Search Engine Optimization

CS174

Chris Pollett

Dec 5, 2007.

Outline

- Factors which contribute to search engine ranking
- Link Optimizations
- On-page Optimizations

Introduction

- You want to make it easy for your customers to find and do business with you online.
- Since a main avenue to your web-site is a search engine, it is important to make the relevant parts of your web-site rank highly (first page or two) with respect to the most popular search engines.
- Today, we are going to look at some techniques for doing this.

Search Engine Algorithms

- You can find which are the most popular search engines by looking at searchengine.com or Neilson's.
- In order, these are Google, Yahoo!, MSN, Ask.com, AOL (uses Google).
- Google's famous PageRank (Brin Page 1998) algorithm assigns a rank, r(P), to a page P as the sum $\sum r(P_i)/|P_i|$ over each page P_i which links to it. Here $|P_i|$ is the number of links going out of page P_i.
- So to avoid chicken and egg problems this is actually computed in a clever iterative fashion.
- Pages are returned from a search based on whether they have the keywords and based on the page rank
- There are also several important tweaks to the algorithm, for instance to handle the problem of nodes without outlinks, the hyperlink matrix has a modification to make it stochastic.
- Ask.com (Teoma) uses another kind of algorithm called HITs (Kleinberg 1998). This works by iteratively computing two scores for a page: (a) an authority score measuring the sum of hubs linking to it and (b) a hub score measuring how good the authorities linked from this site are. It is query dependent in its calculation of (a) and (b).

Factors which contribute to search engine ranking

- From the above we see that we would like lots of links into our site.
- Hopefully, these would be from sites that don't have too many outgoing links.
- However, using a **link farm** or even being associated with the IP of such a site can get one banned from search results.
- We also want to be using the correct keywords to get our potential customers to our product.
- In some rankings it can help if we link to good related sites.
- To get our site noticed at all you can submit your site to Google, Yahoo, DMOZ, etc using their sitemap APIs. That way, the crawlers will be aware of your site.

Link Optimizations

- As we've seen above, links are very important to the ranking of web pages.
- In order for a link to be used in a ranking algorithm it needs to be understood by the crawler or spider sent out by a search engine company to cache pages.
- Links like: product.php?blah=1, product.php?blah=2, etc. will often either not be followed or may be followed once under product.php .
- Links that need posted data to work won't be cached at all.
- Things like links in Javascript, Flash, etc tend not to be followed.
- Sometimes people look at the User-Agent header to see if it is a spider and based on this serve a different page which is more friendly to the spider.
- This is called **cloaking** and is risky as it can get one banned from search results.

URL Rewriting

- To handle the problem of links like: product.php?blah=1, product.php?blah=2

  you can use URL rewriting.
- To do this you need to make sure the mod_rewrite module is loaded and added in Apache.
- Then you need to have a line like:

  RewriteEngine on
- In your .htaccess file you can then add a sequence of lines like:

  RewriteRule reg_expr_pattern substitution
- For example,

  RewriteRule ^product/(.*)/$ /product.php?blah=$1
- We can of course use more general reg_exps than above

Site Map

- A site map is one page on your website which contains links to all the most common other pages on your website.
- It can help ensure that a spider will hit all the pages that you want on your site.

RSS Feeds

- RSS (really simple syndication) is an XML language.
- Having a RSS feed is a way of publishing what's new on your website.
- People, aggregator,etc will peridiocally download the feed file. For instance, every hour, so it is a very useful tool.
- An example file looks like:

```
<?xml version="1.0"?>
    <rss version="2.0"
        xmlns:xlink="http://www.w3.org/1999/xlink">
      <channel>
      <title>Computer Science Department, San Jose State University -
    Old News Items </title>
      <link>http://www.cs.sjsu.edu/</link>
      <description>Archived of selected news items from the Department of Computer Science at SJSU.</description>
      <language>en-us</language>
      <lastBuildDate>Thursday, 16 February  2006</lastBuildDate>
    <item>
    <title>Job opportunity for CS students or graduates.</title>
    <pubDate>Monday, 8 May 2006</pubDate>
    <description>    JOB OPPORTUNITY!
       PageBites, a Palo Alto startup, is hiring. If you are interested
    </description>
    </item><!-- could add more items -->
  </channel>
</rss>
```

On-page Optimizations

- Besides trying to improve the links on your site and links to your site, you also want to make sure that keywords associated with your product can be found when the spider hits the page.
- One way to find which keywords related to your product are frequently searched on is to look at a site like: inventory.overture.com or wordtracker.com
- Some engines give more weight to keywords contained in certain places.
- Some good places to put important keywords are the title and meta tags, in the first h1 and h2 tags of your document and in links.
- If possible you also want links which point to your site to contain the keyword.