号及内容替换为空
Left = Pos(html,"<")
Right = Pos(html,">",Left)
Loop
Return html //返回替换后的数据
/*功能:去除，回车符，换行符，制表符
*f_stripspace
*形式参数: html(需要处理的字符串)
* char(9)代表的是回车,char(10)制表 ,char(13)换行
*/
do while pos(html,char(9))>0
html=replace(html,pos(html,char(9)),1,'')
loop
do while pos(html,char(10))>0
html=replace(html,pos(html,char(10)),1,'')
loop
```

```
do while pos(html,char(13))>0
 html=replace(html,pos(html,char(13)),1,'')
loop
return html //返回替换后的数据
```

以上的代码实现了基本的一些冗余代码的去除，很多时候抓到的数据本身就存在问题，解决问题的办法就是数据库的二次数据整理。

6 数据入库

PowerBuilder 的强大之处就在于对数据库的开发，只要编写相应的 SQL 语句，就可以把数据存入数据库，加之采用了关系型数据库 Microsoft SQL Server2000，在数据的整理上更是便于维护。

6.1 建立库

针对本文的实例，下面给出了相对应的代码。

```
--建立数据库(yp)
Create database web_data
Go
Use yp
go
Create table yellowpage
(
Num int primary key  identity(1,1),/* num 字段每增加一条记录，序列自动增加 1 */
url char(1000),/* 抓取的网址 */
Name varchar(1000),/* 厂家名称 */
Business varchar(1000),/* 主营 */
Address varchar(1000),/* 公司地址 *、
Phone varchar(100) /* 电话 */
)
Go
```

以上是建立数据库和表的代码，大部分字段都采用了 varchar 这种数据类型，原因是这种类型会自动的去除得到字符左右两边的空格[4]。

6.2 数据入库

用 PowerBuilder 对得到数据写入数据库的过程。

```
INSERT INTO yellowpage
            (  url
            name,
            address,
            business,
            phone )
  VALUES ( :url,
            :name,
            :address,
          :business
            :phone)  ;
if sqlca.sqlcode=0 then   //事务处理
commit using sqlca; //执行数据入库
else
messagebox('错误',sqlca.sqlerrtext)//运行错误提示
rollback using sqlca;//回滚 sql，执行失败
end if
```

代码中在 values 后面插入的是相应的变量，随后的是一个事务的运行，检测数据最终是否完整的插入了数据库。这点很重要，这个可以检测数据的完整性。

6.3 数据的二次整理

数据在导入了数据库之后，很多的数据仍需要细心的去挖掘、发现、修改。主要的数据问题有两方面。

6.3.1 处理冗余的 HTML

编写的用 SQL 函数去除冗余 HTML[2][5]。

```
--函数:striphtml
--功能：去除冗余 html
create function rep_html(@str varchar(500))
returns varchar(500)
as
begin
declare @sHTML nvarchar(100)
set @sHTML =@str
declare @i int
set @i = patindex('%<%>%', @sHTML)
while @i > 0 /*循环读取*/
begin
/*删除 “<”和 “>”之间的内容*/
  set@sHTML=stuff(@sHTML,@i,charindex('>', @sHTML, @i) - @i + 1, '')
```

```
    set @i = patindex('%<%>%', @sHTML)
  end
  return @sHTML /*返回替换后的字符串*/
  end
```

6.3.2 整理数据的完整性。

如果本身抓取到的数据内容上存在问题，比如电话号码中存在汉字,这是不可以的，而且有的电话号码的长度不够(可能是网站本身的数据存在错误)。针对这类问题，需要编写相应的 SQL 来清除。本程序中最为关键的就是电话中存在相应的汉字，因为汉字是由双字节组成的，而数值则占用单字节。

```
--函数名： delChinese
--功能：删除汉字
原代码略……。
```

7 结论

本文介绍了基于样本实例的 WEB 数据抓取的方法，实现了对 WEB 信息高效的数据抓取，同时利用 SQLserver 数据库对 WEB 信息进行抽取和存储，在实际中对用户非常有用。但如何利用尽量少的样本，取得比较好的效率；如何有效的管理和应用众多规则；如何提高系统在特征上的选取、规则的生成、数据抓取的智能化程度；如何更好的从 HTML 网页中或者其它的信息实体中得到更多有效数据，仍需要进一步的研究。

参考文献

1 罗敖生.基于 ASP 实现网上数据的自动抓取.江苏广播电视大学学报, 2004,6:60－61.

2 张勇毅,姚华.PowerBuilder+SQL Server 数据库应用系统开发与实例.北京:人民邮电出版社, 2007: 260－275.

3 余金山,冯星红,李肖.PowerBuilder 10 参考手册.北京:科学出版社, 2005:180－196.

4 张惠颖,曲著伟.基于子树匹配的交互式 WEB 数据抽取方法.计算机工程, 2006,9:78－80.

5 李玉波.韩光林.SQL Server 完全自学手册.北京:机械