Keyword stuffing and the big three search engines

Herbert Zuze
Faculty of Business, Cape Peninsula University of Technology, Cape Town, South Africa, and
Melius Weideman
Faculty of Information & Design, Cape Peninsula University of Technology, Cape Town, South Africa

Abstract

Purpose – The purpose of this research project was to determine how the three biggest search engines interpret keyword stuffing as a negative design element.

Design/methodology/approach – This research was based on triangulation between scholar reporting, search engine claims, SEO practitioners and empirical evidence on the interpretation of keyword stuffing. Five websites with varying keyword densities were designed and submitted to Google, Yahoo! and Bing. Two phases of the experiment were done and the response of the search engines was recorded.

Findings – Scholars have indicated different views in respect of spamdexing, characterised by different keyword density measurements in the body text of a webpage. During both phases, almost all the test webpages, including the one with a 97.3 per cent keyword density, were indexed.

Research limitations/implications – Only the three biggest search engines were considered, and monitoring was done for a set time only. The claims that high keyword densities will lead to blacklisting have been refuted.

Originality/value – Websites should be designed with high quality, well-written content. Even though keyword stuffing is unlikely to lead to search engine penalties, it could deter human visitors and reduce website value.

Keywords Keyword stuffing, Spamdexing, Search engines, Webpage, Ranking, Web sites, Visibility

Paper type Research paper

Introduction

The increasing use of web sites as marketing tools has resulted in major competition for high rankings amongst web sites. Furthermore the openness of the web facilitates speedy growth and success but also leads to many web pages’ lack of authority and quality (Castillo et al., 2011). Visser and Weideman (2011) noted that the “value” of a web site can be determined by the page on which it ranks. For this reason search engine optimisation (SEO) has become a priority for most web sites, especially those with commercial intent.

Web sites and web pages are blacklisted (Yung, 2011) and excluded from search engines’ (SEs) indices due to “black hat” (unethical) techniques employed on them. One of these techniques is keyword stuffing: the practice of repeating important keywords on a web page to the point where the excess goes against the grain of readable English text (Weideman, 2009). Yung (2011) also mentions that SE share the list of blacklisted web sites, which means that if a web site is blacklisted by one SE, others will follow suit.

The purpose of this study was to provide experimental evidence of the way SEs view keyword stuffing. Various sources, ranging from scholars to SEO practitioners,
indicated different views on this phenomenon, characterised by different levels of keyword density in the body text of a web page being considered acceptable.

Research objectives
This study had the following objectives:

- to identify the time Google, Yahoo! and Bing each take to index a web page;
- to investigate the reactions of Google, Yahoo! and Bing when indexing web pages with varying keyword density; and
- to determine the level at which keyword density in a web page is regarded as keyword stuffing by the SEs.

Research questions
This study is based on the following research questions:

- Approximately how long do Google, Yahoo! and Bing take to index a web page?
- Does keyword density affect web page indexing time?
- What is the view of SEs and academic experts on the definition and understanding of keyword stuffing?
- How can “keyword richness” and/or “keyword poorness” be measured?
- How does a web developer know if web page content was interpreted as SE spamindexing?

Literature review
SEs
SEs have been defined as programs that offer users interaction with the internet through a front end, where users can specify search terms or make successive selections from relevant directories. The SE will then compare the search term against an index file, which contains web page content. Any matches found are returned to the user via the front end (Weideman, 2004).

When searching for information on the web, users specify keywords and/or keyword phrases by entering them into the search box. Both “link popularity” (quality and quantity of links) and “click-through popularity” are used to determine the overall popularity of a web site, influencing its ranking.

During this process of determining which web pages are most relevant, the SE selects a set of pages which contain some or all of the query terms, and then calculates a score for every web page. This list is then sorted by the score obtained and returned to the end-user (Egele et al., 2009).

