URL Re-writing Best-Practice

Posted At : 30 March 2011 11:00 | Posted By : Gareth
Related Categories: iis, SES URLs, google, coldfusion, apache

Introduction

A while back, I wrote this article about how Google recommended NOT using URL re-writing.

The Google article provoked a lot of debate and certainly managed to infuriate a lot of SEO experts.

Google's advice seemed pretty conclusive after that: don't use URL re-writing. However, Google released their updated SEO guide later, and the waters muddied. Page 8 had the following on URLs:

> If your URL contains relevant words, this provides users and search engines with more information about the page than an ID or oddly named parameter would

Now this isn't directly saying that Google indexes the words in URLs. The example is showing the URL being displayed in search results for a user to see, so it might be misconstrued. It does however re-open the argument on whether we should add keywords in URLs for SEO, and that inevitably leads to URL re-writing.

I've a number of greenfield projects starting soon, and I've been mulling over whether to use re-writing again. The unresolved issue of Google indexing words in URLs will obviously play an important role in any decision, but the usability argument is possibly more strong. There doesn't seem to be much in the way of research or best-practices, so I've written this article to start a discussion on it.

As I'll explain below, I'd like things to be as evidence-based as possible, with links to relevant authorities. Most of the tips are my own thoughts on the matter, but I've tried to back them up with reasoning.

I welcome comments with evidence or reasoning, whether they back up or argue against what I've written. That way, we can put together a strong set of guidelines which will hopefully benefit everyone.

Does Google index words in the URL?

I did X and Y happened

Now I have to declare a certain scepticism towards SEO Experts. I know I'm generalising here, but too many of them seem keen to promote their services as a *black art*. It's almost as if they're saying 'Give me lots of money, and I'll cast spells to get you to the top of Google'.

Now I'm sure the majority of their techniques work, but there's a lack of scientific evidence behind them. Too much of it seems to be anecdotal, e.g. 'I did A and went up the rankings' or 'I did B and went down the rankings'.

One of the sites I manage is for an amateur music ensemble. Between June and September last year nothing happened, as they closed down over the summer. There was no change of content on the site and no active link building or other SEO techniques. However, I monitored the keywords in Google Webmaster Tools, and they yoyo'd up and down the rankings manically over that period.

What this shows, is that there's no baseline. If your rankings are constantly moving while you're doing nothing, then how can you reliably know what affect doing A or B has?

To complicate things further, SEO experts will normally implement several different optimisation strategies at the same time, so trying to pinpoint which work and which don't is even harder.

Of course there's a reason why the rankings are constantly changing in my example above: external forces. Even if other sites aren't adding or removing links to your site, your competitors may be changing. Google is always tweaking it's algorithm, and users click different results. As you start looking into it, there's lots of external factors affecting your position before you look at what you have control of.

Black-box test

Back to the original question: do we know if Google indexes the words in URLs?

Well, knowing exactly **how much** attention Google pays to keywords in URLs is probably impossible unless they tell us. However, if we want a simple **yes or no** answer, then we might be able to test it.

I've devised a process for this. It's probably not 100% conclusive, but a decent start none the less.

Unfortunately I don't have a suitable site to test this on. If you try this yourself, please let us know the results.

1. You need a site that's relatively mature, i.e. already appears in Google's index.
2. Choose a **rare** but **real** keyword. Something like an obscure science term. For this example, I'll use *'enthalpy'*, but obviously use something different yourself. The main requirement, is that this word isn't used anywhere else in your site.
3. Create 2 new pages on your site. Each page should be of similar length, with similar mark-up, similar content and a similar number of internal links to it.
4. The first page should contain your keyword somewhere in the text. It should only appear once in the text, and not in the URL, metadata or anywhere else. The first page is a control page.
5. The second page should contain the word in the URL. Something like *www.mysite.com/my-article-on-enthalpy* would do. What's vitally important, is that the keyword doesn't appear anywhere in the html.
6. Don't tell anyone about this, and don't actively try to get external links to the pages.
7. Once a day, do a Google search using the term *'site:www.mysite.com enthalpy'*. This will restrict the search to your site, and return what pages are indexed for the keyword.
8. When the first (control) page starts being returned in results, you know that Google has re-indexed your site.
9. If the second page starts appearing in the results too, we know Google index words in URLs. If not, we can assume they probably don't.

