Make Dynamic URLs Search Engine Friendly

by
Peter Lavin

December 22 2003

Overview

Using a database to dynamically create web pages makes for a much improved site in many ways. In the process, dynamic URLs with query strings in the format “http://www.mysite.com/main.php?category=books&subject=biography” are often created. Such URLs are not very search engine friendly. Search engines are much better at indexing static pages and do not do a good job of following hyperlinks that contain query strings. The advantages of a dynamic site are overwhelmingly obvious so what is to be done? Well you *can* have your cake and eat it too. With a little extra effort you can create a dynamic site that is easily crawled by webbots. You will even reap the additional side benefit of increased security for your site because, in the process, you will be masking how you access your database.

We will show how this can be done using an Apache web server with the module “mod_rewrite” installed. Code examples will use PHP to access a MySQL database.

Introduction

The robots used by search engines have problems with dynamic pages. You may review Google's comments on this subject at http://www.google.com/webmasters/2.html. However, dynamic URLs can be converted into static URLs so that they can be indexed. For example, the dynamic URL "http://www.mysite.com/main.php?category=books&subject=biography", can be rewritten as "http://www.mysite.com/pagebooks-biography.htm ". This article will show you how to do this and in so doing make your dynamic, database-driven website search engine friendly.

The steps involved are:

1) make sure your web server supports "mod_rewrite"
2) create a ".htaccess" file
3) upload this file to the correct directory on your server
4) test the results by typing the modified URL into the address box of your browser
5) modify your original code to write your links in a search engine friendly form.

Check Your Server

In the introduction we alluded to something called "mod_rewrite". This is a module that is usually compiled into Apache web servers and by all accounts it is fairly complex to configure. But this is the concern of server administrators and need not worry us here. "mod_rewrite" makes use of a file called ".htaccess" to perform its rewrites and creating this file is all we will need to do. ".htaccess" tells the server how to convert between dynamic and static URLs. Writing this file usually requires some knowledge of regular expressions. Like many of us, you may wish that you had greater facility with regular expressions but have never quite gotten around to learning them properly. Well don't worry - creating the ".htaccess" file will require absolutely no knowledge of regular expressions.

To ensure that your web host supports "mod_rewrite" you could send an e-mail to tech support or find out for yourself by creating the following text file:

```
<?php
      phpinfo();
?>
```

Save it as "info.php", upload it to your server and invoke it by typing "http://www.mysite/info.php" into the address box of your browser.

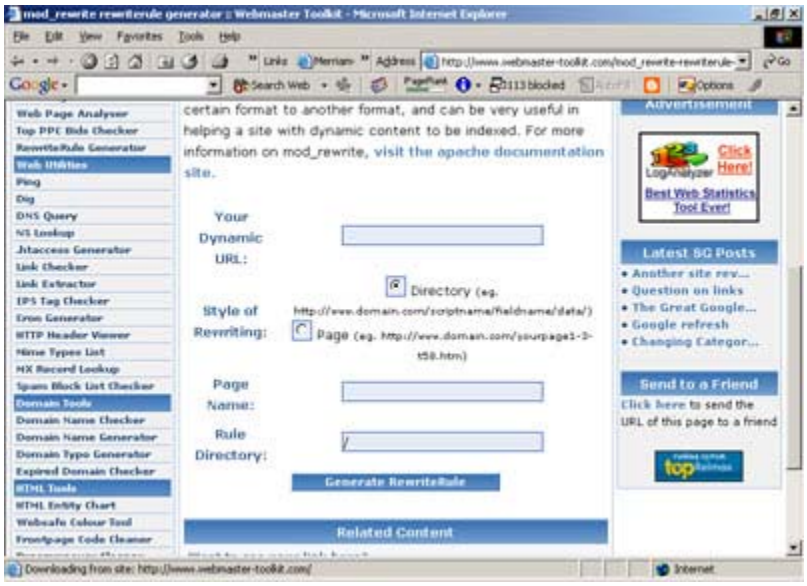
The function “phpinfo” is very useful for determining how your server is configured, for looking at environment variables, cookies and the like but right now we are only concerned with “mod_rewrite”. Look for the heading “Apache” and then “Loaded Modules”. You should find “mod_rewrite” amongst the many modules listed there. If so you are all set to begin and if not, speak to your web host.

