CHOOSING SEARCH ENGINE OPTIMIZATION TECHNIQUES FOR DESIGNED INTERNET RESOURCE

Raspopov A.V.
Research supervisor: Associate Professor Sherstnev V.S.
Tomsk Polytechnic University
rsppv.av@gmail.com

Search engine optimization (SEO) is a key element of successful Internet resources promotion. SEO is the methods used to boost the ranking or frequency of a website in the results returned by a search engine, in an effort to maximize user traffic to the site [1]. The SEO methods following makes website easier for search engines to crawl, index and understand its content.

The designed Internet resource is a web application for publishing content from Oracle Universal Content Management system. All kind of unstructured data can be used as the content but in this case there are mostly scientific documents of TPU. Oracle Universal Content Management system enables organizations to share, manage, and distribute business information using a Web site as a low-cost access point [3]. However, the UCM system provides access only to users of internal computing network within an organization — the Intranet. There is necessity to supply access to stored scientific documents for Internet users. Thus, it is required to develop the web application and ensure website content indexing by search engines.

Therefore, to achieve the goal of successful indexing of designed web resource content, the following objectives were set:

- To learn a search engine architecture;
- To consider the factors those affect the order of search engine results;
- To examine search engine optimization techniques;
- To choose acceptable SEO methods regarding website promotion;
- To ensure that chosen SEO methods provide successful indexing of the developed web application.

There is necessity to learn a search engine architecture to gain a complete understanding of the optimization process. All search engines usually consist of following components:

- Web crawling module;
  - Crawler;
  - Spider;
  - Indexer;
- Database or search engine index;
- Web server;
  - Client web application;
  - Relevance ranking module.

It is necessary to consider these concepts in more detail. The web crawling module finds information on the existing Web pages, builds lists of the words and adds them to a search engine index. The web crawling module includes the crawler, spider and indexer.

The crawler is software that follows or *crawls* links throughout the Internet. However, crawlers can only follow links from one page to another and from one site to another. That is the primary reason why inbound links to web site are so important.

The spider is a special software robot that is responsible for downloading web pages just like a browser, i.e. a clean HTML page without *Cascading Style Sheets*, *JavaScript*, images, etc. Also, the spider finds all links on the page, collects them and transfers to the indexer.

The indexer is a program that reads and analyzes the information from downloaded web pages. The main aim of the indexer is lexical and morphological content analysis [2]. When the indexer scans an HTML page, it only considers the words within the page and location of these words on the page. Words occurring in the title, subtitles, meta tags and other important positions are noted for special consideration during a subsequent user search [5]. After content analysis the indexer builds a list of found words and saves it in search engine database, also known as index.

A search engine index is similar to an index list of a book. The index contains a word set and pages with occurrence of each word. One of the issues is effectively storing huge data sets. For example, the word *flamingo* appears in about 492,000 out of about 2 billion pages known to Google search engine [8].

The web server includes a client web application and relevance ranking module. It is responsible for web pages searching in index, relevance ranking and subsequent displaying the list of results returned by a search engine in response to a search query. According to computing dictionary on *Dictionary.com* relevance is «a measure of how closely a given object (file, web page, database record, etc.) matches a user's search for information» [4].

After learning search engine architecture it becomes clear that website SEO is a key item in search engine work. Obviously, SEO influences mostly relevance ranking module and search engine results page (SERP). Let's consider the main factors that affect the order of search engine results. The on-page factors include the keywords (words in user search query) location and density in the text of a Web page. The off-page factors contain the number and quality of inbound links to the website. Each search engine has its own algorithm of hyperlink analysis. For example, Google has PageRank and Yandex uses Thematic Citation Index. To raise the position of the projected online resource it is necessary to increase the quantity and quality of links from foreign resources [6]. Quality of the link depends on its weight. The weight depends on authority of source website. The webpage weight influences the position of website in SERP. However, a webmaster cannot manipulate off-page factors that is why all SEO techniques use on-page factors.

