Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.229 – 234

RESEARCH ARTICLE

Enhancing Search Engine Optimization using Web Content Mining

[1]Mr. S.Balamurugan, [2]Dr. S.Thirunirai Senthil

[1]PG Student, Department Of Computer Science and Engg, PRIST University, Puducherry, INDIA
[2]Professor, Department Of Computer Science and Engg, PRIST University, Puducherry, INDIA

ABSTRACT-***Most of the website suffers from technological issues, some which are conscious and some which are not aware. Its capacity depends upon the background done in web content mining, when changed domain or web mining or just common mistakes often have done by webmasters. Deciding to hire an SEO is a big decision that can potentially improve your site and save time, but you can also risk damage to your site and reputation to research the potential advantages as well as the damage that an irresponsible SEO can do to your site. It is not just an implementation for measuring privacy of data but can also be used as a tool for business and market research, and to access and also improve the effectiveness of a web site. By understanding how and when your projection and clients are interacting with your content and applications, you can successfully optimize your online presence according to both their needs and your business goals.***

Keywords— ***Content development, SEO, optimization, technical, versions, webmaster***

1. Introduction

Web mining is a fast growing investigation area which consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the finding of client access patterns from Web usage logs. Web structure mining discovers useful knowledge from the construction of hyperlinks. Web content mining aspires to extract useful information or knowledge from web page contents. Web content mining is associated but unusual from data mining and text mining. It is linked to data mining because a lot of data mining techniques can be practical in Web content mining. It is interconnected to text mining because much of the web contents are texts. Nevertheless, it is also quite dissimilar from data mining because Web data are principally semi-structured or unstructured, even as data mining contract primarily with structured data. Web content mining is also different from text mining because of the semi-structure environment of the Web, while text mining focuses on unstructured texts. Web content mining thus needs creative purpose of data mining or text mining techniques and its own exceptional approaches. In the earlier period, there was a speedy expansion of behavior in the Web content mining area. This is not amazing because of the extraordinary growth of the Web contents and noteworthy economic benefit of such mining. Yet, owing to the heterogeneity and requires of construction of Web data, automated detection of unexpected knowledge data still present many tough research problems.

Every time you enter a query in a search engine and press it, you obtain a catalog of web results that hold that query term. Generally users are likely to visit websites that are at the top of the list as they receive those to be more relevant to the query. Have ever wondered why some of websites rank better than the others, it is because of a powerful web marketing technique called Search Engine Optimization (SEO). SEO is a technique which helps search engines find and rank your website superior than the millions of other websites in response to a search query. SEO therefore helps you get traffic from search engines and plays a vital role in website.

2. How Search Engines Work

The first basic truth you need to know to learn SEO is that search engines are not humans. While this might be obvious for everybody, the differences between how humans and search engines view web pages aren't. Unlike humans, search engines are text-driven. Although technology advances rapidly, search engines are far from intelligent creatures that can feel the beauty of a cool design or enjoy the sounds and movement in movies. Instead, search engines based on the Web, looking at particular site items (mainly text) to get an idea what a site is about. This brief explanation is not the most precise because as we will see next, search engines perform several activities in order to deliver search results – crawling, indexing, processing, calculating relevancy, and retrieving. First, search engines based on the Web to see what is there. This task is performed by a piece of software, called a crawler or a spider. Spiders follow links from one page to another and index everything they find on their way. Having in mind the number of pages on the Web (over 10 billion), it is impossible for a spider to visit a site daily just to see if a new page has appeared or if an existing page has been modified, sometimes crawlers may not end up visiting your site for a month or two. To check what a crawler sees from your site. As already mentioned, crawlers are not humans and they do not see images, Flash movies, JavaScript, frames, password-protected pages and directories, so if you have tons of these on your site, you'd better run the Spider Simulator below to see if these goodies are viewable by the spider. If they are not viewable, they will not be spidered, not indexed, not processed, etc. - in a word they will be non-existent for search engines. After a page is crawled, the next step is to index its content. The indexed page is stored in a giant database, from where it can later be retrieved. Essentially, the process of indexing is identifying the words and expressions that best describe the page and assigning the page to particular keywords. For a human it will not be possible to process such amount of information but generally search engines deal just fine with this task. Sometimes they might not get the meaning of a page right but if you help them by optimizing it, it will be easier for them to classify your pages correctly and for you – to get higher rankings. When a search request comes, the search engine processes it – i.e. it compares the search string in the search request with the indexed pages in the database. Since it is likely that more than one page (practically it is millions of pages) contains the search string, the search engine starts calculating the relevancy of each of the pages in its index with the search string. There are various algorithms to calculate relevancy. Each of these algorithms has different relative weights for common factors like keyword density, links, or meta tags. That is why different search engines give different search results pages for the same search string. What is more, it is a known fact that all major search engines. Periodically change their algorithms and if you want to keep at the top, you also need to adapt your pages to the latest changes. This is one reason (the other is your competitors) to devote permanent efforts to SEO, if you'd like to be at the top. The last step in search engines' activity is retrieving the results. Basically, it is nothing more than simply displaying them in the browser – i.e. the endless pages of search results that are sorted from the most relevant to the least relevant sites.

