Search engine optimisation: Application to an academic digital library

S. Dhansay and M. Weideman
e-Innovation Academy, Cape Peninsula University of Technology

Abstract
The purpose of this research was to identify and evaluate techniques used to ensure a high listing as well as building up a high rank for an academic digital library on Google. The first step was to identify different techniques. They were implemented and their impact on listing and ranking measured. The techniques that were examined and evaluated include metadata tags, number of pages, impact of in-links, length of time the domain has been available and the use of directories. The effects of these techniques were measured using the scoring of the PageRank algorithm as used by Google. Certain limitations were encountered including the limited space of time to build up directory listings and the frequency of Google crawling websites, as this has a direct influence on the results. The research conducted provides evidence that using keywords in the right position as well as gaining links from websites that contain relevant and quality content improves the listing on Google.

Keywords
Search engine optimisation, Google, academic digital library

INTRODUCTION
As the Internet has grown, the importance of being listed in a top position on a search engine has become increasingly important in drawing visitors to a particular website. According to Sullivan (2006), Internet marketing research company comScore states that 6.4 billion search engine queries were made by United States users during March 2006. According to
McCarthy (2006), no longer do users rely on sites with links but prefer to use search engines to get to a particular page. This particular style of surfing the Internet has surpassed direct link navigation via a website.

When a user enters a query, the search engine could potentially return millions of results. This could produce a problem for websites not listed near the top since 73% of users do not look further than the first page (Jansen and Spink, 2006). This highlights the importance of being listed as high as possible on a search engine result page.

It was therefore considered important to find techniques to assist search engines in indexing and ranking of websites.

The author plans to identify the different techniques in order to optimise an academic digital library and plans to take an existing academic digital library and make recommendations on what techniques to implement to gain a higher ranking.

LITERATURE SURVEY

PageRank

PageRank is an algorithm used by Google to rank webpages (Vise and Malseed, 2005). There are over 200 factors used by Google to allocate a webpage its ranking (Evans, 2007). Its importance is highlighted by Google (2002) who states that “The heart of our software is PageRank™.”

Thelwall (2002) states that the fundamental concept of PageRank is that the more links to a webpage, the more useful its content is. The author also claims that the original PageRank algorithm has been changed since its inception but is still a core factor used by Google. This is confirmed by Google (2002) who states ‘we have dozens of engineers working to improve every aspect of Google on a daily basis; PageRank continues to play a central role in many of our web search tools.’

Once a set of results is returned from a search query, the level of importance is computed by PageRank. The higher the PageRank figure on a scale of 1 to 10, the higher it will be listed (De Jager, 2005). These findings are confirmed by Evans (2007) who, after analysing results from a search engine optimisation competition, found that the top ten webpage’s PageRank ranged from a PageRank of 4 upwards. The lowest PageRank in the top 40 was 4. The author concludes by saying that PageRank is an important indication; however, a high PageRank will only increase the probability that your webpage will rank high as there are other factors taken into account by Google.

TECHNIQUES

Fishkin and Pollard (2007) conducted a study whereby 37 SEO experts were invited to rate and vote what they considered to be the most important factors in optimising a website. These factors are listed in order of priority:

- title tag,
- anchor text,
- global link popularity,
age of site (domain),
internal link popularity,
content matter in relation to keyword usage,
links from websites about the same topic,
keyword use in text,
global link status of linked site, and
rate content relationship links.

TITLE TAG
Making use of the title tag is important but not obvious to web designers (Dawson and Hamilton, 2005). There are three important reasons for this:
- Google’s search algorithm gives them a higher value,
- users tend to notice the words in the title tag and recognize their importance, and
- the title tag is the default name for bookmarks in web browsers.

The use of the title tag should be regarded as the single most important entry point. Google only displays the first 60 characters in a title tag (Dawson and Hamilton, 2005).

According to Noruzi (2007), the title tag is important as search engines make this information available to users and reduce long titles. The author states that use of keywords at the start of the title is important. It should also be noted that search engines will penalize websites that abuse the title tag by repeating words. Search engines see the title tag as the most important positioning tool (Noruzi, 2007).

The same author suggests that when a search query is made, the search engine looks at the title first and once a keyword is found the search engine then ranks the pages accordingly. The users will also see the content of the tag. The author conducted a pilot study and the results indicated that Google shows on average 54 characters, with a maximum of 59 characters. It should be noted that words truncated from the title are still searchable and retrieved. The author recommends a title of no more than 10 words and not exceeding 60 characters. Important keywords need to be placed at the beginning of the title tag.