SEs are used to find information on the web, whether relevant or not (Alimohammadi, 2003). The SE index is updated frequently either by human editors or by computerised programs (called spiders, robots or crawlers) (Weideman, 2004). The value of a web page’s content to users is inspected by SEs using several complex algorithms. SEs achieve this by making use of crawlers to identify keywords in the natural readable content of a web page (Ramos and Cota, 2004). Recent figures indicate that Google dominates the market, with Bing and Yahoo! a distant second: Google claims 66.4 per cent of the market, vs Bing and Yahoo! with a total of 29.1 per cent between them.
As a result, the researchers selected Google, Bing and Yahoo! to use for this study (Karma Snack, 2011).

SEO
SEO is a process by which both on-page and off-page elements of a website are altered in an attempt to increase the website's ranking on search engine result pages (SERPs). It has been proven that inlinks, keyword usage and hyperlinks are some of the most important on-page elements to be considered in this regard. However, according to Wilson and Pettijohn (2008) spamdexing techniques, such as repeating keywords several times to improve the rank of the page without adding value to the content of the page, are being employed by numerous savvy website developers. It is, however, claimed that SEs penalise the pages that appear to use these techniques. Inevitably legitimate pages are sometimes unduly penalised or completely removed from a SE index. Kimura et al. (2005) mentioned that there are often SEO spamdexing websites that contain little or no relevant content and whose sole aim is to increase their position in the SE rankings. This is particularly true with malicious website developers whose techniques are to achieve undeservedly high rankings by exploiting algorithms used by SEs (Shin et al., 2011).

Following the user search behaviour, the best measurement in terms of the number of words should be established: it should be sufficient to permit SEs a rich harvest of keywords, but not too many as this might scare off human readers (Visser and Weideman, 2011). Zhang and Dimitroff (2005) also found that websites need to have keywords appearing both in the page title and throughout the page body text in order to attain better SE results.

SEs analyse the frequency of the appearance of keywords compared to other words in the web page: generally speaking, the higher the frequency of a keyword the better the chances of the word being deemed relevant to the web page. When taken to the extreme, however, SEs could view it as spamdexing. Like several other scholars Zhao (2004) is of the opinion that a keyword density of 6-10 per cent in the body text of a page is the best for satisfying crawler demands. Todaro (2007) recommends that keyword density be kept to 3-4 per cent of the body text per page. The author further reiterates that the overuse of keywords in the body text raises a red flag with web crawlers and might disqualify the site from being indexed.

Using a keyword more often than SEs deem to be normal could be considered search spamdexing (Wikipedia, 2010). Charlesworth (2009) is of the opinion that keywords which appear twice in 50 words have a better ratio than the use of four keywords in 400 words. Appleton (2010) states that a benchmark of 3-7 per cent keyword density is acceptable and points out that anything more than this strays into spamdexing territory. However, Kassotis (2009) stated that for every 100 words in a web page, one keyword should be used. Kassotis (2009) further states that the general rule is a keyword density of 2-2.5 per cent. These claims indicate that scholars disagree on the ideal keyword density inside body text, as read by SE crawlers.

Web page indexing and ranking
The most important reason for designing a web site is to provide content that is relevant to what searchers or end-users are looking for (Weiss and Weideman, 2008). High-quality content enables a website to become reputable and referenced by other websites and ultimately attain a better position on the SERP. The implication is that if a website is not listed on the top half of the first page of results, it is virtually invisible.
as far as the average user is concerned (Chen, 2010; Weideman, 2008). Higher rankings mean more click-through and often, higher income (Zahorsky, 2010).

Crawlers primarily do text indexing and link following: if a crawler fails to find content or links on a site it leaves without recording anything. They constantly crawl the web, returning with new and updated pages to be indexed and stored. Erdelyi et al. (2011) found that recent results and crawlers’ visits have concentrated on the definition of new features, hence ignoring other important factors in the domain of machine learning techniques that affect SEO results.

**Web page indexing period**
Although there is no official rule that stipulates the time taken for a web site to be indexed, Apexpacific (2010) and Zaslavsky (2010) pointed out that major SEs take in excess of three weeks and even up to six weeks to index a web page. However, pay-per-inclusion engines usually index within two to seven days, depending on the payment plan. Moreover when discussing Google’s perspective, Mathews (2011) mentioned that it should take up to two weeks for a site to be indexed.

