# 正则表达式的研究及在 Web 中的应用

唐惠丽,郑小妹
(上海师范大学 计算机系,上海 200234)

**摘 要**:文中利用正则表达式能够完成对字符串的匹配,替换的功能,以抽取 HTML 文档中的信息为例,介绍了正则表达式的理论和在 Web 中的不同使用方法。以达到从大量数据中挖掘出某些特定信息的目的。其原因是正则表达式是代表具有特殊意义字符的字符串,它能实现将某个字符模式与所预先定义的字符串模式进行匹配,从而抽取出所需的字符串。所以正则表达式使字符串的模式匹配变得更加容易。对于处理字符串的应用程序而言,它起着很重要的作用,应用十分广泛,是一个不可缺少的工具。

**关键词**:正则表达式;模式匹配;. NET
**中图分类号**:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)02-0082-03
doi:10.3969/j.jssn.1673-629X.2013.02.018

# Research of Regular Expressions and Application in Web

TANG Hui-li,ZHENG Xiao-mei
(Department of Computer Science,Shanghai Normal University,Shanghai 200234,China)

**Abstract**:Use regular expressions to complete the string match,replace function,taking extract information in a HTML document for example,introduced the theory of regular expressions and different method applied in Web,reaching the purpose to mine specific information from massive data. Because the regular expression represents the character string has special significance,it can deliver a character-mode and match the search string,and find the information you need. So the regular expressions make pattern matching of string easy. It plays an important role for application program to deal with string,is an indispensable tool.

**Key words**:regular expression;pattern matching;. NET

## 0 引 言

互联网的快速发展使得人们能够在全球范围内共享信息。目前 Web 上的数据大部分都是以 HTML 形式出现的,所以人们能够通过浏览器浏览信息。但 HTML 存在不清晰的语义信息、模式不明确等缺点,这给基于 HTML 结构的信息抽取带来了困难。但是利用正则表达式的匹配、替换和提取等功能使得 Web 信息抽取过程变得容易。目前正则表达式在数据搜集、页面优化、信息抽取等整个 Web 信息抽取的过程中得到应用。对于处理字符串的许多应用程序而言,正则表达式是不可缺少的工具[1]。

## 1 起源与发展

正则表达式最早是由数学家 Stephen Kleene 提出的,设计于五十年代。正则表达式最初用于描述“正则集”,它们是一些神经生理学家研究的模式。在最近的五十年中,正则表达式逐渐从模糊深奥的数学概念发展为在各类工具和软件包中应用的主要功能。尽管数十年来很多 UNIX 工具都支持正则表达式,但仅仅是近十年来,它才在大部分 Windows 开发者工具包中得到体现。在 Microsoft Visual Basic 6 或 Microsoft VBScript 中,情况不够理想,正则表达式仍难以使用。但由于. NET Framework 的出现,正则表达式的支持发展到极点,所有 Microsoft 开发者和所有 . NET 语言都可以使用正则表达式[2]。

## 2 正则表达式简述

正则表达式的英文是 regular expression,意思是符合某种规则的表达式,可以将其理解为一种对文字进行模糊匹配的语言[3]。正则表达式用一些特殊的符号(称为元字符)来代表具有某种特征的一组字符以及指定匹配的次数,含有元字符的文本不再表示某一具体的文本内容,而是形成了一种文本模式,可以匹配符合该模式的所有文本串[4]。正则表达式可以快速地分析大量的文本以找到特定的字符模式;提取、编辑、替

**收稿日期**:2012-05-18;**修回日期**:2012-08-24
**基金项目**:上海市教育科技创新项目(12YZ074)
**作者简介**:唐惠丽(1975-),女,黑龙江人,讲师,硕士,从事 WEB 技术研究与开发。

换或删除文本子字符串[5]。

## 3　正则表达式的功能

在程序语言中引入正则表达式,可以完成以下功能[6]。

1)测试字符串的某个模式, 验证用户输入的有效性。

2)在文本中使用一个正则表达式来标识某些特定的字符,然后对其进行删除、替换等操作。

3)利用正则表达式搜索字符串的模式,然后从字符串中提取一个子字符串。

## 4　正则表达式的应用

文中介绍正则表达式在一个后台管理系统中的应用。

1)需求分析。

许多购物网站,在每逢节日期间要进行产品促销。而且在节日期间要更换不同的页面,以促销不同类别的商品。所以做一个关于节日专题的模块在后台管理系统中是非常必要的。功能如下:

■首先创建关于某个专题的多个模版便于替换;