ANCHOR TEXT
Anchor text is the wording associated with a link on a webpage. If a webpage links to an academic library and the wording associated is ‘Click here to visit Mancunia Academic Library’ with ‘Mancunia Academic Library’ containing the link then this technique is being used to its full advantage (Swart, 2006).

The popularity of the use of anchor text stems from the fact that the anchor text provides a summary about the subject matter being linked and the queries that potentially users of a search engine would use (Eiron and McCurly, 2003). A study indicated that there is a link between anchor text and a search query being used.

According to Brin and Page (1998), the co-founders of Google, their PageRank algorithm treats anchor text in a special way. The authors state that other search engines associate the link that the page is located at but Google associates it with the page it links to. The authors
believe that this has advantages such as the page on which the link resides provides a more accurate assessment than the actual page, and the use of anchor text in the algorithm yields better results.

Brin and Page (1998) state that they use anchor text when retrieving a website using a search query. An example is given whereby the first result returned for the search query ‘Bill Clinton’ did not have a title and they state that the result relied on anchor text to be a reliable way of determining if the link was worthy of being listed.

KEYWORDS
According to Banks (2006), one way of achieving a high listing on Google is to carefully choose and place keywords related to the content of the page. Keywords need to be identified and then placed accordingly. The recommended number of keywords is three to five words per page. The same author claims that the closer the keywords are to the top left of the page, the more weight Google gives to them.

The placement of keywords is important and various placements on the page account for different weights (Kiritzinger, 2007). Meadhra (2004) states that the placement of keywords in the following locations indicates the highest to lowest in sequence of priority:
- domain name,
- page title,
- headings,
- body,
- meta tags,
- links, and
- alt: text.

Banks (2006) claims similar results by stating that search engines place increasing emphasis on:
- titles,
- headings, and
- emphasised text.

The author also states placing the keywords in the body helps. These techniques will be further examined and explained:

Headings
The ‘heading’ tag defines different levels for headings. Figure 1 is an example of the six levels that can be used with its HTML output.

Figure 2 indicates that <h1> will display the largest text moving downwards in size with <h6> displaying the smallest text (W3C, 2007). However, it is difficult for web designers to deceive search engines as HTML maps closely to how the content is displayed – certain search engines do place more weight on the <h1> tag and, therefore, it should contain primary keywords.
Alt
According to the W3C (2007) the alt attribute is used in a set of tags like img, area, input and applet. The alt attribute allows alternative text to be displayed if the object cannot be displayed like an image. They recommend using the tag so that it can be indexed so that search engines will not miss information on your page.

Domain Name
According to Dawson and Hamilton (2005), using a well-established domain has an advantage if the domain has existed for years and is already known to Google. Evans (2007) states that domain age is seen as an important factor and results gathered from a study suggest that there is truth in this.

A search on Google for the word ‘academiclibrary’ automatically suggests to the user to change the search query to ‘academic library’. This would indicate that Google has the technology to able to detect invalid concatenations and can also determine how to separate words. It should also be noted that the word ‘academiclibrary’ was bold in the URL displayed on the results page. The domain http://academiclibrary.btol.com/ was the first result returned and the domain http://www.academiclibrary.com/ is the ninth result returned. More interestingly, the domain http://www.awc.cc.az.us/library/ is placed in eighth position and the word ‘academiclibrary’ is not in the domain. This could be as a result of other search engine optimisation techniques being implemented.

META TAGS
Dawson and Hamilton (2005) state that the value attributed to meta tags is not high as they are open to abuse. Google tends to gather its own summary of the webpage and displays it in the search result. To improve what is displayed they recommend placing a description near the top of the page and Google will include it in the abstract displayed on the results page.

FIGURE 1: HTML code to display headers (below)

FIGURE 2: Figure 1 as displayed by a web browser (right)

```html
<h1>This is header 1</h1>
<h2>This is header 2</h2>
<h3>This is header 3</h3>
<h4>This is header 4</h4>
<h5>This is header 5</h5>
<h6>This is header 6</h6>
```
They do suggest (with evidence) that Google does interpret metadata from certain trusted sites with domain names that have ‘gov’ and ‘edu’. One of the biggest problems should a metadata approach be introduced, is who will create the metadata. It is clear that this technique is open to abuse which is why it is not trusted by Google in general with the exception of trusted sites.