The researchers, as well as the interviewees during this research and several other scholars, including Zhang and Dimitroff (2005), found empirical evidence indicating that the indexing period for web pages is not fixed. This research will also attempt to measure the average time in days taken to index a given web page.

**Web site visibility**
According to Weideman (2009) visibility is defined as the ease with which a SE crawler can find a web page. After finding the information, it is then further defined by the degree of success the crawler has in indexing the page. Borchardt and Weideman (2008) stated that web site owners should concentrate on visibility since there is increasing competition for high rankings between web pages on the WorldWideWeb. A well-designed web page with high visibility can be easily found and has a substantial amount of relevant, easy-to-index information on the page to be indexed by visiting crawlers.

Web site owners invest resources in order to influence their online visibility (Berman and Katona, 2011): if the quality of sites corresponds with their estimation for visitors, then SEO assists as a mechanism that improves the ranking by correcting measurement errors.

**Web site usability**
Usability involves being able to easily and successfully interact with a web site. It further measures the extent to which a visitor can easily and quickly use web page resources. Usability also includes the following factors:

- ease of learning;
- subject satisfaction;
- ease of use;
- efficiency of use;
- memorability; and
- error frequency and severity.

The objective of usability is to eliminate any hindrances impeding the experience and process of online communication (Eisenberg et al., 2008). Visser and Weideman (2011)
are of the opinion that the inclusion of usability attributes will enhance conversion (i.e. visitors converting casual content views or web site visits into desired actions); therefore effective web site design should incorporate usability as a priority.

**Spamdexing**
Spamdexing has grown to become one way for SEO practitioners to attract the SE to quickly index and highly rank a web site (Abernethy et al., 2009). Due to the deterioration of content quality on the web, preventing spamdexing is a top priority for the SE industry (Abernethy et al., 2009; Henzinger et al., 2002). This is particularly true for the high value of top rankings for commercial web sites. SEs reward both qualitative and quantitative web sites with solid, informative and useful content with good rankings for specific search terms or phrases (Visser and Weideman, 2011). Spamdexing as a general classification for unethical SEO techniques includes keyword stuffing, cloaking, invisible text and many other techniques.

**Keyword density**
Malaga (2009) mentioned that keyword density measures the extent to which a certain word or phrase appears on a site or a web page. The ideal value in terms of the number of keywords expressed as a percentage of the total number of words should be established: it is advisable to provide SEs with a rich harvest of keywords, but not too many as this might discourage human readers (Visser and Weideman, 2011).

Weideman (2009) indicates that there should be a balance between the use of a keyword or phrase for both the user and the SE crawlers. According to Ricca et al. (2004) keywords have different scores in a web page, that is, more specific keywords weigh more and have a better score than other keywords, but as the keyword quantity increases so does the probability of keyword stuffing.

Google’s (2010) web master guidelines state that web pages should not be loaded with irrelevant keywords. However, they do not suggest any specific figures or percentages. Similar to Google, Yahoo! does not clearly specify its interpretation of keyword density.

**Summary**
Researchers differ on the ability of a web page to meet the correct design and submission procedures. Web site visibility seems to be a measurement which manifests in changing rankings for the same web page on the same SE. Keyword stuffing is not defined in exact terms by any of the large SEs, and it appears as if designers have to aim for an elusive midpoint between keyword stuffing and keyword-rich web page body text. Furthermore indexing time is unpredictable: this has become evident through the work of Borglum (2009), Weideman (2009), and Zhang and Dimitroff (2005). These scholars found differing values in the amount of time taken by SEs to index a web page; they identified a range of 1-90 days.

**Methodology**
The researchers implemented triangulation and gathered data via a combination of a literature review, personal interviews and the results of an empirical experiment.