■在模版里加上要促销的商品。

2)模块的设计。

图 1 为专题创建过程。

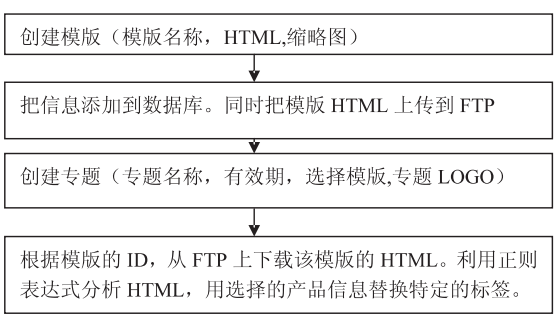


图 1　专题创建过程

3)实现过程。

这部分详细介绍正则表达式不同的使用。

■利用正则表达式实现页面中信息的验证。Web 页面验证功能用正则表达式的控件验证功能来实现。

■除了 ASP. NET 验证控件,在. NET 中使用正则表达式的情况就是使用 System. Text. RegularExpressions 命名空间的类[7]。

(一)创建主题需要的模版页面。

表 1 为模版 UI 设计。

本页面使用正则表达式的控件验证功能:正则表达式的验证控件[8](RegularExpressionValidator)是一种较为灵活的验证方式,可以使用正则表达式的强大功能,实现对复杂字符串验证功能。下面是RegularExpressionValidator 的用法。

表 1　模版 UI 设计

| | | |
|---|---|---|
| 模版名称 | 文本框(六一儿童节,母亲节,端午节……) | 最多 64 个字 |
| 上传 HTML | 文本框 HTML 内容:<br><htmlxmlns = \" http://www. w3. org/1999/xhtml\" ><head><br><title>Untitled Page</title></head><body><h1>模版测试页面</h1><mt:text id = \" text1 \" >ab</mt:text><br /><mt:text id = \" text2 \" >cd</mt:text><br /><mt:link id = \" link1 \" linkurl = \" http://www. myetone. com/\" >测试链接</mt:link><br /><mt:link id = \" link2 \" linkurl = \" http://zone. tone. com/\" >linkurl</mt:link><br /><mt:image id = \" image1 \" imageurl = \" http://www. tone. com/images/logo. gif \" ></mt:image></body></html>\" | |
| 上传模版缩略图 | 选择要上传的模版图片 | |

例如:把模版的名字的长度限制在 64 个字以内。<asp:regularexpressionvalidator id = " RegularExpressionValidator1" runat = " server" CssClass = " little" ControlToValidate = " templatename " ErrorMessage = " RegularExpressionValidator" ValidationExpression = " ^(. | \n){0,64} $ " Display = dynamic>内容不超过 64 个字 * </asp:regularexpressionvalidator>

这个页面的功能是把页面上的模版主题和模版的缩略图的存放地址添加到数据库的模版表中。同时把模版的 HTML 上传到 FTP。其中编写的 HTML 中,标签<mt:product id = \"text1 \" runat = server></mt:product>表示该模版要添加一个商品。

(二)创建专题页面。

表 2 为专题页面。

表 2　专题页面

| | | |
|---|---|---|
| 专题名称 | 文本框(六一儿童节,母亲节,端午节……) | 最多 64 个字 |
| 有效期 | 文本框(例如 10 天,20 天……) | |
| 选择模版 | 单选按钮,为该主题选择一个模版 | |
| 专题 LOGO | 上传一个专题 LOGO 的图片 | |
| 下一步 | 按钮链接到下一步页面 | |

这个页面的功能为:把页面信息添加到数据库中。选择模版,获得模版的 ID。同时把 templateID 传到“下一步”页面。

```
int templateID = 0;
if( Request. Form[ " RadioName" ] ! = null)
{
templateID = int. Parse( Request. Form[ " RadioName" ]);
}
```

“下一步”页面见表 3。

表 3 “下一步”页面

| | | |
|---|---|---|
| 选择产品 | 多选按钮,可以选择多个产品在专题页面上显示 | |
| 发送 | 点击这个按钮提交新的 HTML | |

根据 templateID,从 FTP 上获得该模版的 HTML。然后分析 HTML,利用正则表达式找出标签<mt:product></mt:product>部分,然后用选择的产品信息来代替该部分标签。形成新的 HTML,然后上传到 FTP。从而生成专题页面。

```
int templateID = Util.GetIntParam(this.Request, "templateid");
string mb,str3, str2;
int start,stop;
//输入用户名和密码,登陆 ftp
string host = Config.FTPHost;
string ftpuser = Config.FTPUser;
string ftppassword = Config.FTPPassword;
//根据 templateid 得到模板的 html
String path1 = Config.TemplatePath + "Template" + templateID.ToString();
string str = templateID.ToString();
mb = UpdateFiles.UFileInfo.GetHtmlfile(host, ftpuser, ftppassword, path1, str);
```