LINKS
The list compiled by Fishkin and Pollard (2007) contains five factors that relate to links. This emphasises the importance of these different factors:

- global,
- internal,
- same topic,
- status of linked site, and
- rate of links.

An internal link connects a webpage within the same website. A website representing a high internal link structure together with relevant and high quality page content can improve the ranking (Visser et al., 2006). De Jager (2005) claims that internal linking can have considerable impact on a website’s PageRank and is viewed the same as external links.

External links connect a webpage to another that is hosted on a different domain. According to Visser et al. (2006), external links provide more of an interest to search engines as they tend to indicate worth and credibility of the website’s content. The authors also suggest an increased number of links can improve the ranking and is known as link popularity.

According to Thelwall (2002), PageRank implements the principle that pages with high quality content are more likely to be linked than pages that represent low quality. The author states that it is recognised that increasing the number of external links will increase its rank in search engines. It is impossible for an algorithm to guess what is the most trustworthy page from a set of results so Google does this based on external links as this would indicate a well respected page. A study conducted by Thelwall (2002) on links suggested the following:

- a home page will most likely get the highest ranking,
- the website’s link structure should be shallow,
- both internal and external links should be coded in standard HTML so that they are more likely to be picked up by search engines, and
- place the most important links on the home page as the home page will most likely have a higher PageRank.

De Jager (2005) stated that it is important that external links come from websites that have related and relevant content. This can be done by conducting a search of keywords relevant to your website’s topic and requesting links from the websites returned. Requests for links may not always be accepted as these websites may be your competition.

One way of gaining external links is to be listed in web directories such as DMOZ, Yahoo and del.icio.us. Evans (2007) stated that being listed on DMOZ is difficult. It can take anywhere between six months and a year to be listed due to human volunteers deciding on
whether a website should be listed. The author concluded by saying that this technique is used by successful SEOs.

RESEARCH METHOD
Two websites have been chosen in order to prove the validity of the techniques discovered and explained in the literature review. The search query ‘Academic Digital Library’ was specified on Google and from the results returned, the first website was chosen.

A second poorly designed website with regard to search engine optimisation was chosen.

The reason for having chosen a highly listed and poorly designed website is to determine which techniques the websites implemented and if these techniques have an effect on a high listing on Google.

Recommendations will then be made for the poorly designed website on how to improve its listing on Google.

In order to measure and evaluate these techniques, specific tools will be used:
- Google toolbar,
- Site link analyser,
- Google back links checker,
- Search engine rankings,
- Keyword density checker,
- IBP, and
- Lynx web browser.

These tools are freely available and help gather statistics in order to measure search engine optimisation techniques. The Google toolbar will be used to read off the PageRank value that Google assigns to the webpage. Site Link Analyzer provides statistics about website internal links plus the anchor text used. The Google Back Links Checker provides statistics on the number of external links to a particular webpage. Search Engine Rankings provide the listing position on a search engine for a given specific keyword phrase. It also checks directory listings. The keyword density checker provides a count of the use of the most common words on a webpage. IBP is a search engine optimization suite providing various options. In the research conducted the rank checker was used from this suite. All these tools provide real time data.

FINDINGS
A search was conducted on Google for ‘Academic Digital Library’ and the first result returned was a website named Libweb at http://lists.webjunction.org/libweb/. This website is a global directory for libraries from around the world.

A second website chosen was Caltech Library Services (California Institute of Technology) at http://library.caltech.edu/. This is an online digital library and classified by Libweb as an academic library. A search was done for the given phrase ‘Academic Digital Library’ and the website was not listed within the first 500 results by Google.
TABLE 1: A comparison of techniques