Pitfalls of URL re-writing

Now the reason Google wrote their article discouraging URL rewriting, is that they can be easy to mess up, often without realising. If you do use URL rewriting, then you need to be aware of potential problems.

Duplicate Content

Duplicate content is the most likely side-affect of URL rewriting. To put it very simply, if the same page can be accessed by 2 different URLs, then you've duplicated content.

The first mention of this was in Google's Webmaster Guidelines. If you aren't already familiar with these guidelines – you should be!

As this article discussed, when Google detects duplicate content, it generally won't punish it in search results, unless it appears to be trying to gain an unfair advantage. If several URLs point to the same page, it will pick one to use.

However, there's a few considerations if you duplicate content:

1. You lose control over what URL appears in the results
2. Google could fail to detect it, or incorrectly view it as malicious
3. We don't know how other search engines (like Bing) deal with duplicate content

Tip 13 below actually shows a very easy way to solve the duplicate content issue, by setting the canonical URL. However, it may only work with Google, so you should still try to avoid the issue in the first place.

Parameter order

Consider the following two URLs:

- *www.mysite.com/index.cfm?param1=foo&param2=bar*
- *www.mysite.com/index.cfm?param2=bar&param1=foo*

It's obvious here, that both URLs should produce the same page. All that's changed is the parameter order, which is perfectly acceptable.

However, if URL rewriting was being used, we might have the following:

- *www.mysite.com/param1/foo/param2/bar*
- *www.mysite.com/param2/bar/param1/foo*

In this case, there's nothing to indicate to a spider that the folder structure is actually parameters. It looks like two different hierarchies that display the same content. Would Google see this as malicious?

To be safe, you would want to make sure that rewritten URLs always used the same order.

Access to underlying dynamic URL

Now in many setups, your Apache/IIS rules will convert a URL like *www.mysite.com/param1/foo/param2/bar* into *www.mysite.com/index.cfm?param1=foo&param2=bar*

If you think about it, there's nothing to stop the page being accessed directly from *www.mysite.com/index.cfm?param1=foo&param2=bar*

This means you automatically have at least two URLs that produce the same content. If the spider finds both URLs, could it see them as a malicious attempt to duplicate content?

You might think you are using only rewritten URLs, so there's no chance of a spider picking up the dynamic version. However, here are a few examples of how this could happen:

- New developers don't realise they must use the rewritten links
- An automated system like an RSS feed generator uses the dynamic version
- External sites still use dynamic links, as they got them before your site used rewritten links

For a large site, it may be impossible to guarantee only the rewritten URLs are used, so you may want to permanently redirect all dynamic URLs to be safe. Obviously, you want to avoid an infinite loop, so take care writing the rules.

Session variables

Often sites require a session variable to be passed in the URL, in a similar way to the following example:

www.mysite.com/index.cfm?pageID=23&sessionID=12345678

This is probably unlikely, but what happens if your system rewrote this URL as this?

www.mysite.com/pageID/23/sessionID/12345678

Google might index your site ok on the first pass, but what happens a week later when it re-indexes? The session ID has changed, so none of the old links work.

Also, if a link like this got into search results, then the user clicking it might hi-jack the spider's session, instead of starting their own. We've already hit a load of problems, before we start considering if Google views this as malicious duplicate content.

If I'm not making myself clear here: **Never rewrite URL session variables!**

Google is very good at identifying which dynamic URL parameters are essential (like pageID in the above example), and which it can safely ignore like sessions. If you think it's having difficulties, or you just want to double-check, then tip 16 shows how you can fix this in Google Webmaster Tools.

