Optimizing for Search Indexes

Paul Kovacs
Computer Information Systems, Robert Morris University
Moon Township, PA 15108, USA

ABSTRACT

The expansion of Internet technology is challenging information and instructional technology educators to develop courses related to the teaching of Web site design and development. Such courses should follow all of the recommended web site design practices such as site organization, site navigation, page design, text design, graphic design, and accessibility considerations. Whether the purpose of a Web site is to run and promote a business, provide a service, disseminate information, or just establish a web presence, an extremely important additional topic in a Web design course is to assure that the target audience will find the site. Search indexes are essential starting points for users seeking to find Web locations or documents and getting a site listed in a search index can be an invaluable tool for reaching a target market. This paper discusses the working components of a crawler-based search indexes as well as provides optimization techniques that can be included in Web design and development courses.

Keywords: web design, search indexes, optimizing HTML tags, world wide web indexing

1. INTRODUCTION

The public's enthusiasm and business acceptance for the Internet is unparalleled in the history of information technology. The record growth of the Internet has exceeded all previous records for both the brief time of acceptance and the number of businesses embracing this new technology. The new millennium was a crossroads for the Internet and over 50 percent of all Americans homes and more than 90 percent of all college students had internet access (Lehnert 2001).

This expansion of Internet technology is challenging information and instructional technology educators to develop courses related to the teaching of Web site design and development.

Such courses should follow all of the recommended web site design practices such as site organization, site navigation, page design, text design, graphic design, and accessibility considerations.

Whether the purpose of a Web site is to run and promote a business, provide a service, disseminate information, or just establish a web presence, an extremely important additional topic in a Web design course is to assure that the target audience will find the site. There are a number of ways to accomplish this including traditional media advertising , announcements with Usenet newsgroups, e-mail lists, Listserv Postings, reciprocal links agreements, or personal recommendations. However, according to a survey by the Georgia Institute of Technology (Pitkow and Kehoe 1997), more than 86% of all people locate a Web site through a search indexes as opposed to Printed media (62.47%), friends (56.92%), TV (30.20%), email signatures (31.23%), Usenet Newsgroups (32.75%) or other methods. A more

recent Nielsen/NetRating survey reported at searchwatch.com (Sullivan, 2002) indicates that nine out of ten web surfers utilize a search index when they need to find a Web site, and a Consumer Daily Question Study (Felke, 2003) found that search indexes were the most utilized method of obtaining information and finding sites.

In the context of the World Wide Web and Indexing, the term "search index" can describe both crawler-based search indexes and human-edited directories. Crawler-based search indexes, such as Excite, index words or terms in Web documents automatically by transversing or crawling the web. Conversely, directories such as Yahoo, classify locations or Web documents into an arbitrary subject classification scheme or taxonomy that are produced by human beings who visit and consider websites for inclusion. This paper discusses the components of a crawler-based search indexes as well as provides optimization techniques for Web design and development courses. The paper does not discuss Human-edited Search directories.

2. COMPONENTS OF A CRAWLER-BASED SEARCH INDEXES

Crawler based Search indexes have the following three major components:

- Spider
- Index
- Search form

The first component is the spider also called the crawler, robot, or worm. These names are somewhat deceptive because they give the impression of undesirable software such as a virus. This not true and they are not destructive. The spider simply visits a web page, reads it, and then follows links to other pages within the site. http://www.jafsoft.com/searchengines/webots.html provides a list of spiders for the major search indexes (Fotheringham 2003).

The second component is the Index or the database, catalogue or dataset that contains a copy of every web page that the spider locates. If a web page is changed, then new information updates the index. Many times, it can take a while to add new or altered pages to the index. Thus, a web page may be "spidered" but not "indexed." Until the page is actually inserted into the database, it is not available to those searching for it.

The third component is the Search form and allows for the entry of key words or phrases. When submitting a search form, a special program analyzes the search query, and then searches the database or index for pages that contain this keyword or keyword phrase. Next it analyzes every single one of the relevant pages in order to determine how important that keyword or phrase is on that page.

A search index can contain hundreds of millions of web pages in its database. In order to deliver relevant results, special ranking algorithms calculate the order of search query results or the keyword or keyword phrase the searcher enters into the search form.