<table>
<thead>
<tr>
<th></th>
<th>Libweb</th>
<th>Caltech Library Services</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://lists.webjunction.org/libweb/">http://lists.webjunction.org/libweb/</a></td>
<td><a href="http://library.caltech.edu/">http://library.caltech.edu/</a></td>
</tr>
<tr>
<td><strong>Listing</strong></td>
<td>1</td>
<td>Not listed in first 500 results</td>
</tr>
<tr>
<td><strong>PageRank</strong></td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td><strong>Title</strong></td>
<td>Libweb – Library WWW Servers</td>
<td>Caltech Library Services</td>
</tr>
<tr>
<td><strong>Keywords</strong></td>
<td>The most common words used include libraries, library, maintained, Libweb, Germany, servers</td>
<td>The most common words used include library, Caltech, engineering, web, access,</td>
</tr>
<tr>
<td></td>
<td>Use of the word 'library' in a &lt;h2&gt; tag</td>
<td>Use of the word 'library' in a &lt;h3&gt; tag</td>
</tr>
<tr>
<td></td>
<td>There is an internal link for the anchor text Academic Libraries'.</td>
<td>Main headings are images with alt text used</td>
</tr>
<tr>
<td><strong>Domain</strong></td>
<td>No keywords in domain</td>
<td>The keyword library is used in the domain</td>
</tr>
<tr>
<td><strong>Meta tags</strong></td>
<td>The keywords and description meta tags are being used</td>
<td>No meta tags are used</td>
</tr>
<tr>
<td></td>
<td>The keywords chosen include Libweb, libraries, library home pages, library directories</td>
<td></td>
</tr>
<tr>
<td><strong>Internal links</strong></td>
<td>17</td>
<td>110</td>
</tr>
<tr>
<td><strong>External links</strong></td>
<td>1400</td>
<td>725</td>
</tr>
<tr>
<td><strong>Directory listing</strong></td>
<td>Not listed</td>
<td>Libweb</td>
</tr>
<tr>
<td><strong>delicious bookmarks</strong></td>
<td>227</td>
<td>47</td>
</tr>
</tbody>
</table>

**DISCUSSION OF RESULTS**

Using the phrase ‘Libweb’ as a search query, it is listed as number 1 on Google. This is an advantage as traffic will be generated as a result of this. Caltech is not listed in the first 500 results for the given phrase.

Both websites have a high PageRank value. This indicates that Google rates both sites as having quality content.

The title tags of both sites are under 60 characters as recommended. They both contain the word ‘library’ but not ‘digital’ or ‘academic’. Caltech’s title tag is self-explanatory but Libweb’s is not at first glance. The literature review highlighted the importance of using this technique. Neither of the websites use repeated words, which mean they will not be penalised by search engines that would think that spamming techniques are being used. A high value is put on words at the beginning with Caltech preferring to use its university name. A search on Google for the keyword ‘Caltech’ lists this specific website in tenth position.

Caltech opted to use images as its main headings with alt text being used. This makes the wording used on these images irrelevant to Google as it cannot interpret the text being used in these images. Even with alt text being used, it does not have the same effect as using emphasised text or heading tags. This already puts the website at a disadvantage. Libweb opted for a text-based approach using various fonts and weights. Both websites make use of
the header tag in relation to the word 'library'. Libweb uses it in a <h2> tag whereas Caltech uses it in <h3> tag.

The Lynx web browser is text based and provides an indication of how web crawlers see a webpage. The page is readable and understandable to a web crawler. The headings are displayed with the data and links following in a logical order. The main heading's alt text is seen and displayed by the browser.

While conducting a keyword density test Caltech's most popular keywords were library, Caltech, engineering. [ - square brackets, web and access. Table 2 lists the top ten keywords used. Libweb's most popular words were libraries, library, maintained, Libweb, Germany, and servers. It is clear that Libweb places a high value on the word 'library', both singular and plural in its body text whereas Caltech does not. Table 3 lists the top ten keywords used. This could be attributed to the fact that Caltech lists images as headings as opposed to text.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Count</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>library</td>
<td>16</td>
<td>4.36%</td>
</tr>
<tr>
<td>caltech</td>
<td>8</td>
<td>2.18%</td>
</tr>
<tr>
<td>engineering</td>
<td>8</td>
<td>2.18%</td>
</tr>
<tr>
<td>[]</td>
<td>8</td>
<td>2.18%</td>
</tr>
<tr>
<td>web</td>
<td>7</td>
<td>1.91%</td>
</tr>
<tr>
<td>access</td>
<td>6</td>
<td>1.63%</td>
</tr>
<tr>
<td>pdf</td>
<td>6</td>
<td>1.63%</td>
</tr>
<tr>
<td>science</td>
<td>5</td>
<td>1.36%</td>
</tr>
<tr>
<td>search</td>
<td>4</td>
<td>1.09%</td>
</tr>
<tr>
<td>databases</td>
<td>4</td>
<td>1.09%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Count</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>libraries</td>
<td>10</td>
<td>4.35%</td>
</tr>
<tr>
<td>library</td>
<td>8</td>
<td>3.48%</td>
</tr>
<tr>
<td>maintained</td>
<td>8</td>
<td>3.48%</td>
</tr>
<tr>
<td>libweb</td>
<td>5</td>
<td>2.17%</td>
</tr>
<tr>
<td>germany</td>
<td>5</td>
<td>2.17%</td>
</tr>
<tr>
<td>servers</td>
<td>4</td>
<td>1.74%</td>
</tr>
<tr>
<td>public</td>
<td>3</td>
<td>1.3%</td>
</tr>
<tr>
<td>america</td>
<td>3</td>
<td>1.3%</td>
</tr>
<tr>
<td>www</td>
<td>3</td>
<td>1.3%</td>
</tr>
<tr>
<td>university</td>
<td>3</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