假设获得的模版的 HTML 如下:

```
string mb = "<html xmlns=\"http://www.w3.org/1999/xhtml\"><head><title>Untitled Page</title></head><body><h1>模版测试页面</h1>
<mt:product id=\"text1\"></mt:product><br />
<mt:product id=\"text2\"></mt:product><br />
<mt:product id=\"text3\"></mt:product>
<mt:link id=\"link1\" linkurl=\"http://www.tone.com/\">测试链接</mt:link><br /></body></html>";
```

然后分析该 HTML(这里<mt:product></mt:product>是模版的 HTML 中预先定义的标签),把

```
<mt:product id=\"text1\"></mt:product><br />
<mt:product id=\"text2\"></mt:product><br />
<mt:product id=\"text3\"></mt:product>
```

用选择的产品信息来代替。

```
start = mb.IndexOf("<mt:product");
stop = mb.LastIndexOf("</mt:product>");
string s = "</mt:product>";
str2 = mb.Substring(start, stop - start + s.Length)[9];
string str3 = Regex.Replace(str2, @"\<mt:product(?<text1>.*?)id=\""(?<text2>.+?)\""(?<text5>.*?)>(?<text4>.*?)\</mt:product>", "id=${text2}");//这里使用了 using System.Text.RegularExpressions 中的方法 Regex.Replace[10]
string[] sArray = Regex.Split(str5," id=",RegexOptions.IgnoreCase); //这里使用了 using System.Text.RegularExpressions 中的方法 Regex.Split[11]
string controlName="", Newupfilehtml="";
int productid=0;
foreach (string i in sArray)
{ controlName = i.ToString();
  if (Request.Form[controlName] != null)
  {
    string constr = Request.Form[controlName];
    productid = int.Parse(constr);
    ProductSystem prod1 = new ProductSystem();
    ProductData productSet1 = prod1.GetProductByID(productid);
    ProductData.TBL_PRODUCTRow row = productSet1.TBL_PRODUCT[0];
    string st = mb.Replace(str2, "");
    Newupfilehtml = st.Insert(start, table(row));
    //table(ProductData.TBL_PRODUCTRow row)是生成动态 table
    用这个 table 来代替<mt:product></mt:product>
    sub_info.Text += Newupfilehtml;
  }}
```

(三)生成的结果页面。

这个页面就是模版和产品的组合。

## 5 结束语

正则表达式是一种描述文本模式的极有效方法,它使文本模式成为字符串验证和操作的极好资源。.NET Framework 通过 System.Text.RegularExpressions 命名空间[12](特别是其中的 Regex 类)提供了对正则表达式的强大支持。

**参考文献:**


[1] 郭耀译. 正则表达式经典实例[M]. 北京:人民邮电出版社,2010.

[2] 瓦 特. 正则表达式入门经典[M]. 李松峰译. 北京:清华大学出版社,2008.

[3] 百度文库. 正则表达式[EB/OL]. 2012-03-30. http://wenku.baidu.com/view/e5ac68d426fff705cc170a96.html.

[4] 周 峰,王 征. ASP.NET 3.5 网络程序设计案例集锦[M]. 北京:水利水电出版社,2009.

[5] Goyvaerts J,Levithan S. Regular Expression Cookbook[M]. [s.l.]:O'Reilly Media,2009.

[6] 维基百科. Regular expression[EB/OL]. 2004-01. http://en.wikipedia.org/wiki/Regular_expression.

[7] Friedl J E F. Mastering Regular Expressions[M]. 3rd ed. [s.l.]:O'Reilly Media,2006.

[8] 宋鑫坤,陈万米,朱 明,等. 基于正则表达式的语音识别控制策略研究[J]. 计算机技术与发展,2010,20(2):106-109.

[9] 佘石泉,周肆清. 正则表达式在编程题自动阅卷中的应用


(下转第 88 页)

能,接下来就可以发送一个真正的数据报文了。数据报文是以 UDP 数据包的形式来发送,UDP 数据包的组成是按照 RFC 826 文档标准中的内容来实现,UDP 数据包按标准中的各个字段来填充,并计算检验和等信息。Generic RNDIS 设备数据配置流程如图 6 所示。