3. Common Technical SEO problems

Lack of original content: Depending on how severe this problem is, it might spread out link power on your site, lowering rankings or even makes search engines classify you as a scraper site.

Long and unreadable URLs: A common SEO problem among old content management systems is session ID's and other parameters in URL's making it hard for humans to understand what the page is about. Basically, you want a logic URL with words describing the topic. This way, you are likely to receive more clicks from people arriving from search engines.

Redirects or lack of it: If your company has multiple domains and campaign sites it's important to sort out redirects. The most common redirect to use is a 301-redirect. Sometimes, you need to fix old and dead links with 301-redirects.

Menu problems: It's not uncommon with menu links in java script format. However, it's not search engine friendly since its hard and impossible for search engines to find those pages.

No sitemap? If the main navigation or internal linking is not perfect, the importance of a sitemap is even greater. It's usually a great idea to include a XML sitemap for search engine robots and a HTML sitemap for users.

3.1. Uppercase vs. Lowercase URLs

The problem stems from the fact that the server is configured to respond to URLs with uppercase letters and not to redirect or rewrite to the lowercase version. Generally, the search engines are much better at choosing the canonical version and ignoring the duplicates. However, many instances of search engines doesn't always do this properly, which means that you should make it explicit and not rely on the search engines to figure it out for themselves. There is a URL rewrite module which can help solve this problem on IIS 7 servers. The tool has a nice option within the interface that allows you to enforce lowercase URLs. If you do this, a rule will be added to the web.config file which will solve the problem.

3.2. Multiple versions of the homepage

Again, this is a problem encountered in .NET websites, but it can happen quite easily on other platforms. If we start a site audit on a site is .NET, it almost immediately goes and checks if this page exists:

www.pattern.com/default.aspx

This is a duplicate of the homepage that the search engines can usually find via navigation or XML sitemaps. Other platforms can also generate URLs like this:

www.pattern.com/index.html
www.pattern.com/home

Get into the minor details of how these pages are generated because the solution is quite simple. Again, modern search engines can deal with this problem, but it is still best practice to remove the issue in the first place and make it clear. Finding these pages can be a bit tricky as different platforms can generate different URL structures, so the solution can be a bit of a guessing game. Instead, do a crawl of your site, export the crawl into a CSV, filter by the META title column, and search for the homepage title. You'll easily be able to find duplicates of your homepage. To prefer to solve this problem by adding a 301 redirect to the duplicate version of the page which points to the correct version. You can also solve the issue by using the canonical tag, but I stand by a 301 redirect in most cases. Another solution is to conduct a site crawl using a tool like Screaming Frog to find internal links pointing to the duplicate page. You can then go in and edit the duplicate pages so they point directly to the correct URL, rather than having internal links going via a 301 and losing a bit of link equity.

