The New Natural Search Spam:

Faceted Navigation

Many merchants have recently implemented attribute-based (a.k.a. faceted) navigation systems on their websites to make it easier for customers to filter, sort, navigate, and buy.

But when it comes to faceted navigation, what's good for users is actually not so good for search engines… or for your bottom line.

- By Brian Klais, EVP Search, Netconcepts

This past autumn, my wife and I were in the market for a fire pit for our backyard. As a loyal Home Depot shopper, I revved up the website and found the site navigation saved me a ton of shopping time by pre-sorting and filtering their hundreds of fire pit options on relevant attributes, like price, size, and features. I was able to quickly locate the perfect fire pit for our backyard, pick it up from the local store and fire up a lovely backyard party that very evening. But my consumer experience is not uncommon: many merchants have recently implemented similar attribute-based (a.k.a. faceted) navigation systems on their websites to make it easier for customers to filter, sort, navigate, and buy. But in this case, what's good for users is actually not so good for search engines — or for your bottom line.

In fact, in October 2008, Google pronounced the "infinite filtering" and "resulting page duplication" produced by such faceted navigation systems as negative for bots. Google urged that the duplicated pages actually be cleaned up, saying:

> *"A shopping site might allow for finding clothing items by filtering on category, price, color, brand, style, etc. The number of possible combinations of filters can grow exponentially. This can produce thousands of URLs, all finding some subset of the items sold. This may be convenient for your users, but is not so helpful for the Googlebot, which just wants to find everything - once!"*

On the surface, there is irony in such a statement from Google. After all, the big G does continually remind us in their Webmaster Guidelines to make sites built for users, not search engines. Faceted navigation clearly provides user value, and if implemented correctly can produce more granular, segmented landing pages for search engines that match the attributes searchers use in their queries. Indeed, some merchants that use our optimized landing page technology in tandem with faceted navigation systems have seen natural search sales grow by as much as 50% – 100%. However, what Google's statement really says is that while useful for consumers, faceted navigation as normally implemented actually confuses the engines by generating massive amounts of URL permutations that contain mostly duplicated content. This is bot spam.

Let's consider how "infinite filtering" becomes problematic, using my Home Depot fire pit example. The product information page for the "Catalina Creations Celestial Cauldron Fireplace" is actually duplicated at 466 URLs in Google's index, based on different combinations of parameters representing brand, category, subcategory, store and more (Example 1). In addition, the "Fire Pits" category, the primary category this product is linked from, is duplicated 99 times in Google's index (Example 2).

Example 1: 466 different URLs, same meta data   **Example 2**: 99 Different URLs, same meta data

What's happening here? Each page URL produced by the faceted navigation system contains a distinct arrangement of parameters revealing the navigational pathways googlebot crawled to arrive at essentially the same page content. Each category and subcategory has many different facets to it – variations on color, price, construction, brand, etc. – that can be listed in different orders to create more URL variations. Now consider that each product can typically be found in multiple categories, subcategories, sub-subcategories, and pagination assortments for each. Each of these acts as a multiplier, creating more and more unique URLs for content that either is exactly the same or sends the same keyword signals. Either way, the pages generated by these unique URLs contain duplicated content. We refer to this phenomenon as "crawl leak."

What crawl leak means for Home Depot – and for your business – is that natural search marketing and sales performance suffer in the following systemic ways:

1. It creates self-competition between your site's own landing pages. Home Depot may have 460,000 URLs indexed according to a "site:homedepot.com" Google query, but only a fraction are unique landing pages. Which of Home Depot's 466 fire pit pages should Google consider as the "authority" page to show their searchers?

2. It wastes your available "crawl equity." Of the millions of page requests search bots make to your site each month, the vast majority are wasted on duplicate pages that are unfit to acquire incremental searchers, while the authoritative pages get shorted.

3. This fragments your available link popularity (PageRank) between many different versions of the same/similar landing pages rather than aggregate it towards single authoritative pages. Home Depot's powerful PR 8 home page importance score quickly evaporates within 1 click. Needless to say, none of the 466 "fire pit" pages have any PageRank importance. Ouch.

4. Rubbing salt in the wound, many of the indexed landing pages are irrelevant for searchers. Some attributes, like price, are seldom used in search queries. Yet Google reports over 1,700 matching Home Depot "fire pit" pages just within the $100 - $200 price range. So not only does the price attribute contribute to the duplicate page generation problem, it's also generating a page that's accidentally optimized for a nonexistent search market (since the price range filter doubles as the backlink anchor text, and is listed prominently in the landing page title).

This systematic page duplication is like the devaluation of a currency, preventing Home Depot even from realizing the full value of related search engine optimization investments such as content enhancement, HTML improvements, and link building. Is it any wonder Home Depot is ranked 73rd for "fire pits" – far below competitors Lowe's and Walmart, and right beneath total unknown, BurningIssues.org's repurposed article about the pollution produced by fire pits? (Fortunately for Home Depot only geeks like me dig that deep, so that fact will remain our little secret.)

We can even estimate what this situation is costing Home Depot. Google reports over 400,000 searches occur each month for the search phrase "fire pit." Home Depot is likely capturing something just north of 0% of this natural search market right now. Obtaining even a modest 5% CTR on this market would produce 20,000 qualified searchers. Converting 3% at $200 AOV equals $120,000 of incremental revenue per month. And I'd wager across their website, there are another 500,000 to 1,000,000 related long tail search phrases in addition to "fire pits" that Home Depot's pages are trying to compete in, but are being similarly depressed into the bowels of Google's rankings. Make no mistake: crawl leakage on large-scale retail brand websites is a very costly problem, where brand absence in long tail keyword markets is shockingly noticeable.