Search Indexes such as Excite (http://www.excite.com), AltaVista (http://www.altavista.com), and Hotbot (http://www.hotbot.com) index Web pages by starting from a historical list of URLs, especially of documents with many links elsewhere, such as "What's New" pages, server lists, and the most popular sites on the Web. The spider may then decide to parse the document and insert it into its database. Some Search Indexes spiders only index the home page of the Web site while others index all of the links in the site. The Web page's title, body, and other elements all play a role in the indexing. Some indexes also parse the META tag, or other special hidden tags.

3. ELECTING KEYWORDS AND PHRASES

As mentioned in the preceding section, when searching for a Web location or document, searchers enter various keywords in the search field of a search form. The search index then looks for Web pages that match those keywords. It follows that the starting point of search index optimization is selecting crucial keywords (one word) or keyword phrases (more than one word) that describe the Web site or page.

Assemble a list of relevant keywords. This list can contain 10 to 25 words or phrases arranged in order of importance. Do not in-

clude keywords into the list that have nothing to do with the Web site and do not repeat keywords. In assembling the list, think of what a user would enter when searching for the Web site. If there are not enough words to describe the site, a thesaurus may be helpful.

The Keywords selected should not be too general or too restricted. Too general Keywords can result in a large number of sites found. Conversely, keywords too restricted may reduce the number of sites found, but not many searchers will use that keyword.

Several Web resources for selecting a keyword list are:

- Overture's Search Term Suggestions Tool http://inventory.overture.com/d/searchinventory/suggestion/
- Word Tracker http://www.wordtracker.com/
- Lexical FreeNet Connected thesaurus http://www.lexfn.com/
- iBoost Keyword Generator http://www.iboost.com/tools/keyword_generator.htm
- Ranks.nl Keyword Density, Placement, Prominence Analyzer http://www.ranks.nl/tools/spider.html

Keyword density refers to the ratio of a keyword or phrase to the total words in a Web page. The keywords should be between 1% and 7%. This means that there should be 1 to 7 keywords for every 100 words on the Web page. A Web developer can manually count the keywords or can acquire a software product to do the count. Two software products that will analyze keyword density are:

- WebPositionGold by FirstPlace Software http://www.webposition.com/
- GRKda Keyword Density Analyzer by GRSoftware http://www.grsoftware.net/search_engines/software/grkda.html

4. OPTIMIZING HTML TAGS

A search index does not view a Web page as a human does but rather reads only the text-based code contained in the HTML tags. Thus, optimization for a search index is really optimization of selected HTML tags. Of course, in order to accomplish this, it is necessary to have some understanding of HTML. Many Web design courses use WYSIWYG *(what you see is what you get)* web authoring software such as Microsoft FrontPage or Macromedia Dreamweaver. Although this software allows for Rapid Application Development (RAD), it can also hide the HTML from the Web designer. Furthermore, some Web authoring programs produce HTML code that is full of incorrect or unnecessary tags. As a result, when optimizing the page for a search index, it may be necessary to rewrite parts of the code itself.

In section 6. CLASSROOM IMPLEMENTATION of this paper, an example is provided as to how optimization techniques are incorporated in a course at Robert Morris University. This course teaches HTML and uses HTML for the development of a Web site not web authoring software. When students complete the class, they may want to learn and use a web authoring tool. This approach assures that the student fully understands the HTML language and if using Web authoring software, should be able to rewrite HTML code if necessary.

The title tag

The <TITLE>tag contains the purpose for the Web page. Just about every crawler-based index will look at this tag for keywords and their placement in it. Use the keyword from the keyword list in making the title as descriptive as possible. The more descriptive the title, the more correctly the site will appear on a query to the search index database. Additionally, a descriptive title will be helpful when a visitor bookmarks or saves the Web site because the keywords in this tag determine the name the site is list under in a visitor's hotlist.

Keep the title short because too many keywords will lessen the relevance or the keywords that are included. It is advisable to use no more than 6-10 keywords. Additionally, do not put the name of the Web site first unless, of course, the name contains the essential keyword phrase.

For example:
<title>Inflight Training Solutions</title>, is better written as

<title>Training for the Corporate Flight Attendant by Inflight Training Solutions in Philadelphia, Pennsylvania</title>.

META tags

Meta tags are self-contained text written into the HTML code. They are not visible in a browser window and are placed in the HEAD section of the Web page between the <HEAD> and </HEAD> tags. Although there are a number of META tags with various uses, only three name attribute meta tags are of importance for search index optimization, the Description, the Keyword, and the ROBOT.