3.3. Query parameters added to the end of URLs

This problem tends to occur mostly in ecommerce websites that are database driven. There of a chance of occurrence on any site, but the problem tends to be bigger on ecommerce websites as there are often loads of product attributes and filtering options such as color, size, etc. In these cases, the URLs which the users click on are relatively friendly in terms of SEO, but quite often you can end up with URLs such as this:

www.pattern.com/product-category?colour=12

This pattern would filter the product category by a certain color. Filtering in this capacity is good for users but may not be great for search, especially if customers do not search for the specific type of product using color. If this is the case, this URL is not a great landing page to target with certain keywords. Another possible issue that has a tendency to use up TONS of crawl budget is when said parameters are combined together. To make things worse, sometimes the parameters can be combined in different orders but will return the same content. For pattern:

www.pattern.com/product-category?colour=12&size=5
www.pattern.com/product-category?size=5&colour=12

Both of these URLs would return the same content but because the paths are different, the pages could be interpreted as duplicate content. Before going further, want to address another common, related problem: the URLs may not be SEO friendly because they are not database driven. This isn't the issue I'm concerned about in this particular scenario as there is more concern about wasted crawl budget and having pages indexed which do not need to be, but it is still relevant. The first place to start is addressing which pages you want to allow crawling and indexing. This decision should be driven by your keyword research, and you need to cross reference all database attributes with your core target keywords. Let's continue with the theme from Go Outdoors for our pattern:

Here are our core keywords:

- Waterproof jackets
- Hiking boots
- Women's walking trousers

In the website, each of these products will have attributes associated with them which will be part of the database. Some common patterns include:

- Size (i.e. Large)
- Color (i.e. Black)
- Price (i.e. £49.99)
- Brand (i.e. North Face)

Your job is to find out which of these attributes are parts of the keywords used to find the products. You also need to determine what combination (if any) of these attributes are used by your audience. In doing so, you may find that there is a high search volume for keywords that include "North Face" + "waterproof jackets." This means that you will want a landing page for "North Face waterproof jackets" to be crawl able and index able. You may also want to make sure that the database attribute has an SEO friendly URL, so rather than "waterproof-jackets/? Brand=5" you will choose "waterproof-jackets/north-face/." You also want to make sure that these URLs are part of the navigation structure of your website to ensure a good flow of Page Rank so that users can find these pages easily.

3.4. Soft 404 errors

A soft 404 is a page that looks like a 404 but returns a HTTP status code 200. In this instance, the user sees some text along the lines of "Sorry the page you requested cannot be found." But behind the scenes, a code 200 is telling search engines that the page is working correctly. This disconnect can cause problems with pages being crawled and indexed when you do not want them to be. A soft 404 also means you cannot spot real broken pages and identify areas of your website where users are receiving a bad experience. From a link building perspective (I had to mention it somewhere!), neither solution is a good option. You may have incoming links to broken URLs, but the links will be hard to track down and redirect to the correct page.

Fortunately, this is a relatively simply fix for a developer who can set the page to return a 404 status code instead of a 200. Whilst you're there, you can have some fun and make a cool 404 page for your user's enjoyment. To find soft 404s, you can use the feature in Webmaster Tools which will tell you about the ones have detected: Again, this is an easy redirect for developers to get wrong because, from a user's perspective, they can't tell the difference. However, the search engines treat these redirects very differently. Just to recap, a 301 redirect is permanent and the search engines will treat it as such; they'll pass link equity across to the new page. A 302 redirect is a temporary redirect and the search engines will not pass link equity because they expect the original page to come back at some point. To find 302 redirected URLs, recommend using a deep crawler such as Screaming Frog or the IIS SEO Toolkit. You can then filter by 302s and check to see if they should really be 302s, or if they should be 301s instead. To fix the problem, you will need to ask your developers to change the rule so that a 301 redirect is used rather than a 302 redirect.

3.5. Broken/Outdated sitemaps

Even as not essential, XML sitemaps are very useful to the search engines to make sure they can find all URLs that you care about. They can give the search engines a nudge in the right direction. Unfortunately, some XML sitemaps are generated one-time-only and quickly become outdated, causing them to contain broken links and not contain new URLs. Ideally, your XML sitemaps should be updated regularly so that broken URLs are removed and new URLs are added. This is more important if you have a large website that adds new pages all the time. Bing has also said that they have a threshold for "dirt" in a sitemap and if the threshold is hit, they will not trust it as much. First, you should do an audit of your current sitemap to find broken links. This great tool from Mike King can do the job. Second, you should speak to your developers about making your XML sitemap dynamic so that it updates regularly. Depending on your resources, this could be once a day, once a week, or once a month. There will be some development time required here, but it will save you (and them) plenty of time in the long run. An extra tip here: you can experiment and create sitemaps which only contain new products and have these particular sitemaps update more regularly than your standard sitemaps. You could also do a bit of extra-lifting if you have dev resources to create a sitemap which only contains URLs which are not indexed.