URL Re-writing tips

Now for all the various pitfalls, and the uncertainty about SEO, there's still one good reason you might want to use re-writing: Usability.

Lets be clear, this is re-writing for the human end user, not for search engines or any other software – they're *probably* better off with dynamic URLs.

As a side note, it's interesting to note that the URL Rewrite module with IIS 7 refers to rules as 'User-friendly URL'.

1. Think about the user first

Many people rewrite URLs for the end user, but don't really consider what the end-user wants to do.

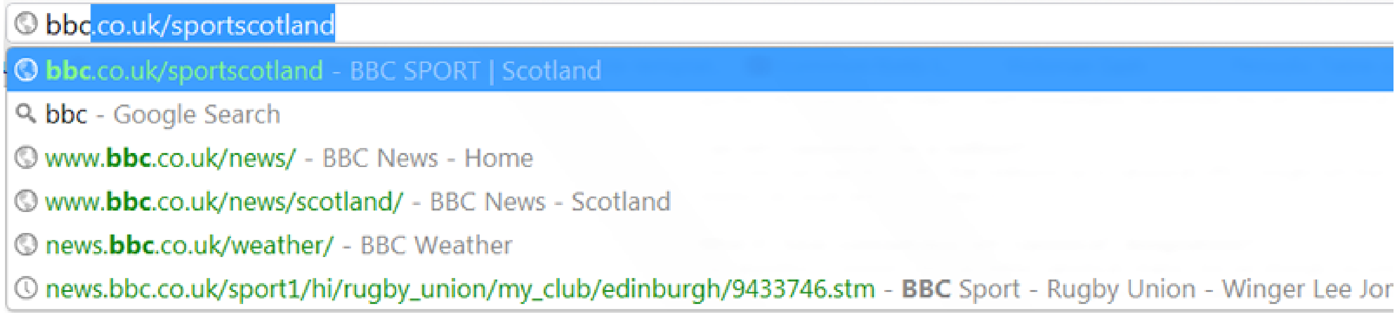
1. A URL should be easy to remember. Consider the situations when a user would have to remember a URL. This could be viewing it on a 10 second TV advert, a quick glance at a poster, being told over the phone by a friend or viewing it in someone else's browser bar. Being easy to remember saves them having to write it down or record it in some other form.
2. The URL should be easy to type. Users don't just type URLs in browsers. There's a growing number of people using smartphones and iPads etc., where they don't have a proper keyboard.
3. The URL shouldn't confuse the user. This point sounds a bit odd, but some URLs can have all sorts of funny characters in them, which might be difficult to relay over the phone etc.

2. Use <title> tags

This tip doesn't strictly relate to URL re-writing, but consider that most browsers show the page title of matching pages in your history when you start typing the URL.

If a proper page title is used, then you don't need to put so much detail in the URL. You can keep the URL short, but the full page title will still be displayed to help the user remember.

The following screen-shot demonstrates this better:



bbc.co.uk/sportscotland

bbc.co.uk/sportscotland - BBC SPORT | Scotland

bbc - Google Search

www.bbc.co.uk/news/ - BBC News - Home

www.bbc.co.uk/news/scotland/ - BBC News - Scotland

news.bbc.co.uk/weather/ - BBC Weather

news.bbc.co.uk/sport1/hi/rugby_union/my_club/edinburgh/9433746.stm - BBC Sport - Rugby Union - Winger Lee Jor

In this case the URL is short (*www.bbc.co.uk/sportscotland*), but the browser shows the full page title *"BBC SPORT | Scotland"*.

In general, you should use a unique page title for each unique page on your site. Using Google Webmaster Tools will let you know if Google has found any duplicates.