The description tag provides a description of the Web page in place of the explanation the search index would ordinarily create. This description should be a normal sentence that provides information about the site. The keyword tag provides keywords for the index to associate with the Web page and should include the keywords and phrases that are relevant to the page content. The ROBOTS tag sets indexing controls.

The format of these three tags is as follows:

<META NAME="*name*" CONTENT= "*content*">

The name attribute is the name of the tag in this case description, keyword, or robot and the content attribute provides keywords or information.

An example of the description and keyword tag for a Web site designed for a company called Inflight Training Solutions, which provides training for corporate flight attendants may be as follows:

<META Name="description" Content="Inflight Training Solution of Philadelphia prepares professionals for a career in aviation as a Corporate Flight Attendant">

<META name="keywords" content=" flight, attendant, careers, training, corporate flight attendant training, flight attendant training, flight attendant career training, corporate aviation training, flight training, flight attendant school, flight attendant curriculum">

The Robots Meta Tag indicates to a spider which pages to index, which pages not to index, which links to follow, and which links not to follow**.**

The formal syntax for the ROBOT META tag content is:
CONTENT=ALL - Index the page and all of its links (default)
CONTENT=INDEX - Index the page
CONTENT=NOINDEX – Do not index the page
CONTENT=FOLLOW - Follow the links
CONTENT=NOFOLLOW – Do not follow the links

For example, to specify that all of the pages of a Web site should be indexed the format is:

<META NAME=”ROBOTS” CONTENT=”ALL”>

A second example specifies no indexing of a particular page:

<META NAME=”ROBOTS” CONTENT= ”NOINDEX”>

A third example illustrates that the indexing spider should neither index this document nor analyze it:

<META NAME="ROBOTS" CONTENT= "NOINDEX, NOFOLLOW">

Note that the case used in these examples is mandatory. According to the HTML 4.0 specification, specify index-related keywords and the ROBOTS name values in uppercase.

A Web site that will aid in generating meta tags is AnyBrowser.com http://www/anybrowser.com/MetaTagGenerator.html

The header tags

HTML supports six levels of headers, numbered <H1> through <H6>, with <H1> being the largest and most prominent, and <H6> being the smallest. Do include the keyword phrases in these tags, especially the first top-level tag (whether it is <h1> or <h2>.

The body tag

The portion of the Web page or the text that is visible in the browser is contained be-

tween the <body> tags. This text should be both descriptive and rich in keywords.

The alternate text tags

The <IMG> tag allows for inserting an image or graphic into a Web document and one of the properties available with this tag is the ALT text tag. The Alt text tag allows text to appear in place of an image. This is important not only for users who have non-graphical browsers but also because many Web sites have pages that are primarily graphics. As a result, there is little text that crawlers can use to index the site. They cannot determine what the pictures mean. However, if using the ALT tag to include descriptive remarks concerning the purpose of the site and some good keywords, the site get a better placement in the index database.

Example: Normal image tag:
<IMG SRC="InflightTrainingLogo.gif">

Image with alt tag and description
<IMG SRC="InflightTrainingLogo.gif" ALT="Inflight Training Solutions – Training for the **Corporate** Flight Attendant">

5. ADDITIONAL WEB DESIGN CONSIDERATIONS

This section will focus on what to avoid when considering indexing spiders.

Little or no body text or content
When there is little or no body text on a Web page, some crawlers will take this to mean that there is no content on the website worth indexing.

Large file size

This refers to the total size of the HTML code. If a Web site has a great amount of text, the file size could be very large. An indexing spider may give up if the site takes a long time to read. This is another reason why it is a good idea to break up large amounts of text to separate pages.

URL redirection

When directing a Web site's URL to another Web host, a spider may not be able to follow the re-direct to the destination page and may not index the page. If a re-direction is necessary, it is essential to create a spider optimized doorway page on the index page of the site that contains the re-direct.

JavaScript/Cascading Style Sheet

JavaScript and Styles are a jumble of coding that crawlers do not understand and can have an effect on indexing. Although both are widely used in site design, it is something to consider if there is difficulty indexing the site. It is recommended to place the script and the styles in a separate text file external to the Web page and then link it into the page.

Flash
In Web design, Flash can be a fantastic element. However, Flash can also be destructive to spiders. If using Flash, a consideration may be to create an entrance page without flash that is crawler optimized.

Fancy navigation
A fancy navigation structure can make it difficult for a spider to transverse other pages of the website.