Create a “.htaccess” File

Open the source code of a page on your site and find an URL with a query string similar to: “http://www.mysite/main.php?category=$category&subject=$subject“. If your URL is relative, rather than absolute, change it to an absolute URL and then copy it to the clipboard. You will have a much clearer understanding of what’s happening if you do it this way. Now open your browser and go to http://www.webmaster-toolkit.com/.

Take a moment to appreciate what is available here. I’m sure you’ll want to return so bookmark the site.

On the left, under the “SEO Tools” heading find and click the link “Rewrite Rule Generator”. Scroll down until your screen looks like this:



Now paste your dynamic URL into the appropriate textbox and decide if you want to represent your dynamic URL as a page or a directory. It’s up to you but make sure that you don’t create a directory name that matches an existing directory and that you don’t generate a page name that is longer than 255 characters. If you choose to generate a page name set the appropriate radio button and enter a page name into the textbox – something that best describes your page. We will be using the page name style of static URL in our examples. Click the generate button and you should have your “.htaccess” file in seconds. Copy the returned

textarea and paste it into your favourite text editor. The text returned when entering the word “type” into the page name textbox and using our example URL, “http://www.mysite/main.php?category=$category&subject=$subject“, is as follows:

```
Options +FollowSymLinks
RewriteEngine on
RewriteBase /
RewriteRule type(.*)-(.*)\.htm$ /main\.php?category=$1&subject=$2
```

This is all you will need for your “.htaccess” file. This file will intercept all page requests within the directory in which it is located and convert specified URLs with a static format to ones using a query string. That may sound like the opposite of what you want to achieve but read on and things will become clearer.

As promised, you’ve now created a “.htaccess” file without any reference to regular expressions. Don’t leave the webmaster-toolkit site just yet. Have a look at what the rewritten URL should be. Our example looks like this:

```
http://www.mysite.com/type\$category-\$subject.htm
```

Copy it into a text file, remove the backslashes that precede each dollar sign and save it to a text file called “format.txt”. You’ll need this when revising your PHP scripts.

Upload “.htaccess” to Your Server

Now that you have created the “.htaccess” file you need to upload it to your server. It must reside in the same directory as the page that invokes your dynamic URL. Likewise, the script that is invoked is also assumed to be in this same directory. In our example we use the root directory of the server.

If you are only familiar with file naming conventions under Windows the name of this file will strike you as odd. Why start a file name with a dot? On Unix/Linux systems files of this type are hidden. For this reason you need to make sure your FTP programme is set up to view hidden files. Usually this is done using a site configuration option. With my FTP programme I need to enter “-al” into a “Remote File Mask” textbox. If you can’t configure your FTP programme most operating systems have an FTP utility that is invoked by typing “ftp” at the command line. Once you’ve connected to your site, to view hidden files you need to list them using the “-al’ option. This is done by typing “ls –al”. This is an important point because you want to be able to see your file on the server, especially if you are using a graphical programme to transfer it. If you successfully transfer it and it’s not correct you won’t be able to see it to delete it. Be warned that a misconfigured “.htaccess” files can make your site inaccessible from a browser.

Make sure that you transfer your file in ASCII mode. When automatically transferring files, most FTP programmes determine the transfer mode by looking at file extensions. Under Windows this may mean that a file named ".htaccess" will transfer in binary mode. Configure your software so that files with this "extension" will upload as text files or, alternately, transfer the file manually.

Test the Results

If there are any problems we want to know about them before proceeding. There is no point in making the wrong modifications to our code! Type the rewritten URL into the address bar of your browser substituting actual literal values for the PHP variables. In our example, we know that there is a category called "books" and a subject called "biography". That would make our URL:

```
http://www.mysite.com/typebooks-biography.htm
```

If the right page is returned then you've done everything correctly.

If not here are a couple of suggestions. Make sure you have uploaded ".htaccess" into the right directory. Still have problems? Try the original, "unrewritten" URL in the address bar. If that also doesn't work then the format of your URL is incorrect. Go back, recopy it from your code and re-enter it into the rewrite rule generator.

