网页源码抓取方法的设计与实现①

Design and Implementation of Capturing the Source Code from Web Pages

王 伟 (西安财经学院 信息与教育技术中心 陕西 西安 710061)

摘 要: 分析了嵌入式框架在实际应用中存在的不足，设计并实现了一种基于正则表达式的抓取网页源码方法。并在 asp.net 环境下，用 C#语言实现了该方法。这种方法有生成页面简单、快速的优点，提高了网页的可读性、安全性，生成的页面也更利于设计者使用。

关键词: 网页 源码 正则表达式 抓取 嵌入式框架

1 引言

在网页设计中，嵌入式框架 iframe[1-3]是经常用到的一种技术,通过 iframe 可以在一个网页中引用另外一个网页的内容。通过这种网页的嵌入技术，设计者只需采用简单的网页标记<iframe> </iframe>，就可以达到栏目丰富化、功能多样化、信息实时化等网页设计要求。但是，iframe 在带给设计者如此多好处的同时，也存在其不足之处。

2 iframe存在的不足

随着信息技术的飞速发展，信息安全问题是不容忽视的。目前，大量的有害信息，如木马、病毒、非法网站等都利用 iframe 进行传播[4,5]。用户在访问正常网站的同时，无法了解当前浏览的网站采用了多少个 iframe，这些引用的信息是从哪里链接过来的用户更加难以判断，一不小心就会“中招”。所以，大多数的浏览器对 iframe 的使用都有着不同程度的限制,多数安全防护产品也禁止用户浏览 iframe 的内容。这样，对很多采用 iframe 技术来设计网页的设计者来说，用户根本无法浏览完整的内容。

另外，即使用户的浏览器、安全防护软件等均允许浏览 iframe，留给设计者的还有一个需要解决的问题，那就是:“如何将引用的网页，按需合理地安排在自己的网页中？”。这个问题也是令设计者非常头痛的。比如，在一个网页中引用某个新闻网的新闻条目时，如何去掉被引用页面中的标题、格式、栏目、广告等等“垃圾”信息？Web 页面不像传统的文本那样整齐、干净，其中包含了大量噪声[6]。一般来说，采用 iframe 需要被引用网站提供给引用者一个“专门的网页”，待设计网页的风格、栏目、排版等内容要根据这个“专门的网页”来确定，这样的解决办法对设计者来说显然有很大的局限性。

3 解决思路

采用抓取网页源码的技术[7]，分析被引用页面的浏览器输出源码，找出需要引用部分代码的共性和差异。将相同部分的代码转换为正则表达式的常量，将不同部分的代码转换为正则表达式的变量，从而通过正则表达式实现对文字内容的过滤，并将过滤后的内容进行重整后输出[8]。用这样的方法，不但可以灵活地选择网页内容，而且不用担心浏览器、安全防护软件等对 iframe 的限制。用户通过浏览器所看到的内容，是经过筛选后的重新生成的静态页面，这样也保证了客户端的安全性。

4 正则表达式

正则表达式(Regular Expression)是指一个用来描述或者匹配一系列符合某个句法规则的字符串的单个字符串。正规表达式最初出现于理论计算机科学的自动控制理论和形式语言理论中。以后其它一些领域

① 基金项目:国家自然科学基金(10771129);陕西省自然科学基金(SJ08ZP14)
收稿时间:2008-10-29

尤其是数学上对它进行了扩展研究,数学上使用正则集合的数学符号来描述此模型,目前正则表达式被广泛地使用于各种 Unix 系统及多种计算机程序设计语言中。正则表达式实际上是一种生成字符串的字符串[9]。利用正则表达式对引用页面进行“清洗”，可以非常准确地匹配网页代码的特征。

5 实现方法

众所周知，C#语言提供的 WebClient 类[10]可以从 URI 标识的任何本地、Intranet 或 Internet 资源发送数据，以及可以从这些资源接收数据[11]。本文正是通过 ASP.NET 环境的 C#语言编程，利用 C#所提供的 WebClient 类，实现对指定网页的源码抓取，并以 JavaScript 的格式输出[12,13]，在需要插入引用页面的地方，只需加入相应的 Javascript 脚本即可实现调用。

下面以 http://news.xaufe.edu.cn/中的“新闻中心”栏目为例，给出一个网页抓取、页面调用的实现方法步骤。

Step 1. 分析源码，得出正则表达式。本例中，每条新闻除链接地址、新闻标题、发表时间这三部分外，代码完全相同。定义变量 id、text、date 分别表示这三部分内容。