**Interviews**
Interviews were conducted with five well-known SEO/digital marketing companies in Cape Town, based on a structured questionnaire.
Experimental research
Five new web sites were created, identical in style but with differently worded content. New domains were registered where these web sites were hosted, to exclude the possibility of some domains already having been indexed prior to the experiment. Consistent, relevant and similar content was placed on every web site to meet all the appropriate white hat regulations with the exception of keyword density. In each case the content related to the sales and support of second-hand laptop computers. The five respective homepages had increasing densities for one keyword: “laptops” (see Table I).

The web site homepages were submitted to Google, Yahoo! and Bing, within an hour of each other on the same day. A daily check was then carried out to establish when, if at all, these sites were indexed. This was done using specific operator searching, and by searching for specific content known to be present on the web sites. The dates of all these checks were recorded, from which the data listed elsewhere was drawn.

After collecting the experimental results, certain anomalies were noted regarding some theoretical claims in prior research. As a result the researchers decided that the first experiment was to be considered as a first phase (Phase 1) and to extend the work to include a second experiment (Phase 2). This enabled the researchers to design a second phase of the experiment with intentionally high keyword densities.

For Phase 2 the following were changed in order to reinvestigate the response of the SE crawlers:

- the keyword density of all the web sites was increased, while “laptops” remained as the main keyword; and
- a fourth page per web site was added (and all menu structures updated) to ensure that duplication between Phase 1 and 2 would not be possible.

To further ensure that crawlers would not view the content as being duplicated, all Phase 1 web site files were deleted from the hosting servers and Phase 2 files were uploaded. The submission to Google, Yahoo! and Bing followed the same procedure as Phase 1. For each one of the two experimental phases, a total monitoring time of 67 days was allowed. Table II lists the Phase 2 keyword densities.

During the Phase 1 experiment, five web sites providing information about second-hand laptop sales and accessories were designed in simple Hyper Text Mark-up Language (HTML) with no Flash files, frames, JavaScript or excessive graphics. This was done to ensure that crawlers would not blacklist these sites due to the use of black hat technologies. The research was centred on the homepage whilst the other pages had similar content and were intended to provide increased site content that would ensure that visiting crawlers would consider them to be real web sites with content of value. The web sites consisted

<table>
<thead>
<tr>
<th>Domain</th>
<th>Word count</th>
<th>Keyword count</th>
<th>Keyword density (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.getlaptops1.co.za">www.getlaptops1.co.za</a></td>
<td>330</td>
<td>13</td>
<td>3.94</td>
</tr>
<tr>
<td><a href="http://www.getlaptops2.co.za">www.getlaptops2.co.za</a></td>
<td>330</td>
<td>20</td>
<td>6.06</td>
</tr>
<tr>
<td><a href="http://www.getlaptops3.co.za">www.getlaptops3.co.za</a></td>
<td>330</td>
<td>28</td>
<td>8.48</td>
</tr>
<tr>
<td><a href="http://www.getlaptops4.co.za">www.getlaptops4.co.za</a></td>
<td>330</td>
<td>40</td>
<td>12.12</td>
</tr>
<tr>
<td><a href="http://www.getlaptops5.co.za">www.getlaptops5.co.za</a></td>
<td>330</td>
<td>90</td>
<td>27.30</td>
</tr>
</tbody>
</table>

Table I. Phase 1 web site keyword densities for the keyword “laptops”
of three pages each:

- the home page;
- the catalogue page; and
- the contact page.

These web sites are hosted at the following domains:

- www.getlaptops1.co.za
- www.getlaptops2.co.za
- www.getlaptops3.co.za
- www.getlaptops4.co.za
- www.getlaptops5.co.za

Web site submission. The web sites were manually submitted to Google, Bing and Yahoo! within an hour of each other on the same day, again to ensure that they all had the same opportunity to be exposed to SE crawlers.

Results and analysis

The data were analysed following the categories specified below:

- interviews with the SEO practitioners;
- literature review; and
- data collected from the five experimental web sites.