Free Web hosting
A free hosting service for the website will often place the site on the same server as many other websites with all of them using the same IP address. With a Web site having the same IP address as other sites already in an indexing database, the index may not accept the site, or the site may be slow or inaccessible because another website on the server is controlling all the resources. Furthermore, if a Web site has the same IP address of a site banned from an indexing service, the indexing service may also ban the site. It is worth it to acquire a unique IP address and a reliable hosting service.

6. CLASSROOM IMPLEMENTATION

This search index optimization topics previously presented are implemented in an undergraduate and graduate course at Robert Morris University titled HTML/Internet. The first half (based on a 14 week semester) of the course is traditional lecture/discussion/and hands-on and the remaining class time is based on a project-based format.

The lecture/discussion/hands-on portion teaches the student HTML syntax. It is textbook driven and covers the anatomy of a Web page as well as the tags necessary for hyperlinks, colors, graphics, tables, and forms. Also discussed are Cascading style sheets and elementary JavaScript concepts. Students received a lecture on the tag and then are required to complete an exercise at the end of the chapter.

The second portion of the class is devoted to having the each student combines the major principles of the systems development life cycle (SDLC) to develop a personal web site and post it to a Web server. The site requires a home page and at least four hyperlinks to content pages.

Before the students begin working on their personal site, they are given an overview of search indexes including the popular search indexes and the components of the search indexes as discussed previously in this paper. Additionally, a class is devoted to visiting various search indexes and determining the steps involved in submitting a site to that particular index. With this accomplished, each student is required to create a keyword and phrases list for their site. The Web resources for selecting a keyword list mentioned previously in this papers (3. ELECTING KEYWORDS AND PHRASES) can be used. The instructor reviews this keyword list and when approved, the student may create the page layout design for their home page. Items such as site color scheme, graphics, text, and navigation considerations are determined so that the page is appealing and compelling to a target audience. Additionally, in order to assure that the page is optimized to work with a crawler-based search index, selected keywords are incorporated from the student's list with attention given to the title tag, header tag, and image tags. Included are also the keyword, description, and ROBOT metatags. When the page is completed, the code of each student's home page is checked by the instructor to assure that all tags are optimized as well as submitted by the student to either WebPositionGold or GRKda Keyword Density Analyzer to check the key word density.

The points discussed in section 5. ADDITIONAL WEB DESIGN CONSIDERATIONS of this papers are also addressed. For example, if the student wishes to use JavaScript or cascading styles sheets in their site, they must place the script and the styles in a separate text file external to the Web page and then link it into the page. Additionally, if the student wants to incorporate Flash, an entrance page without flash that is crawler optimized must be included.

The remaining pages of the site are then constructed and checked for optimization, as was the home page. When the site is completed, it is posted to the server at Robert Morris University or to a hosting service of the student's choice.

7. CONCLUSION

This paper discusses design concepts related to OPTIMIZING a Web site for crawler-based search indexes in order to attract worldwide interest or gain "hits" to the Website. After attracting new visitors, the next step is to keep those "hits" continuing. One way to assure this is to keep the Web site updated on a regular basis. Another way is to keep up to date with Web services. Finally, be sure to include an e-mail address on all Web pages and encourage visitors to send any comments or suggestions.

8. REFERENCES

All About Search Engine Robots andSpiders. (2002) [Online].http://www.searchtools.com/robots

Felke, Terry A, 2003, Web Developer Foundations. Scott Jones Publishers, El Granada, CA.

Fotheringham, John A. 2003, Search Engines Robots. [Online] http://www.jafsoft.com/searchengines/webbots.html/

Lehnet, W, 2001, Web 101: Making the Net Work for You. Reading, MA: Addison Wesley. University Press.

Marckini Fredrick W. 2001, Search Engine Positioning. Republic of Texas Prining.

Nobels, Robin and Susan O'Neil, 2001, Streetwise Maximize Web Site Traffic: Build Web Site Traffic Fast and Free by Optimizing Search Engine Placement. Adams Media Corporation

Pitkow, J. and C. Kehoe, 1997, "Results from the Seventh WWW User Survey", [Online]. http://www.gvu.gatech.edu/user_surveys/survey-1997-04/bulleted/use_bullets.html

Sekhar, Chandra, 2001, Internet Marketing and Search Engine Positioning – A " Do it Yourself" Guide. Southern Pub Group.