传媒大学媒体中心资源批量获取工具的制作

2013年10月25日 17:00:49　阅读数：3849

中国传媒大学媒体中心（ http://media.cuc.edu.cn/ ）是中国传媒大学媒体资源最集中的地方，各种电影，电视剧，音乐等等，可以说是丰富多彩。然而它有一个缺点，就是只能在线看，不能下载。这导致想把自己喜爱的视频保存下来是比较困难的。为此我课余时间进行了一个小研究，做了一个MFC的小程序，可以实现媒体中心中资源URL的提取和保存，在此记录一下自己的制作过程。

该工具主要涉及以下三个技术：

1.发送HTTP请求，获取网页的源代码

2.查找具有特定标记的字符串，并提取出来

3.数据写入xml文件

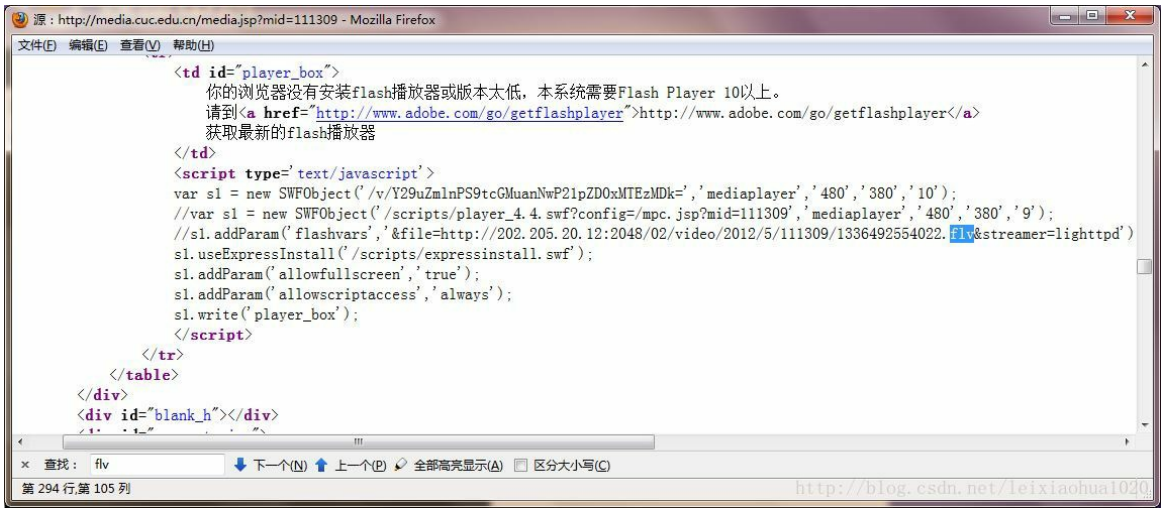
下面先看看实际情况

首先登录媒体中心，打开一个视频，截图如图所示：

查看一下网页的源代码，看看有没有视频URL。在网页里搜索了关键字“.flv”，竟然找到了。

地址就是：

http://202.205.20.12:2048/02/video/2012/5/111309/1336492554022.flv



只要把上述地址粘贴到迅雷，快车里面就可以下载视频了。

现在有一个问题，就是下载一个视频可以这样翻网页源代码找一找，但是每次这样操作有点太过麻烦了。因此需要编程实现一个小工具。当输入一个视频播放网页的地址的时候，就可以输出该视频实际的下载地址。当然，最好可以批量输入页面地址，然后批量解析视频的实际地址。

其实这个东西已经实现出来了，最终界面如下所示：

在这个工具中，贴入视频所在的网页，就可以解析出视频的标题以及视频的地址。而且下半部分还提供了批量解析的功能，输入视频ID（随后会解释）的范围，就可以探测出该范围内所有的视频资源，同时输出成XML或者TXT。