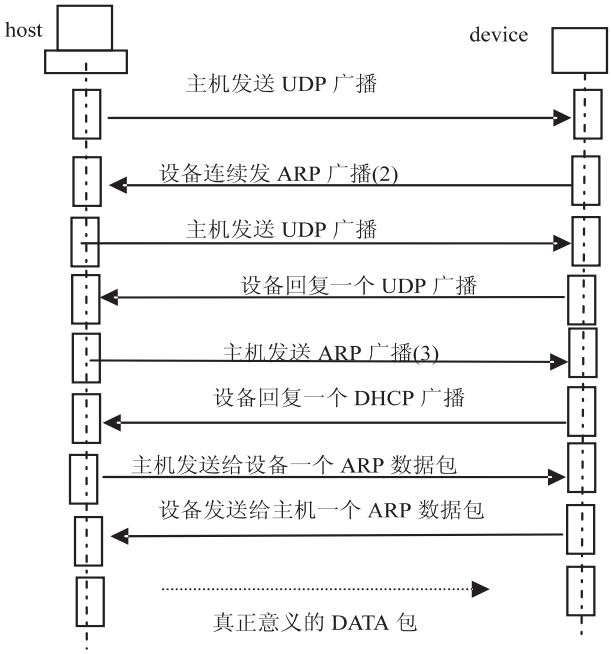


图 6 虚拟网卡配置流程

关于在 Generic RNDIS 设备和主机之间的报文格式可以参考微软的 RNDIS 协议,里面有详细的描述,而 UDP 有关的结构定义则可以参考 UDP 协议及相应的标准文档,这里将不再详细描述。只给出其方法实现。

### 2.5 实验结果

在 dopod 手机(该手机使用 Windows mobile 5.2 OS 内置了 RNDIS 功能)平台上实现了上述程序[12],在 dopod 手机上测试了程序的功能实现和性能分析,分别在手机端(RNDIS 设备)和 PC 机(USB 主机)运行客户端程序和虚拟网卡程序,分别传送 100M 字符数据和文件数据,分别存储在手机 SD 卡中,然后将这 100M 数据再修改后再回送给 PC,最后用 BUS HOUND 来捕获 USB 级的通信数据信息,用 Wire Shark 来捕获网络层的通信数据信息,经分析和实验表明该程序可保证正确无误地传递相应字符数据和文件数据。所以,在主机端,WIN XP 平台已经把 Generic RNDIS 设备识别为一个具有收发功能的以太网网卡,只要设备初始化完成就可以发送接收以太网数据。

## 3 结束语

只要设备上有对 RNDIS 功能的支持,程序在设备上面实现是可行的,PC 机上的程序会把设备虚拟成一个以太网网卡来传输数据,程序实现了微软 ActiveSyc 工具的同步功能,微软的 ActiveSyc 是基于 TCP 协议的,本程序是基于更为简单的 UDP 协议,而 UDP 协议不保证可靠传输,程序实现时要加上回复确认机制,以实现 TCP 级别的可靠传输。因此本程序比 ActiveSyc 更为简单易行,当前智能手机市场的激烈竞争和 3G 网络的发展,微软也把原来的 Windows mobile 升级为 Windows Phone 以对抗 Android 手机系统,随着 Windows Phone 手机市场份额的增加,基于 USB 接口的 RNDIS 数据通信也将有更广泛的应用。

**参考文献:**

[1] Microsoft Corporation. Remote NDIS Specification Rev1.1[S/OL]. 2002. http//www. microsoft. com.
[2] Microsoft Corporation. NDIS Specification Rev1.0 [S/OL]. 1995. http//www. microsoft. com.
[3] Microsoft Corporation. Remote NDIS (RNDIS) and Windows [S/OL]. 2009. http//msdn. microsoft. com.
[4] 虞科华. 基于 RNDIS 规范的 USB 网络设备设计[J]. 测控技术,2007,26(5):111-113.
[5] Johannes Libusb-0.1 API Documentation [S/OL]. 2010. http://www. libusb. org/.
[6] 刘 静,耿国华. 基于 USB2.0 的高速大容量数据采集存储系统[J]. 计算机技术与发展,2011,21(2):143-146.
[7] USB implementer´s forum. USB specification rev2.0[S]. CompaqIntelMicrosoftNEC,1998.
[8] 王云飞. USB 系统研究[D]. 北京:清华大学,2001.
[9] Axelson J. USB 开发大全[M]. 李鸿鹏,郑瑞霞,陈香凝译. 北京:人民邮电出版社,2011.
[10] 马 伟. 计算机 USB 系统原理及其主从机设计[M]. 北京:北京航空航天大学出版社,2004:39-59.
[11] UDP 协议[S]. IETF RFC 768,1980.
[12] 谭浩强. C 程序设计[M]. 北京:清华大学出版社,1997.

(上接第 84 页)

[J]. 计算机技术与发展,2007,17(7):244-246.
[10] MSDN LIBRARY. . NET Framework Regular Expressions [EB/OL]. 2006-06-09. http://msdn. microsoft. com/en-us/library/hs600312(VS.80). aspx.
[11] Friedl J E F. 精通正则表达式[M]. 余 晟译. 北京:电子工业出版社,2009.
[12] 沙 金. 精通正则表达式-基于. NET/ASP/PHP/JSP/JavaScript[M]. 北京:人民邮电出版社,2008.

# 正则表达式的研究及在Web中的应用



作者： 唐惠丽， 郑小妹
作者单位： 上海师范大学 计算机系,上海 200234
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(2)