Google, Yahoo! and Bing analysis

The researchers noted with great concern that the argument behind the penalisation of a site or removal from the index of the respective SE was indicated by all the three SEs; unfortunately the extent to which the penalty was implemented was not justified. Each SE has its own guidelines but the three big ones seem to agree on not tricking the SE or the user. Harsh consequences for failure to adhere to the guidelines were mentioned by all three SEs, with the worst penalty being removal of a web site from the index. All three also briefly noted the penalisation issue and addressed it in an umbrella scenario without providing specific examples. Google went on to describe keyword stuffing under its quality guidelines, however, nothing was mentioned about the point at which a site is blacklisted or removed from its index for repeating keywords. Yahoo! and Bing did not go into detail about keyword stuffing even though they stated some points in passing as part of the techniques that must be avoided. The failure by all three SEs to address this issue might be a reflection of the possibility that their algorithms could fail to counteract the keyword stuffing problem.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Word count</th>
<th>Keyword count</th>
<th>Keyword density (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.getlaptops1.co.za">www.getlaptops1.co.za</a></td>
<td>330</td>
<td>100</td>
<td>30.3</td>
</tr>
<tr>
<td><a href="http://www.getlaptops2.co.za">www.getlaptops2.co.za</a></td>
<td>330</td>
<td>132</td>
<td>40</td>
</tr>
<tr>
<td><a href="http://www.getlaptops3.co.za">www.getlaptops3.co.za</a></td>
<td>330</td>
<td>170</td>
<td>51.52</td>
</tr>
<tr>
<td><a href="http://www.getlaptops4.co.za">www.getlaptops4.co.za</a></td>
<td>330</td>
<td>232</td>
<td>70.3</td>
</tr>
<tr>
<td><a href="http://www.getlaptops5.co.za">www.getlaptops5.co.za</a></td>
<td>330</td>
<td>321</td>
<td>97.27</td>
</tr>
</tbody>
</table>

Table II.
Phase 2 web site keyword densities for the keyword “laptops”
Interview results
All the interviews were conducted in September 2010 and the five interviewees had at least seven years of experience in the SEO industry. The interviewees held strategic positions within their companies and their positions were SEO expert, SEO strategist, CEO, SEO manager and SEM manager. The information gathered was therefore based on experience rather than academic literature. Four of the interviewees had previously been involved in web site design and marketing and the diversity and exponential growth of the industry led them to venture into SEO. Their responses showed a high level of tactical maturity and understanding of SEO methodologies and the evolution of the industry. However, the interviewers wished to remain anonymous. The interviews included questions about their personal experience in the SEO industry, their understanding of this industry and their knowledge of keyword density and spamdexing. The questions were:

- How many years have you been in the SEO industry and why did you consider being in this field?
- What is your understanding of keyword stuffing?
- How often do you carry out experiments to check whether SEs’ indexing algorithms have changed?
- How do you measure the richness or poorness of a keyword in the body of a web page?
- How do Google, Yahoo! and Bing interpret keyword stuffing and do their algorithms adhere to their respective document guidelines and procedures?
- Do you think SEO practitioners and web site developers understand spamdexing?
- Do you think there is a crossover point from keyword-rich text to spamdexing and how do you interpret it?

After the interviews were concluded, the interviewees were furnished with copies of the test web sites (Phase 1) to enable them to assess whether or not the respective web sites would be indexed. For brevity’s sake, domain names will be abbreviated as GLPS1, GLPS2, etc. from this point onwards. Figure 1 indicates that all the interviewees expected the GLPS1, GLPS2 and GLPS3 sites to be indexed. Among the interviewees two were unsure whether GLPS4 would be indexed whilst the remaining three agreed that this web site would be indexed. None of the interviewees differed about GLPS5 in terms of indexing; they all agreed that the site would not be indexed.