Modify Your Code

Now that you've tested your URL and it works, modifying your code is all that remains to be done. Every instance of a dynamically created URL must be revised. Just to clarify, in our example that would be all URLs that invoke the page "main.php" using a query string with two parameters, the first named "category" and the second named "subject". Any other dynamic URLs that do not match this pattern will have to have their own separate rewrite rule. But let's keep it simple and look at a code example, again referring to our sample URL.

Original Code

Assume that our database has been opened and the result set of a query has been returned into the variable "$rs". Iterating through this result set using a "while loop" creates the dynamic URLs. This is done with the following code:

```
<?php
while($row = @ mysql_fetch_array($rs)){
      $category = $row["category"];
      $category = URLencode(htmlentities($category,ENT_QUOTES));
```

```
      $subject= $row["subject"];
      $subject = URLencode(htmlentities($subject,ENT_QUOTES));
      /*format for following HTML result
http://www.mysite.com/main.php?category=books&subject=biography
      */
      echo "<a
href=\"http://www.mysite.com/main.php?category=$category&subject=$subje
ct\">";
      echo "$row[description]</a><br>\n";
      }
?>
```

The fields “category” and “subject “ are self-explanatory. “description” is simply the text that will appear as the clickable hyperlink. You can see in this example how a query string with two parameters has been created. The first parameter is separated from the page itself by a “?” and the second by an “&”. This is the line of code that will need modification.

Revised Code

```
<?php
while($row = @ mysql_fetch_array($rs){
      $category = $row["category"];
      $category = URLencode(htmlentities($category,ENT_QUOTES));
      $subject= $row["subject"];
      $subject = URLencode(htmlentities($subject,ENT_QUOTES));
      /*format for the URL rewrite is as follows
      http://www.mysite.com/type$category-$subject.htm
      */
      echo "<a href=\"http://www.mysite.com/type$category-
$subject.htm\">";
      echo "$row[description]</a><br>\n";
      }
?>
```

Notice that a comment showing the format we are aiming for has been inserted into the code. This is the text that was saved in the file “format.txt”. It will serve as a handy reference so that mistakes are avoided. You can see that all that has been changed is the one line that actually creates the query string.

Conclusion

With minimal effort, a website can have all the advantages of being created dynamically from a database without compromising its ability to be indexed by search engines. This is achieved by using “mod_rewrite”, a “.htaccess” file and by making minor adjustments to the original scripts. The robots used by search engines have no trouble following the resulting “static” HTML page. No knowledge of configuring “mod_rewrite” was required nor any knowledge of regular expressions.

Resources

Google on dynamic pages - http://www.google.com/webmasters/2.html.
Generate a ".htaccess" file - http://www.webmaster-toolkit.com/.

About the Author

Peter Lavin runs a Web Design/Development firm in Toronto, Canada. For more information visit http://www.softcoded.com/. He may be reached at peterlavin@sympatico.ca.

My Web Pages Are Not Currently Listed

Home

All About Google

Webmaster Info

Index
Getting Listed
Not Listed
Incorrect Listing
Rank Questions
Guidelines
Facts & Fiction
SEOs
Frequent Questions

Find on this site:

Search

A. My web pages have never been included in the Google index.

Google is a mechanized search engine, which employs robots known as 'spiders' to crawl the web on a monthly basis and find sites for inclusion in the Google index. Please review the basics of submitting your site to learn more.

1. Reasons your site may not be included.

- **Your pages are dynamically generated.** We are able to index dynamically generated pages. However, because our web crawler can easily overwhelm and crash sites serving dynamic content, we limit the amount of dynamic pages we index.

- **You employ doorway pages.** Google does not encourage the use of doorway pages. We want to point users to content pages, not to doorways or splash screens.

- **Your page uses frames.** Google supports frames to the extent that it can. Frames tend to cause problems with search engines, bookmarks, emailing links and so on, because frames don't fit the conceptual model of the web (every page corresponds to a single URL). If a user's query matches the site as a whole, Google returns the frame set. If a user's query matches an individual page on the site, Google returns that page. That individual page is not displayed in a frame -- because there may be no frame set corresponding to that page.

