Subject: HttpRequest : problem with multiple redirections ?
Posted by jibe on Thu, 07 Aug 2014 09:12:11 GMT
View Forum Message <> Reply to Message

Hi,

I want to get some data about books on various websites : ISBNdb, GoogleBooks, Worldcat... This works well with all of them, but not always with Amazon.fr : with it, using the same URL, I do not get always the same content! Sometimes, it is the one I get with Firefox, and sometimes it is another. This seems to be aleatory, and any content I get, there is never any error (HttpRequest::GetError() returns 0).

Trying with wget (I am using Linux), I see that with this URL, there is 3 redirections. Could it be the problem ? And if not, what could it be ?

My code is very simple : I'm supposed to know the URL for the book (if not, I make a search on the ISBN of the book, and I always find the right URL, even on Amazon.fr - if the book is known on the site, of course !)

```
String content;
HttpRequest http;
...
http.Url(url);
content = http.Execute();
```

An url showing this problem :
http://www.amazon.fr/14-Jean-Echenoz/dp/2707322571/ref=sr_1_1/278-1397759-3160153?ie=UTF8&qid=1372075436&sr=8-1&keywords=9782707322579

Any idea ?

Subject: Re: HttpRequest : problem with multiple redirections ?
Posted by mirek on Thu, 07 Aug 2014 16:02:57 GMT
View Forum Message <> Reply to Message

Hi,

I have tried and it seems OK. Hard to say more without knowing the "BAD" content.

Generally, the problem might be on server side, perhaps they are doing A-B testing or something.

Another issue is that redirections go bad - many sites are using cookies in redirection process; U++ HttpRequest tries to emulate the browser as much as possible, but of course it does not store cookies persistently, only across redirections.

In any case, I recommend putting HttpRequest::Trace into code to trace all HTTP comms in log, then perhaps you can compare them to Firefox logs (headers and such) or perhaps Chrome logs...

Mirek

Subject: Re: HttpRequest : problem with multiple redirections ?
Posted by jibe on Fri, 08 Aug 2014 13:02:26 GMT
View Forum Message <> Reply to Message

Hi Mirek,

Thanks for your reply. The "bad" content is not so bad, it seems to be another similar (outdated ?) page about the same book. The problem is that it's not organized the same way, so I don't retrieve the data I need, or I should parse it a different way.

I will try tracing the requests and see what can be done.

What is surprising is that I get (sometimes) this bad content only when I get directly the URL. The first time I look for a book, I make a search on the site, obtain a list of books, select the right one and follow the link. This link is the URL that I store and use next times, but curiously, I get always the right page when I first search the book rather than using the stored URL!

I just wanted to have other's opinion about this : anyway, I can workaround the problem either parsing the "bad" content when I get it, or doing the search of the book first rather than use the direct url. I'll let know if I find the reason of this bad content.

Thanks for your advices.

Subject: Re: HttpRequest : problem with multiple redirections ?
Posted by mirek on Sat, 09 Aug 2014 07:51:03 GMT
View Forum Message <> Reply to Message

jibe wrote on Fri, 08 August 2014 15:02Hi Mirek,

Thanks for your reply. The "bad" content is not so bad, it seems to be another similar (outdated ?) page about the same book. The problem is that it's not organized the same way, so I don't retrieve the data I need, or I should parse it a different way.

I will try tracing the requests and see what can be done.

What is surprising is that I get (sometimes) this bad content only when I get directly the URL. The first time I look for a book, I make a search on the site, obtain a list of books, select the right one and follow the link. This link is the URL that I store and use next times, but curiously, I get always the right page when I first search the book rather than using the stored URL!

I just wanted to have other's opinion about this : anyway, I can workaround the problem either

parsing the "bad" content when I get it, or doing the search of the book first rather than use the direct url. I'll let know if I find the reason of this bad content.

Thanks for your advices.

Are you using the same HttpRequest for both? In that case, it would mean cookies are responsible... HttpRequest preserves cookies even for successive calls. You can also try if that is the issue by using "CopyCookies" (copies cookies from one HttpRequest to another).

Mirek

Subject: Re: HttpRequest : problem with multiple redirections ?
Posted by jibe on Mon, 11 Aug 2014 07:34:16 GMT
View Forum Message <> Reply to Message

Hi, Mirek,

Yes, it's that :)

I tried to remove cookies on my browser, and I obtain the "bad" page (curious site, giving an almost similar page with a very different code - all CSS classes and id are different ! - depending on the cookies...).

What is done in my application is that : the first time it looks for the book by the ISBN, obtain a list of the corresponding books (normaly only one, as 2 different books cannot have the same ISBN), then follow the link to get the page. I keep this URL in the database. It's sometime later that, if we use the link, we get the "bad" page. But in this case, I think that the cookie is no more available, as the application has been stopped...

Probably, I should keep the cookie in the database ? Well, I will see : probably a workaround will finaly be simpler.

Thank you for your help !