4. Search Engine Optimization Process

Data/information extraction: Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.

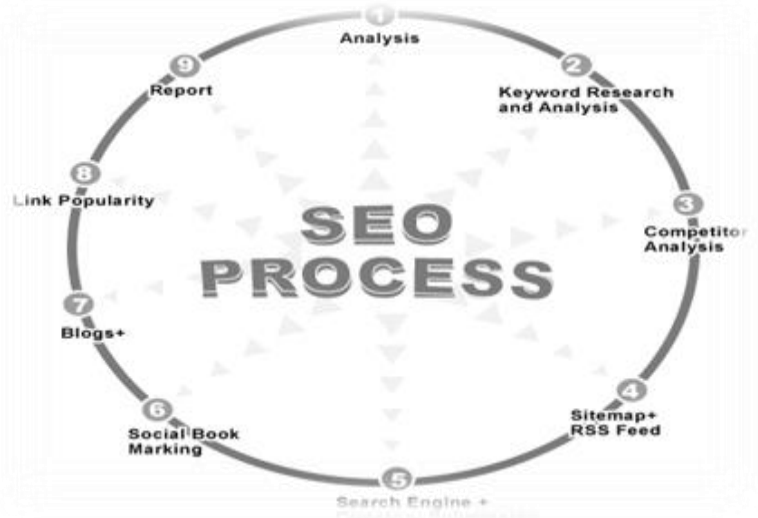
Web information integration and schema matching: Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.

Opinion extraction from online sources: There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking. We will introduce a few tasks and techniques to mine such sources.

Knowledge synthesis: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.

Segmenting Web pages and detecting noise: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem. A number of interesting techniques have been proposed in the past few years.

All these tasks present major research challenges and their solutions also have immediate real-life applications. The tutorial will start with a short motivation of the Web content mining. We then discuss the difference between web content mining and text mining, and between Web content mining and data mining. This is followed by presenting the above problems and current state-of-the-art techniques. Various patterns will also be given to help participants to better understand how this technology can be deployed and to help businesses. All parts of the tutorial will have a mix of research and industry flavor, addressing seminal research concepts and looking at the technology from an industry angle. SEO is a technique which helps search engines find and rank your site higher than the millions of other sites in response to a search query.



Identification of keywords: To identify the keywords that people will likely use to search with.
SEO Copywriting: To develop text that search engines will like.
HTML Coding: To write minor modifications of HTML code for SEO.
On-page optimization: To develop CSS et al. for the purpose of SEO.
Developing Optimized Pages: To create particular pages that will improve your optimization.
Search Engine and Directory Submission: To find the right directories for your site and have your link added.
Link Building: To get links to your site in the places that your customers will see.
Post Optimization Analysis: We prepare ranking reports, submission reports, traffic reports, and log file reports.
Study of website to be optimized: To identify what the website will be designed to achieve.

5. Operation of Major Search Engines

Although the basic principle of operation of all search engines is the same, the minor differences between them lead to major changes in results relevancy. For different search engines different factors are important. There were times, when SEO experts joked that the algorithms of Bing are intentionally made just the opposite of those of . While this might have a grain of truth, it is a matter a fact that the major search engines like different stuff and if you plan to conquer more than one of them, you need to optimize carefully. There are many patterns of the differences between search engines. For instance, for Yahoo! and Bing, on-page keyword factors are of primary importance, while for many links are very, very important. Also, for many sites are like wine – the older, the better, while Yahoo! generally has no expressed preference towards sites and domains with tradition (i.e. older ones).

6. Strategy for SEO

Many SEOs and other agencies and consultants provide useful services for website owners, including:

- Review of your site content or structure
- Technical advice on website development: for pattern, hosting, redirects, error pages, use of JavaScript
- Content development
- Management of online business development campaigns
- Keyword research
- SEO training
- Expertise in specific markets and geographies.

SEO or Search Engine Optimization delivers measurable results that increase website traffic and online revenues. You need Search Engine friendly web design, rich unique on-page and off-page content and a balanced link profile to succeed.