All SEO techniques affecting the search crawler can be divided into three groups: black, grey and white hat optimization techniques. Black hat techniques use optimization methods that search engines do not approve. In contrast, white hat techniques optimize a website without affecting the search crawler. White methods are divided into internal — work with the website structure — and external — all kinds of advertising. Grey hat optimization techniques also use licit methods of the Internet resources promotion. Despite this in some cases a resource can be locked by search engines if it uses grey hat techniques.

Taking into account the future of the designed Internet resource, it is appropriate to choose white hat optimization techniques avoiding the risk of blocking the designed web application. The designed software is an enterprise application and provides access to documents for external users. In this case external white hat methods (registration in catalogs, social networks, advertising) are not appropriate. Therefore, the reasonable choice for developers is to implement internal white hat methods, i.e. to improve the structure of the designed Internet resource. One of the features of the web application is indexing pages with description of each document that is why the structure is a key element or framework of the website.

To create clear structure for search engines it was decided to use the following internal white hat methods:

- Using a unique set of keywords for each page provides greater audience and more precise description of each document;
- Placing keywords in meta descriptions, tags *title, b, i, alt* and *title* at pictures and headers *h1, h2, h3,* etc.;
- Proper and unique filling of meta tags without using quotation, end of line marks and special characters provides additional user traffic to a website;
- The publication of new documents on the Internet resource ensures exclusive content and its regular updating;
- Appropriate configuration of file *robots.txt*. This text file is a document located on a website and used by search engine crawlers. It contains indexing parameters of the website for all or a particular search engine;
- Using semantic markup, also known as semantic HTML, and CSS file that includes decor of the page. The one cleans the web page of *trash*;
- Applying links with additional attribute *rel = nofollow* forbids indexing of untrusted content, paid links and useless pages for search engine [7].

Eventually, the web application was developed and chosen optimization techniques were implemented. Developed application was deployed on a webserver. The location of the application is *catalog1.vt.tpu.ru*. Embedded SEO methods promote successful content indexing. For example, according to *Google Webmaster Tools* and *Yandex.Webmaster* 4117 pages were indexed by Google and 5573 pages were indexed by Yandex on January 15, 2014.

Search engine optimization is a powerful tool that provides accessibility and intelligibility of information in the Internet. The Internet contains a lot of information, but the best information is the one that turns into useful knowledge. Therefore, the information should be structured and understandable, that is performed using SEO.

References

1. Search-engine optimization | Define Search-engine optimization at Dictionary.com // Dictionary.com – Free Online English Dictionary. [2014]. URL: http://dictionary.reference.com/browse/search-engine%20optimization (access date: 11.01.2014)

2. Internet search engines: Yandex, Google, Rambler, Yahoo. Composition, functions,work principles // Search engine marketing «Seonews». [2005-2014]. URL: http://www.seonews.ru/masterclasses/poiskovyie-sistemyi-interneta-yandeks-google-rambler-yahoo-sostav-funktsii (access date: 9.01.2014)

3. Oracle Fusion Middleware System Administrator's Guide for Content Server // Oracle Documentation. [2010]. URL: http://docs.oracle.com/cd/E21043_01/doc.1111/e10792/c01_introduction001.htm#i1048669 (access date: 7.01.2014)

4. Relevance | Define Relevance at Dictionary.com // Dictionary.com – Free Online English Dictionary. [2014]. URL: http://dictionary.reference.com/browse/relevance (access date: 27.12.2013)

5. How Internet Search Engines Work. Curt Franklin // HowStuffWorks “Learn How Everything Works!”. [1998-2014]. URL: http://computer.howstuffworks.com/internet/basics/search-engine.htm (access date: 27.12.2013)

6. What is TIC? // Help section of Yandex.Webmaster. [2004-2014]. URL: http://help.yandex.com/webmaster/recommendations/intro.xml (access date: 8.01.2014)

7. Rel = “nofollow” // Webmasters Tools Help. [2014]. URL: https://support.google.com/webmasters/answer/96569?hl=en (access date: 14.01.2014)

8. Anatomy of a Search Engine: Inside Google // Search Engine Watch. [2014]. URL: http://searchenginewatch.com/article/2064446/Anatomy-of-a-Search-Engine-Inside-Google (access date: 13.01.2014)