批量解析输出成XML如下所示：

```html
<URLList>
    <URL id="4">
        <name>恐惧拉斯维加斯</name>
        <link>http://202.205.20.12:2048/02/video/2008/6/恐惧拉斯维加斯/恐惧拉斯维加斯.flv</link>
        <type>vod</type>
        <protocol>http</protocol>
    </URL>
    <URL id="5">
        <name>一球成名2CD1</name>
        <link>http://202.205.20.12:2048/02/video/2008/7/一球成名2CD1/一球成名2CD1.flv</link>
        <type>vod</type>
        <protocol>http</protocol>
    </URL>
    <URL id="6">
        <name>一球成名2CD2</name>
        <link>http://202.205.20.12:2048/02/video/2008/7/一球成名2CD2/一球成名2CD2.flv</link>
        <type>vod</type>
        <protocol>http</protocol>
    </URL>
    <URL id="7">
        <name>深海寻人</name>
        <link>http://202.205.20.12:2048/02/video/2008/7/深海寻人/深海寻人.flv</link>
        <type>vod</type>
        <protocol>http</protocol>
    </URL>
    <URL id="28">
        <name>Oh_My_Friend</name>
        <link>http://202.205.20.12:2048/02/video/1970/1/28.BIGBANG_3rd_MINI_Oh_My_Friend_MV/BIGBANG_3rd_MINI_Oh_My_Friend_MV.flv</link>
        <type>vod</type>
        <protocol>http</protocol>
    </URL>
    <URL id="30">
        <name>once in a lifetime</name>
        <link>http://202.205.20.12:2048/02/video/2008/9/30.once_in_a_lifetime/once_in_a_lifetime.flv</link>
        <type>vod</type>
        <protocol>http</protocol>
    </URL>
</URLList>
```

或者TXT格式：

```plain
视频ID:4    视频名称：恐惧拉斯维加斯    视频地址：http://202.205.20.12:2048/02/video/2008/6/恐惧拉斯维加斯/恐惧拉斯维加斯.flv
视频ID:5    视频名称：一球成名2CD1    视频地址：http://202.205.20.12:2048/02/video/2008/7/一球成名2CD1/一球成名2CD1.flv
视频ID:6    视频名称：一球成名2CD2    视频地址：http://202.205.20.12:2048/02/video/2008/7/一球成名2CD2/一球成名2CD2.flv
视频ID:7    视频名称：深海寻人    视频地址：http://202.205.20.12:2048/02/video/2008/7/深海寻人/深海寻人.flv
视频ID:28    视频名称：Oh_My_Friend    视频地址：
http://202.205.20.12:2048/02/video/1970/1/28.BIGBANG_3rd_MINI_Oh_My_Friend_MV/BIGBANG_3rd_MINI_Oh_My_Friend_MV.flv
视频ID:30    视频名称：once in a lifetime    视频地址：
http://202.205.20.12:2048/02/video/2008/9/30.once_in_a_lifetime/once_in_a_lifetime.flv
```

介绍完毕。现在简要介绍一下单个视频解析url的制作过程。

第一步：发送HTTP请求，获取网页的源代码

曾经写过一篇发送HTTP请求获取网页源代码的文章： C++发送HTTP请求获取网页HTML代码

第二步：查找具有特定标记的字符串，并提取出来

曾经写过一篇查找字符串并提取出来的方法的文章： C++从文件中查找特定的字符串，并提取该字符串

在此需要综合前两篇文章的方法，实现对特定url的网页源代码的请求，以及对特定字符串的查找和提取。

首先观察一下网页源代码，发现视频地址是在一对<script>标签里：

```html
<script type='text/javascript'>
                    var s1 = new SWFObject('/v/Y29uZmlnPS9tcGMuanNwP21pZD0xMTEzMDk=','mediaplayer','480','380','10');
                    //var s1 = new SWFObject('/scripts/player_4.4.swf?config=/mpc.jsp?mid=111309','mediaplayer','480','380','9');
                    //s1.addParam('flashvars','&file=http://202.205.20.12:2048/02/video/2012/5/111309/1336492554022.flv&streamer=lighttpd');
                    s1.useExpressInstall('/scripts/expressinstall.swf');
                    s1.addParam('allowfullscreen','true');
                    s1.addParam('allowscriptaccess','always');
                    s1.write('player_box');
</script>
```