3. Keep it short

Consider the following URL:

www.mysite.com/my-great-article-on-iphones-with-reviews-and-prices-and-other-stuff

That's a lot to remember. It's also a lot to type on a normal keyboard, let alone a smartphone. The following URL would be a lot easier:

www.mysite.com/iphones

Companies like *bit.ly* and *tinyurl.com* have made a whole business model out of short URLs, so they can be very beneficial.

4. Lowercase everything

It's just easier to type if everything is lowercase. Typing an upper-case character on my phone requires an extra button press. That adds at least an extra second per (upper) character. Everything uppercase looks like you're shouting, so don't use it either.

Now sometimes a user might type the wrong case for a character. If you really want to be kind, you can put in a re-write rule that permanently re-directs the user to the lower-case version.

5. Don't use variables or IDs

You see a lot of re-written URLs in similar formats to this:

www.mysite.com/htc-android-desire-3188476

It's obvious that the *3188476* is probably the product ID, and has been appended to help do a database lookup.

The start of the URL is fine, but why should the user have to remember (and type) *3188476*. I can't remember my National Insurance number, and I've had it over 15 years - why should I have to remember your product number?

In a similar way, the URL below contains data only relevant to the system - not the user:

www.mysite.com/index.cfm/productID/3188476

As long as your URL is unique, it should only contain information relevant to the user. Let your system work out what it has to lookup.

6. No special characters, especially the underscore

I know what the underscore character is, I'm sure you do too. My 80 year old Grandmother doesn't!

You'll find a lot of non-techy people in a similar position. It's not even obvious on my keyboard, as it sits *above* the dash symbol on the same key.

Even for someone like me who knows what the underscore is, finding it on my smartphone requires an extra press.

The same applies for other special characters, spaces and punctuation marks - just do without them. They slow things down, and in some cases will end up URL encoded.

7. Use a dash as separator

So what should we use to mark word boundaries?

Well ideally, you would just have single word in your URL e.g. *www.mysite.com/iphone*, but that's not always possible.

If you do have multiple words, then using a dash (-) seems to be the generally accepted way to go. Most people know what a dash is, and (on my phone at least), you don't need an extra press to enter it.

And if you were thinking about not using a separator - don't! I once met someone with dyslexia who had difficulties reading camel case, let alone lowercase words without breaks.

8. No file extensions

A site I visit regularly has URLs of the form:

www.mysite.com/page/Home/0,,10284,00.html

Why anyone would think this abomination of a URL is good idea is beyond me. It's obvious they're using a server-side technology, but somehow think (wrongly) that turning variables into folders, and slapping '.html' on the end will make it search engine friendly.

The myth that Search Engines don't index dynamic pages has often led developers to needlessly change file extensions from cfm, php or cgi etc. to html.

However, there is an argument that users may forget the extension, or be confused or wary if it's not html. But should you use htm or html? Sometimes the final 'l' can be forgotten or missed when cutting+pasting, or they just forget if it's htm or html.

Either way, the user doesn't need any file extension, so just do without it. For that matter, do without a trailing / too - it's only another character to type.

9. What about form completed and other action pages?

By these, I mean pages you wouldn't go to directly. Suppose you have a form at *www.mysite.com/contact*, that submits to a confirmation page at *www.mysite.com/contact-sent*.

Now you wouldn't want the user to navigate directly to *www.mysite.com/contact-sent*. First, it might cause an error if you haven't anticipated this. Even if you have anticipated this, it doesn't make sense for the user to go directly to it.

This example is very simplistic, and a simple solution would be for the form to submit to itself. However, that may not be possible when using more complex forms like wizards or shopping carts.

So should you rewrite action pages? The answer is probably *'it depends'*.

Rewriting would keep things consistent, but might tempt users to enter the URL directly. In other cases, you may be required to pass cart, session or other variables in the URL, so there's probably not much point rewriting.

What do people suggest?

10. Should I rewrite image and other asset file URLs?

I don't think there's much user benefit from rewriting image URLs unless you're actually encouraging people to rip them.