Libweb opted to use the keyword and description meta tags whereas Libweb totally disregarded those tags. It is however recommended that these are included as other search engines or directories might use them. Dawson and Hamilton (2005) do indicate that Google trusts data from certain websites with a domain 'edu' as in Caltech's URL.

The importance of links highlighted in the literature review reflects what appears in the results. Libweb has just less than double the number of external links Caltech has. Google uses this as an indicator to rate the web page's quality. The link structure of both sites is consistent in use of standard html links, with the home page being an important central point providing links to other web pages and subsections.

LIMITATIONS OF THE STUDY

The results retrieved using the various tools were taken at a specific time and are subject to change due to changing data both externally and internally with regard to both websites.
CONCLUSIONS AND RECOMMENDATIONS
When the findings of the literature review are combined with the results from the research conducted there are two important aspects with regard to search engine optimisation.

The first aspect is keywords, as related to the target market of the website. The recommended number is three to five words. Once these words are established they would need to be strategically placed in various places as recommended by Meadhra (2004):

- domain name,
- page title,
- headings,
- body,
- meta tags,
- links, and
- alt tags.

The top five keywords used by Caltech as found by using the keyword density tool are:

- library,
- caltech,
- engineering,
- [ , and
- web

The keywords library and Caltech do seem relevant to the website. However the use of square brackets and ‘web’ do not seem to have any relevance with regard to identifying a specific target market. To a lesser extent the keyword ‘engineering’ is not relevant unless the website is targeting engineering academics. The digital library caters for all disciplines and not just engineering. The use of keywords such as academic and digital are recommended.

The top two keywords are used in the title. The full use of headings is not being made use of with just one reference in a <h3> tag. It is recommended that all keywords be used in header tags. The main image depicting the header makes use of the alt tag using the exact keywords used in the image. Other images on the page do not use any alt tags, including one with the Caltech name on it. It is recommended that these images be given descriptive names by making use of the keywords.

The domain name is a sub domain of the Caltech website with the word being used ‘library’.

It is clear that the webpage title tag plays an important aspect and search engines seem to emphasise this. Placing keywords in the body is important. There is no evidence to suggest that Google trusts meta tags from certain domains. It is however recommended to include them as other search engines might make use of them. If images are used then alt text needs to be descriptive so that search engines can pick up keywords. These techniques are vital as they assist Google in categorising a website.

The second aspect is proving that the content represents quality and the users of Google will find the information useful. This can be achieved by building a collection of internal and external links. The type of links highlighted by Fiskin and Pollard (2007) include:
• global,
• internal,
• same topic,
• status of linked site, and
• rate of links.

A comparison of external links from the findings shows that Libweb has double the number of external links that Caltech has. Therefore, Google sees Libweb as representing information with higher quality. This is just one part of links that is important. The literature review highlighted the importance of anchor text and how Google uses this to categorise websites.

In order for Caltech to improve its listing the following recommendations are made.

To prove that Caltech’s content represents quality, more external links need to be established. Directories need to be used as this specific technique is used by successful SEOs. It is also important that these use relevant and meaningful anchor text. Partnerships need to be established with websites in a similar field to that which this website represents. It is also important that all websites use relevant keywords when using anchor text.

A clear difference with regard to keyword usage has been highlighted and this would need to be improved if a digital academic library target audience has been identified by Caltech. It is clear that correct usage of these keywords will have to be implemented if Google is to use them when retrieving results.

The importance of using search engine optimisation techniques is important as a high listing on a search engine yields traffic to a website. Correct keyword usage and placement along with content quality links will ensure that this happens.

APPENDICES/ANNEXURES

Due to space restraints the appendices/annexures could not be included. These can be obtained by contacting the editor.

REFERENCES


Thelwall, M. 2002. ‘Can Google's PageRank be used to find the most important academic Web pages?’ *Journal of Documentation*, 99(2): 205–217.