- Increasing web traffic
- Maximizing web conversion
- Growing web revenues
- Website Competitive Analysis
- Keyword Marketing Research
- SE Compatibility Analysis
- On Page Optimization
- Keyword Marketing Research
- Link Popularity Development

7. Conclusion

On-Page SEO is mainly focused around website content, site structure, keyword usage in titles, headings and images, along with internal and external link usage. The key factors to overlook and deal with in order to succeed with SEO

A. What part of On-Page SEO do you start with?

On-page SEO tasks can start with a review of your web content and how to make it continuously updated. Old and statics content are seldom popular for a long period of time among search engines.

If you manage to create outstanding web content, both backlinks and followers will increase.

B. Targeting relevant keywords?

It's important to set-up goals for keywords you aim to target. Try to naturally include appropriate keywords you want to be found for in titles, headlines, image text and meta descriptions and internal links. If done properly, your chances to achieve prominent search engine rankings increase.

C. Solving technical problems

Technical issues often give problems with regards to website indexation and general search engine visibility. Make sure to review your URL structure, set-up redirects for dead pages and add a sitemap so search engines can find all your sub-pages.

D. Being unique and user friendly

Finally, it's important to make sure you site is user friendly and unique. Both your visitors and search engines will certainly appreciate your efforts in this area.

References

[1]Beel, Jöran and Gipp, Bela and Wilde, Erik (2010). "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Scholar and Co.". Journal of Scholarly Publishing. pp. 173–190. Retrieved April 18, 2010.

[2]Brian Pinkerton. "Finding What People Want: Experiences with the WebCrawler" (PDF). The Second International WWW Conference Chicago, USA, October 17–20, 1994. Retrieved May 7, 2007.

[3] Danny Sullivan (June 14, 2004). "Who Invented the Term "Search Engine Optimization"?". Search Engine Watch. Retrieved May 14, 2007.See groups thread.

[4](Document Number 19970801004204) "Documentation of Who Invented SEO at the Internet Way Back Machine". Internet Way Back Machine.Archived from (Document Number 19970801004204) the original on August 1, 1997.

[5]Cory Doctorow (August 23, 2001). "Metacrap: Putting the torch to seven straw-men of the meta-utopia". e-LearningGuru. Archived from the original on April 9, 2007.
Retrieved May 8, 2007.

[6] Pringle, G., Allison, L., and Dowe, D. (April 1998). "What is a tall poppy among web pages?".Proc. 7th Int. World Wide Web Conference.Retrieved May 8, 2007.

[7] Brin, Sergey and Page, Larry (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine".Proceedings of the seventh international conference on World Wide Web. pp. 107–117. Retrieved May 8, 2007.

[8] Thompson, Bill (December 19, 2003). "Is good for you?". BBC News. Retrieved May 13, 2007.

[9] ZoltanGyongyi and Hector Garcia-Molina (2005)."Link Spam Alliances" (PDF).Proceedings of the 31st VLDB Conference, Trondheim, Norway.Retrieved May 9, 2007.

[10] Hansell, Saul (June 3, 2007). " Keeps Tweaking Its Search Engine". New York Times.Retrieved June 3, 2007.

[11]Danny Sullivan (September 29, 2005). "Rundown On Search Ranking Factors". Search Engine Watch. Retrieved May 8, 2007.

[12]Christine Churchill (November 23, 2005). "Understanding Search Engine Patents". Search Engine Watch. Retrieved May 8, 2007.

[13]" Personalized Search Leaves Labs". *searchenginewatch.com*. Search Engine Watch. Retrieved September 5, 2009.

[14]"Will Personal Search Turn SEO On Its Ear? | WebProNews".www.webpronews.com. Retrieved September 5, 2009.

[15]"8 Things We Learned About PageRank". www.searchenginejournal.com. Retrieved August 17, 2009.

Books

[1]Search Engine Optimization: An Hour a Day by Jennifer Grappone and GradivaCouzin

[2]Landing Page Optimization: The Definitive Guide to Testing and Tuning for Conversions by Tim Ash**.**

[3] Building Findable Websites: Web Standards SEO and Beyond **by** Aarron Walter