Step 2. 在 ASP.NET 环境下新建一个 ASP.NET Web 应用程序项目 WebCapture，在该项目下建立 GetNews.aspx 文件。GetNews.aspx.cs 的部分源代码如下：

```
using System;                    //引用命名空间
//引用所需的命名空间
……
//引用正则表达式所需的命名空间
using System.Text.RegularExpressions;
namespace WebCapture
{
public class GetNews : System.Web.UI.Page
{
private void Page_Load(object sender,
System.EventArgs e)
{
  GetWebCode();
}
//获取网页源码
public void GetWebCode()
{
//定义要引用的网页地址
string url=@"http://news.xaufe.edu.cn/";
//定义用于存放输出结果的字符串
string result="";
try{
//创建一个 WebClient 实例
WebClient wb=new WebClient();
//从资源下载数据并返回字节数组
byte[] pagedata=wb.DownloadData(@url);
//将获得的源码进行转换
result=RebulidNews(Encoding.Default.
GetString(pagedata));
}
catch(Exception ex)//捕捉异常
{
result=ex.Message;
}
Response.Write(result);//输出 JS 代码
}
//利用正则表达式对抓取的代码进行处理
public string RebulidNews(string pagedata)
{
//定义输出字符串
string
  result=@"document.writeln(""<table>"");";
//定义正则表达式
//变量 id、text、date 分别表示
//链接地址、新闻标题、发表时间
Regex re = new Regex(@"<a
href=""http://news.
****.edu.cn/(?<id>.*?)""[^>]*target=""_blank""
class=""a2"">(?<text>.*?)</a>(?<date>.*?)</t
d>", RegexOptions.IgnoreCase);
//定义循环变量
//i<10 表示最多抓取 10 条新闻
int i=0;
for(Matchm=re.Match(pagedata);m.Success
&& i<10;m=m.NextMatch())
{
```

```
    //定义满足条件的新闻格式
    result+=@"document.writeln(""<tr><td>?
<a href=\ ""http://news.****.edu.cn/";
    result+=@m.Groups["id"].Value.Trim();
    result+=@"\"" target=\""_blank\"">";
    result+=@m.Groups["text"].Value.Trim().Re
place("          color='#6495ED'>","").Replace("
size=2>","");
    result+=@"</a><font color='#727272'>";
    result+=@m.Groups["date"].Value.Trim();
    result+=@"</font></td></tr>"");";
    i++;
    }
    result+=@"document.writeln(""</table>"");
  ";
    return result;//返回结果字符串
    }
  ……
  }
}
```

Step 3. 将上面的 GetNews.aspx 进行编译后，建立一个静态页面 Test.html 进行测试，源码如下：

```
<html>
……
<body>
……
<scripttype="text/javascript"src="GetNews.asp
  x"> </sc- ript>
……
</body>
</html>
```

Setp 4. 在浏览器中运行 Test.html，满足条件的新闻标题以 Javascript 格式输出为静态页面，设计者可根据页面需要，自由定义输出文字的样式。

6 结束语

这种方法替代 iframe 的传统调用方式，避免了各种软件的限制，具有很好的安全性。利用正则表达式进行循环替换，在大量的数据替换时速度非常快，生成的代码根据设计者需要可随意定义样式，具有很好的灵活性。对于不同的引用网页，只需定义不同的正则表达式即可实现抓取，结果以 Javascript 格式输出，对于任何格式的网页文件都可以直接调用，具有很强的通用性。利用该方法的基本原理，可满足页面调用、搜索引擎、新闻存档等相关需求，具有很好的可扩展性。

参考文献

1 阳富民,李俊,周正勇,等.嵌入式浏览器的设计与实现.计算机工程与科学, 2003,25(4):39－41.

2 胡贯荣,阳富民,周正勇.一种嵌入式浏览器交互模型.计算机工程与科学, 2004,26(12): 12－14.

3 姚琼,孙鹏,胡琳琳,等.一种基于 iframe 的嵌入式浏览器动态数据处理策略与实现.微计算机应用, 2008,29,90:17－21.

4 马贞辉.微软 IFRAME 溢出攻击原理分析.信息网络安全, 2004,12:42－43.

5 刘天颖,张彬,石立桩,等.Web 服务器中恶意 Iframe 插入的防范.农业网络信息, 2008,8:132－133.

6 周源远,王继成,等.WEB 页面清洗技术的研究与实现.计算机工程, 2002,9:48－50.

7 刘瑞虹,曹东启.基于 Intranet 的 Web 信息获取方法和实现.计算机科学, 1999,1:23－27.

8 张亮,周琪云.基于 ASP.NET 的自定义规则动态页面自动生成系统的设计与实现.计算机与现代化, 2008,7:123－126.

9 陆虎,宋余庆,等.一种基于正则表达式匹配的协议分析异常检测方法.计算机应用与软件, 2008,3:89－90.

10 Williams M. Visual C#.NET 技术内幕.北京:清华大学出版社, 2002.

11 张黎明,蒋泽军,等.基于 C#的 Web 应用程序在线升级.微电子学与计算机, 2006,(4):101－103.

12 Ferracegiati FC, Glynn J. .NET 数据服务 C#高级编程.毛尧飞译.北京:清华大学出版社, 2002.

13 邵鹏鸣.C#面向对象程序设计.北京:清华大学出版社, 2008.