C#程序设计及应用

唐大仕

dstang2000@263.net

北京大学

第9章 网络信息获取

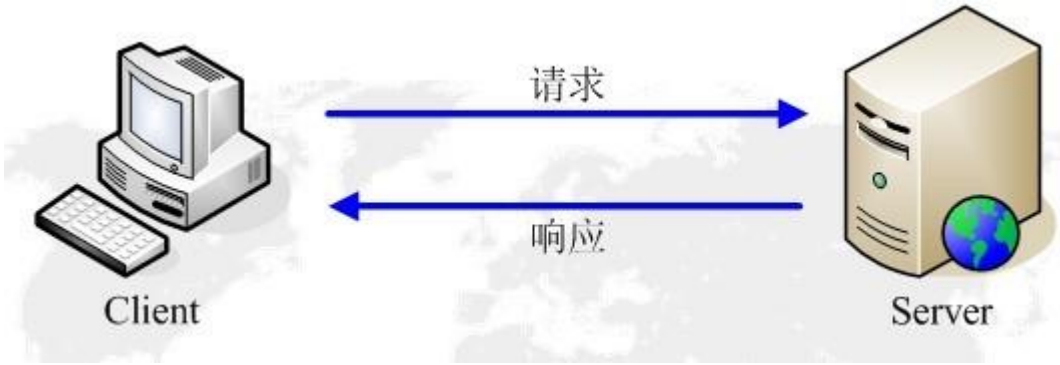
唐大仕
dstang2000@263.net
http://www.dstang.com

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>北京大学-Peking University</title>
</head>
<body>
<script language="JavaScript1.2" type="text/javascript">mmLoadMenus();</script>
<div id="all">
  <!--顶部开始-->
 <div id="top" style="float:left">
    <ul class="logo">
      <li><img src="images/index/3_02.gif" alt="北京大学" /></li>
    </ul>
```

网络信息获取

网络信息浏览

- HTTP协议
- 客户端与服务端
- Request与Response
- Stream
- Get与Post



一些查看工具

- Fiddler2
  - http://www.fiddler2.com/
  - 其他工具
    - 如 NetworkMoniter、Visual Sniffer、httpwatch、WireShark
- Chrome/FireFox等浏览器 F12
  - Chrome 中按F12，或点右键，“审查元素”
  - FireFox中安装FireBug

使用System.Web

- System.Web提供支持浏览器/服务器通讯的类和接口。
- 此命名空间包括提供有关当前 HTTP 请求的大量信息的 Request 类、管理 HTTP 到客户端的输出的 Response 类，以及提供对服务器端实用工具和进程的访问的 HttpServerUtility 对象。
- System.Web 还包括用于 Cookie 操作、文件传输、异常信息和输出缓存控制的类。

System.Net中的类

类	说　明
Cookie	提供对cookie(一种网络服务器传递给浏览器的信息)进行管理的一套方法和属性
Dns	提供简单的域名协议功能
EndPoint	表示网络地址的抽象类
FileWebRequest	与‘file://’开头的URl地址进行交互，以访问本地文件
FileWebResponse	通过‘file://’ URI地址提供对文件系统的只读访问
HttpWebRequest	授权客户向HTTP服务器发送请求
HttpWebResponse	授权客户接收HTTP服务器的回答信息
IPAddress	表示一个IP地址
IPEndPoint	表示一个IP终端(1P地址加端口号)
IPHostEntry	与带有一组别名和匹配IP地址的DNS登录建立连接
WebClient	提供向URL传送数据和从URI接收数据的通用方法
WebException	使用网络访问时产生的异常

WebClient类

- DownloadData 及 DownloadFile
  - 后来又有 DownloadString
  - UploadData 及 UploadFile
  - OpenRead 及 OpenWrite
- 示例：WebClientDownload.cs