而地址开头“http://”前面是“('flashvars','&file=”，地址结尾“.flv”后面是“&streamer=lighttpd”。以这两个字符串作为标志，就能找到视频url地址。

视频标题的开头前面是“”，地址结尾“ ”。以这两个字符串作为标志，就能找到视频的标题。

下面贴上这部分的源代码

注意：本工程中使用了3个CString变量关联到3个Edit Control控件：

CString m_htmlurl;//输入页面url

CString m_videourl;//输出解析出来的视频url

CString m_videoname;//输出解析出来的视频名称

```cpp
void Csocket_http_dialogDlg::OnBnClickedOk()
{
    // TODO: 在此添加控件通知处理程序代码
    //地址----------------
    char stringsearch_before[]="('flashvars','&file=";
    char stringsearch_after[]="&streamer=lighttpd";
    //标题----------------
    char stringsearch_before1[]="<span class=text_bl>";
    char stringsearch_after1[]="</span>  ";
    //url_search_before位置，代表找到了相应字符串
    const char *mark=NULL;
    //开始和结束
    const char *stringstart=NULL;
    const char *stringend=NULL;
    //结果
    char url[200]={0};
    char vname[200]={0};

    char *content_temp=NULL;
    char *string_temp=NULL;
    //-----------------



    CInternetSession session;//建立对话
    CHttpFile *file;
    //CException *e;
    UpdateData(true);
    CString URL = m_htmlurl.GetString();
    if(URL==""){
    AfxMessageBox("网页地址为空！");
    }
    try{
        file=(CHttpFile*)session.OpenURL(URL);//打开文件
    }catch(...){
        file = 0;
    }
    if (file){
        DWORD dwStatus;
        file->QueryInfoStatusCode(dwStatus);
        if(dwStatus == HTTP_STATUS_OK){
            CString content;
            CString data;
            while (file->ReadString(data)){
                content+=data+"\r\n";
            }
            content.TrimRight();
            //MessageBox((LPCTSTR)content);
            //处理数据,数据位于content之中-------------------------
            //获得地址
            mark=strstr(content,stringsearch_before);
            if(mark==NULL){
                AfxMessageBox("没有找到地址...本软件只适用于媒体中心");
                goto end;
            }
            //注意要-1,, 此处获得string开始
            stringstart=mark+sizeof(stringsearch_before)-1;
            //注意此处获得string结束之后的1位，因此最后一位应改为\0
            stringend=strstr(stringstart,stringsearch_after);
            string_temp=url;
            for(content_temp=(char*)stringstart;content_temp!=stringend;content_temp++,string_temp++){
                *string_temp=*content_temp;
            }
            string_temp='\0';
            //获得标题--------------------------
            mark=strstr(content,stringsearch_before1);
            if(mark==NULL){
                AfxMessageBox("没有找到标题...本软件只适用于媒体中心");
                goto end;
            }
            //注意要-1,, 此处获得string开始
```

```
            stringstart=mark+sizeof(stringsearch_before1)-1;
            //注意此处获得string结束之后的1位，因此最后一位应改为\0
            stringend=strstr(stringstart,stringsearch_after1);
            //vname是最终输出
            string_temp=vname;
            for(content_temp=(char*)stringstart;content_temp!=stringend;content_temp++,string_temp++){
                *string_temp=*content_temp;
            }
            string_temp='\0';

            //--------------------------------
            m_videourl=url;
            m_videoname=vname;
            UpdateData(FALSE);

        }else{
            MessageBox("dwStatus!=HTTP_STATUS_OK");
        }
        end:
        file->Close();
        delete file;
    }
    session.Close();


}
```

第三步：数据写入xml文件

曾经写过一篇数据写入成xml的文章： TinyXML：一个优秀的C++ XML解析器

在这里就不多说了，方法类似。

文章标签： http xml 字符串 网页源代码

个人分类： 计算机网络