Making Faceted Navigation Work for You - Not Against You

Breaking into these keyword markets requires cleaning up the search engine spam produced by faceted navigation systems like Home Depot's. This involves re-engineering the site pages to both retain the great usability benefits while maximizing the quantity and relevance of unique landing pages available. Key to this re-engineering process is deciding which URLs to suppress and which to promote for search engines.

Ideally, the way to understand the impact faceted navigation will have on the URLs is to crawl the site while it's still in the testing environment. The scale of these crawl leaks is nearly impossible to predict without a crawl. Crawling the site before launch gives a savvy team the ability to analyze how many URLs are created and what form they will take. Be warned, though: a crawl for a site that contains faceted navigation across 10,000 product SKUs is likely to number in the hundreds of thousands of URLs. (A merchant we work with was on track to flood the engines with over 1 billion URL permutations before embarking on this process. They have a manageable 130,000 unique landing pages indexed now.)

Analyzing such a high volume of URLs requires patience to tease out the patterns and determine which URLs should be made dominant. Is the same title tag duplicated across multiple pages? Is the same product ID number present in many different URLs? Look for patterns like these to find potential areas of content duplication. URL crawl frequency should also be a deciding factor, as higher frequency indicates more inbound linkages, and therefore a page more likely to be dominant. Once the dominant pages are identified for each category, subcategory and product page, you need to decide which tactics to use to suppress the non-dominant pages for natural search while continuing to enable them for human users.

The words used to create the faceted navigation scheme should be well researched during the design phase as well, because they define the keyword theme of the resulting landing pages. Home Depot's 'Outdoor' category page lists 'Fire Pits' as the link to the subcategory page. 250% more searches are conducted for "fire pit" (singular) than "fire pits" (plural). Home Depot could change this link to become more relevant for search engines and be found by more searchers. If changing this link from plural to singular does not make sense for all site users, Home Depot could generate a separate but integrated landing page dedicated to natural search users that does contain this change.

Furthermore, faceted filter links that are not beneficial to natural search performance (such as price) should be similarly evaluated, and tactics put in place to make these landing pages uncrawlable for bots. This can be accomplished using a combination of nofollow tags, meta noindex page tags and/or robots.txt commands. However it can be quite tedious work to change each page. Often algorithmic rules need to be developed by IT to save time deploying – and then monitoring – these changes.

More advanced techniques can be deployed to automatically aggregate the many parameter-filled URL permutations into singular, authoritative version of each category, subcategory, and product level page URL. For Home Depot, this could mean directing all requests for the "Catalina Creations Cauldron Fireplace" product to resolve on the product URL as linked from the "Outdoors > Outdoor Living > Outdoor Heating" page. This would be the authoritative URL, as opposed to the URLs produced by links from within the "Catalina Creations" brand category page, or from the home page, or from the "Outdoors" category page, or from the "Outdoor Living" page, etc.

Be careful though: this powerful form of canonicalization via intelligent redirection does eliminate page duplication much more quickly and forcibly redistributes available link popularity to the greatest number of unique landing pages. However, this technique's power is dangerous because it's difficult to detect if you've accidently nuked important parts of your site.

If you have launched faceted navigation without a focus on plugging crawl leaks, your clean-up job becomes more difficult and time consuming to manage, but not impossible. If you can marshal the resources to retrofit your site, many of the above steps should be taken. Tactics like disallows and meta noindexing can take up to 12 months to achieve desired de-indexation effects, and all these steps need to

be actively monitored for effectiveness. An easier option may be to outsource this work and deploy a new set of landing pages that apply the needed changes automatically for search engines and their users, while leaving the faceted navigation structure as is for all other website users.

During the clean up period, consistent tracking of basic natural search data is critical to determine progress. The team should track indexation of the critical URL types across all major engines, backlink data, PageRank flow, natural search referred traffic and natural search referred sales. More advanced KPIs should be monitored as well. Look for a decrease in the total number of URL types crawled and indexed over a comparable timeframe, before and after. Naturally, if you do not see a decrease, you need to be able to segment URLs crawled by googlebot or other bots, to diagnose further crawl leak issues. Once duplication is removed, you should see all your remaining landing pages crawled with more frequency than before; a greater quantity of your remaining landing pages driving search traffic; and more traffic per page as a result of higher rankings of each page in relevant keyword markets. That means significant increases in brand visibility and higher margin sales.

If you're still with me: congratulations! You can see that the process of eliminating crawl leaks and duplication induced by faceted navigation is advanced work. It can be intimidating and done wrongly very easily. Think twice about trying this at home if your organization lacks the technological capability to execute tactics outlined above, access to advanced crawler analytics, or the human resources to make sense of the data and adjust tactics. But done right, you can make faceted navigation work harder for you to produce not only a stronger natural search presence but increased online sales as well.

About Netconcepts, LLC

Netconcepts, the leading provider of natural search marketing, offers automated natural search technology and value added services tailored to help retailers drive online sales. For the largest retail brands, Netconcepts' GravityStream solution is an award winning, patent pending SaaS technology that powers the natural search channel. GravityMarket is a comprehensive eBusiness solution for mid-market retailers. Founded in 1995, the company is based in Madison, Wisconsin with R & D facilities in Auckland, New Zealand. Netconcepts serves retail brands such as Sketchers, Northern Tool, BNET (a division of CNET), Allen Edmonds and Cabela`s among others. For more information, visit the company Web site at www.netconcepts.com.

Contact

Brian Klais
Executive Vice President of Search
2820 Walton Commons West, Suite 123
Madison, WI 53718
ph. 608-285-6600
fax: 608-285-6601
email: bklais@netconcepts.com