```
string url = @"http://www.baidu.com";
WebClient client = new WebClient();
byte[] pageData = client.DownloadData(url);
string pageHtml = Encoding.Default.GetString(pageData);
Console.WriteLine(pageHtml);
```

WebRequest及WebResponse

- WebRequest myRequest =
  WebRequest.Create("http://www.contoso.com");
- WebResponse myResponse = myRequest.GetResponse();
- Stream requestStream = myRequest .GetRequestStream()
- Stream receiveStream = myWebResponse.GetResponseStream();

示例：获取网络文件内容

- DownloadString.cs
- DownloadStringAndGuessEncoding.cs

一些值得注意的地方

- Credentials: 主要指用户名、密码等
- Header： 头部信息
- Cookie： Cookie信息
- User-Agent: 用户代理（浏览器 ）
- Refer： 由哪个页面进行的访问

对获取到的内容进行处理

- 文本
  - Html
  - Xml
  - Json
- 二进制
  - 图片
  - 媒体

示例

- DownloadImages.cs 下载链接中的所有图片文件

示例：显示纸白银价格

- GoldPriceFetcher
- 要点：
  - WebClient 获取信息
    - 用了client.Headers.Add 信息
  - 用字符串函数进行解析
    - msg = msg.Substring(msg.IndexOf(tag) + tag.Length);
    - string[] words = msg.Split('|');
  - NotifyIcon组件，显示到Windows系统托盘

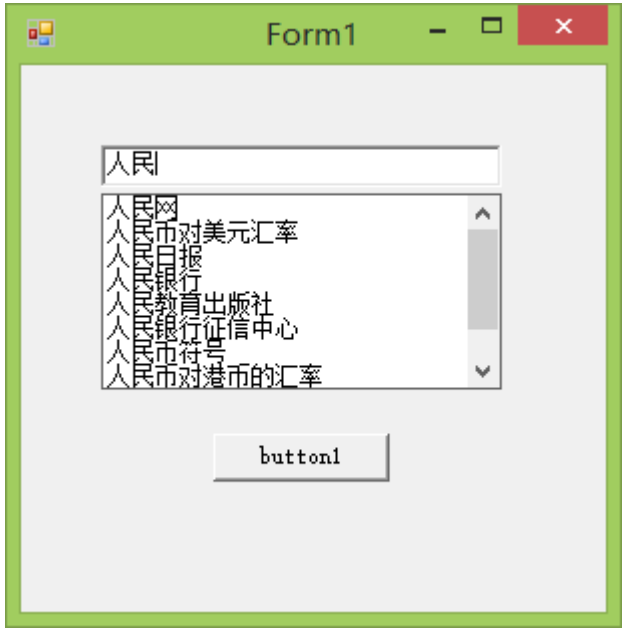
示例：网络爬虫

- SimpleWebCrawler
- 要点：
  - WebClient： 获取一个网页
  - 正则表达式： 解析其中的链接
    - @"(href|HREF|src|SRC)[ ]*=[ ]*[""'][^""'#>]+[""']“
  - Hashtable： 存入链接（其值为true或false表示是否下载过）