  If you are concerned with the description of your site as seen by search engines, please read "Search Engines and Frames". It describes the use of the 'NoFrames' tag, which is used to provide alternative content. If, instead of providing alternative content, you use wording such as "This site requires the use of frames" or "Upgrade your browser", then you are excluding both search engines and people who use browsers with frames turned off. (For example, audio web browsers, such as those used in automobiles and by the visually impaired, typically do not deal with frames, which are a visual mechanism.) You can read about NoFrames in the HTML standard here: http://www.w3.org/TR/REC-html40/present/frames.html#h-16.4

2. Google does not index all of my pages. Why?

Although we index more than 3 billion web pages, we cannot guarantee that we will crawl all the pages on a particular site. However, we are always working to increase the number of pages we crawl and hope to include more pages in our index soon. For more information about how we find and include pages in our index please see http://www.google.com/technology/index.html.

If your site's internal link structure does not provide a path to all your pages, our robot may not see all the pages on your site. Google follows links from one page to the next, so pages that are not linked to by others may be missed.

Google does offer a custom site search service for a fee. If you subscribe to this service, Google will index your pages and provide visitors to your site with a full search option. However, participation in this program does not include all of your pages in the larger Google index. Nor does participation alter your rank in Google search results or increase traffic to your site. Basically, you can't buy your way into our actual search results. You can however, purchase advertising adjacent to Google results. More information about that program can be found here.

B. My web pages used to be listed and now they aren't.

1. Changes from one index to the next.

Each time we update our database of web page (about once a month), our index shifts: we find new sites, we lose some sites, and site rankings change. If your site was dropped from Google and you have not made major changes to it in the last month, we will likely pick it up again in our next index. It's possible your site was simply inaccessible when our robots tried to crawl it.

You may want to check and see if the number of other sites linking to your URL has decreased. This is the single biggest factor in determining what sites are indexed by Google, as we find most pages when our robots crawl the web and jump from page to page via hyperlinks. To find out

who links to your site, use Google's link: tool.

It's also possible your rank decreased because other sites were found and assigned a higher rank. You can be assured that no one at Google has hand adjusted the results to boost the ranking of a site. Google's order of results is automatically determined by several factors, including our PageRank algorithm. Please check out our "Why Use Google" page for more information on how this works.

2. Multiple indices

We update our index about every four weeks. If you happen to enter the same query repeatedly while we are in the process of posting the index at our various data centers around the country, it might seem like you are seeing inconsistent results from Google. What is actually happening is that you are seeing a result from an 'old' version of our index one time and a result from a 'new' version the next. Due to the size of our index, we can not simultaneously post a new index at all of our data centers, which may result in this behavior for a short period of time.

3. Other reasons

If your page does not appear at all, here are some other possible explanations.

- Your site may not have been reachable when we tried to crawl it because of network or hosting problems. When this happens, we retry multiple times, but if the site cannot be crawled, it will not be listed in our current index. If it was a transient problem, your site will likely show up in the next index, which will be completed in a few weeks.

- A technical glitch on our side may have caused us to 'miss' your site. In crawling more than 3 billion pages every few weeks, our system experiences hiccups from time to time. Again, this is a transient problem, and your site will likely show up in the next index. Please be patient with us during this period, as we are not able to modify our index by hand to add sites missed in this way.

- The contents of your page or the links pointing to your page changed significantly and you no longer have a sufficiently high PageRank, or your page had low PageRank to begin with and a small change caused you to be dropped from the Google index.

- Your page was manually removed from our index, because it did not conform with the quality standards necessary to assign accurate PageRank. We will not comment on the individual reasons a page was removed and we do not offer an exhaustive list of practices that can cause removal. However, certain actions such as cloaking, writing text that can be seen by search engines but not by users, or setting up pages/links with the sole purpose of fooling search engines may result in permanent removal from our index. If you think your site may fall into this category, you might try 'cleaning up' the page and sending a re-inclusion request to help@google.com. We do not make any guarantees about if or when we will re-include your site.