The only normal reason would be **if** Google is actively indexing keywords in asset URLs. Just like above, there's no definitive answer to this, but someone could probably concoct a similar test.

11. Sub-folders

Now consider the following 2 URLs:

- *www.mysite.com/phones/android/htc/desire*
- *www.mysite.com/blog/2011/03/26/htc-desire-review*

Would these be better as something like this?

www.mysite.com/htc-desire-review

Again, I'm not sure on this one, so I'd be interested to hear people's views.

The URLs with sub-folders are longer and more complex, but they do add useful categorisation.

If you do use sub-folders, it's important to cater for advanced users who may try shortening the URL. In the first example, a user might try shortening it to *www.mysite.com/phones/android*, expecting to see a list of android phones or manufacturers.

12. Sub-domains

What if the same page can be accessed by both *www.mysite.com/iphones* and *mysite.com/iphones*?

Now the Google article on duplicate content linked above, states that it's unlikely you'll be penalised for this. However, you still lose some control. It's up to Google whether it uses the*www* sub-domain or not in search results.

It's generally best to choose a single domain to promote, although I don't think it matters much if it's the *www* sub-domain or the short version.

Once you've decided, you can set your preference in Google Webmaster Tools. However, it's probably a good idea to set a permanent re-direct rule too, to keep users and other search engines on the preferred domain.

13. Specify the Canonical URL

As mentioned above, Google have provided a great way to avoid duplicate content penalties, by specifying the canonical URL through a link tag in the **page header**.

An example would be:

```
<link rel="canonical" href="http://www.mysite.com/my-page/" />
```

Remember however, that this is a Google invention. I don't know if other search engines such as Bing use it too. For that reason, you may want to avoid some of pitfalls mentioned above.

14. I can just put everything in .htaccess files right?

Have a look at Apache's own docs on **.htaccess** files.

They recommend you don't use them, and more importantly don't even enable them. If you use shared hosting, then .htaccess files may be your only option. Otherwise, put all your rules inside a *<Directory>* section of your main configuration file instead. There are 2 reasons for this:

1. **Security**. This may or may not be an issue for you depending on your set up.
2. **Performance**. There's an unnecessary performance hit just from enabling .htaccess files.

Now Google have confirmed that **website speed** is a ranking factor, so if possible turn them off.

If you're interested in website performance, then please see my other articles here:

- **High Performance Websites**
- **How to tweak IIS to improve your coldfusion sites' performance**

15. What about IIS?

I don't know.

I've been using IIS 7.5 on most servers now, and re-write rules go in the web.config file. I've no idea if this is loaded on every request, if there's a performance hit or if there's any alternative. Any advice would be great!

16. Configure URL parameters in Google Webmaster Tools

This doesn't really apply to URL rewriting, but if any of your pages use URL parameters, you should check this out.

As mentioned above, Google can make a pretty good guess at what dynamic URL parameters are essential for navigation, and which (like session IDs) should be ignored.

However, if you think Google may be making the wrong choices, or you simply want to check it's got them correct, then you can do this in Google Webmaster Tools.

Just go to 'Site configuration' >> 'Settings' >> 'Parameter handling'

Here you can configure the site's parameters as you wish, to help avoid the chances of duplicate content penalties.

What next?

It's quite scary how much I've written about URLs here! Like most things that should be simple, the subtle complexities start coming out when you think about them in depth.

Obviously I'm keen to hear people's feedback. Most of the tips are my own opinion, and I'd prefer if there was some evidence to confirm (or even disprove) them.

I think the next step will be to work out some general re-write rules that can be added to Apache or IIS for each site. Rather than specify each re-write (which would require a lot of maintenance), most page requests should probably go to an index.cfm page.

It would then be up to coldfusion to look-up the URL, and determine what variables (such as articleID or productID) are required to run the page request.

If it works, it may well make a future article here.