```
开始爬行了....
爬行http://www.cnblogs.com/dstang2000/页面！
爬行http://i.cnblogs.com/EditPosts.aspx?postid=250188页面！
爬行/css/login.css页面！
爬行/BotDetectCaptcha.ashx?get=image&amp;c=c_login_logincaptcha&amp;t=d84c10d81781434f8dba4d01b281d56f页面！
爬行http://i.cnblogs.com/EditPosts.aspx?postid=113761页面！
爬行/BotDetectCaptcha.ashx?get=sound&amp;c=c_login_logincaptcha&amp;t=974e56a0db7b49c4b371ce0e864e0419页面！
爬行http://www.cnblogs.com/ContactUs.aspx页面！
爬行/ad.aspx页面！
爬行/register.aspx?ReturnUrl=http://i.cnblogs.com/EditPosts.aspx?postid=113761页面！
爬行/about/intro.aspx页面！
爬行/about/job.aspx页面！
爬行结束
```

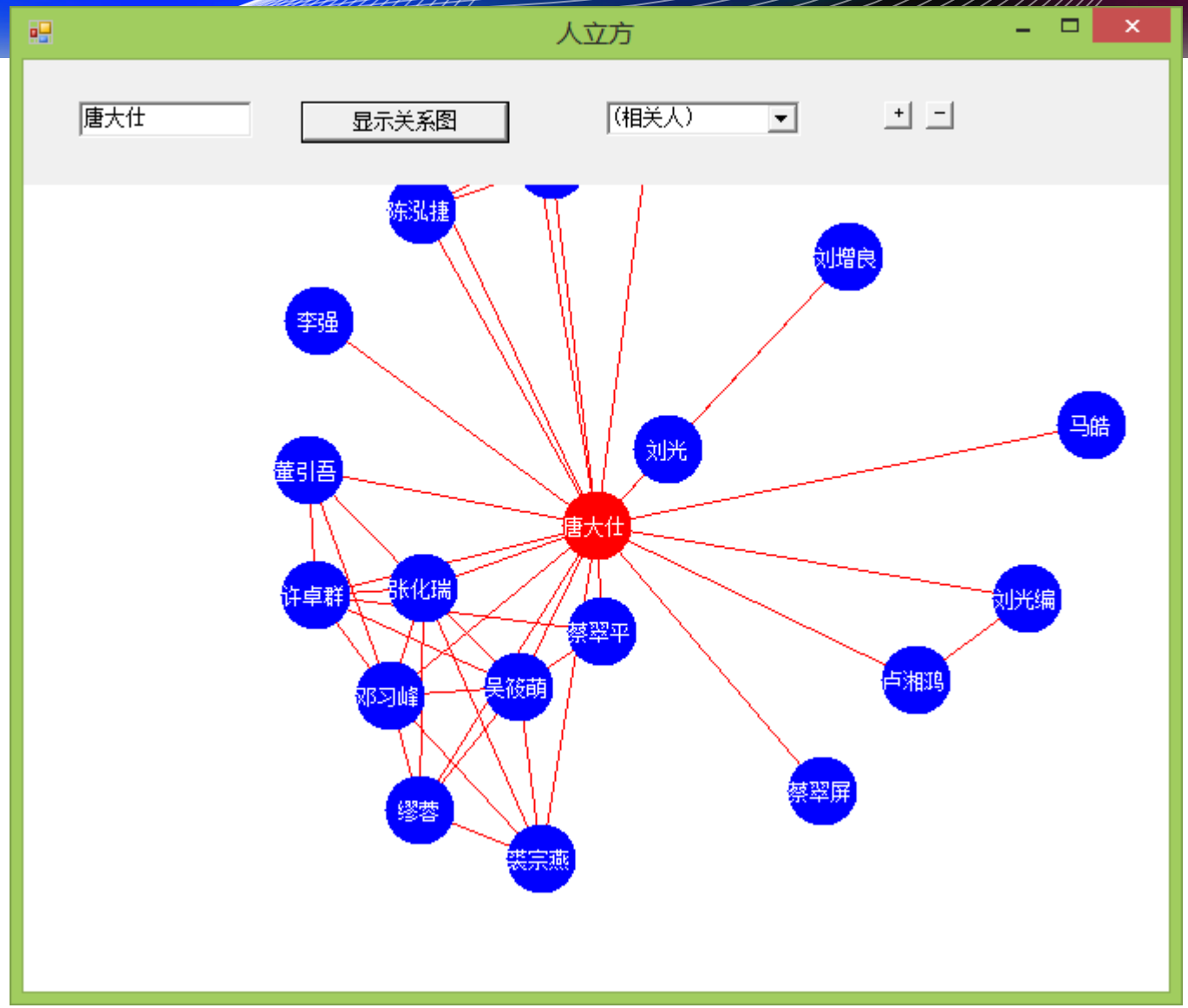
示例：显示百度的建议词

- BaiduSuggestion
- 要点
  - WebClient 下载信息
    - 用了Cookie
  - Regex.Replace:　　取出信息
  - listBox1.Items.Add：　加入到下拉框



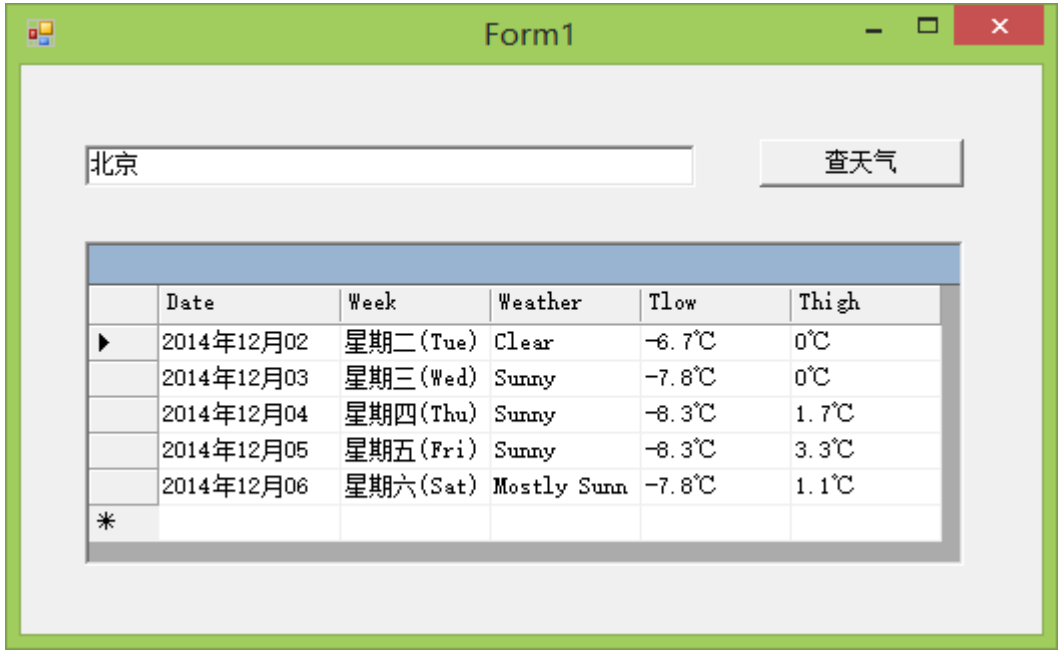
示例：人立方

- Renlifang
- 要点
  - HttpRequest： 获取内容
  - Xpath: 查询其中的信息
  - Graphics: 显示点与线



示例：获取天气信息

- YahooWeather
- 要点：
  - XmlDocument的Load方法：
    - 获取信息
  - SelectNodes及Xpath
    - 解析信息
  - DataTable
    - 内存中存放信息
  - DataGridView控件
    - 显示信息



示例：翻译字幕文件

- BaiduTranslate
- 要点：
  - WebRequst 及 Header信息： 获取翻译词
  - 文件读及写

示例：北大的IP网关

- IPGW
- 综合应用
  - WebRequest：提交数据
  - 正则表达式：解析
  - 加密、解密、文件、Xml、系统（磁盘序列号）