The research established that a keyword density of 5 per cent was supported by three of the interviewees as the maximum keyword density level that is acceptable for both humans and crawlers. Nonetheless one of the interviewees felt that 3 per cent is the best keyword density and above this mark they considered the level to be unacceptable to the end-user. These two figures became interviewees’ measurement of a keyword-rich web site whilst a web site with a keyword density below 3 per cent was considered poorly optimised. A web site with a keyword density above 5 per cent was considered to indicate keyword stuffing. However, one of them saw 12 per cent as being the crossover point to spamdexing, whilst the remaining interviewees did not say what they considered to be the crossover point to keyword stuffing.
Phase 1 indexing results

Web page indexing predictions by interviewees

Figure 1.
Web page indexing predictions by interviewees

Notes: 1=Will not be indexed; 2=Not sure; 3=Will be indexed

Experimental results analysis
After submitting the five homepages to Google, Yahoo! and Bing, each day’s indexing results were recorded (see a snapshot of each in Figures 2 and 3). They were checked using the following methods:

- a string search;
- a site search; and
- the Webmaster Tools (SE analysis for each registered web site) were used to constantly check whether the SE’s crawler did visit a site.

Colour coding was used to represent indexing status for each day in both figures. Red means “not indexed”, green shows “indexed” and blue indicates “cloaked”. The basic expectation, based on the results of the literature survey and the interviews, was as follows:

- GLPS1, 2 and 3 should be indexed by all three SEs, after a period of a few weeks;
Phase 1 indexing results

Phase 1’s shortest indexing waiting time was five days and the longest was 33 days. After recording the indexing results for 67 days, all the web site pages were successfully indexed with the exception of GLPS1, which was not indexed by Google. The research showed the fifth web site as the most favoured one: it was indexed first by Yahoo! and Bing four days after submission. After five more days Yahoo! and Bing registered the remainder of the web sites, including even GLPS5, despite having the highest keyword density of 27.3 per cent.

With respect to Google, Yahoo! and Bing, this research proved that indexing time is not affected by keyword density, given that both experimental phases showed all web sites with their varying keyword densities being indexed by Yahoo! and Bing. However, during the second phase of the experiment, the study showed only one out of five web sites being indexed by Google: this does not prove that the higher keyword density was the factor resulting in other web pages being not indexed.

Table III shows the homepage indexing time in days, recorded over a period of 67 days.

There was a noteworthy anomaly: an Iranian web site, selling DVDs and books, appeared to have copied some content from GLPS1. Their web site appeared high upon a Google SERP, and also showed the GLPS1 text in question on the SERP. However, when visiting this site, the paragraphs of text were invisible to the human user. This implies that the Iranian web site not only scraped content from the GLSP1 web site, but then also used cloaking to show this text to a visiting crawler, but hide it from a human visitor.

Phase 2 indexing results analysis

Previous research has indicated that SEs prefer indexing regularly updated web pages, and that there is no clear updating pattern (Lewandowski, 2008). For example it could

<table>
<thead>
<tr>
<th>Phase 1 website homepage indexing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLPS1</td>
</tr>
<tr>
<td>Google</td>
</tr>
<tr>
<td>Yahoo!</td>
</tr>
<tr>
<td>Bing</td>
</tr>
</tbody>
</table>

Notes: GLPS, getlaptops; NI, not indexed
happen that a web page is indexed today, but measurement of indexing frequency starts tomorrow, leading to a longer waiting time for the next visit. This was catered for in the current research through waiting for a period of 67 days, which exceeds all previously recorded crawler waiting times. The limitation of this method is that some waiting times could appear to be longer than they should be, however, measuring the indexing time was not the main purpose of this research.

Phase 2's shortest waiting time was 19 days and the longest was 29 days. During the Phase 2 experiment Bing and Yahoo! indexed all five sites, including the one with the highest keyword density of 97.3 per cent. Google only indexed one of the five sites, with a keyword density of 40 per cent. After 67 days all results were recorded and the experiment was terminated.

Table IV shows the homepage indexing time in days recorded over a period of 67 days.

**Phases 1 and 2 indexing analysis**

In conclusion the indexing percentage of the 15 homepages during the Phase 1 experiment was 93 per cent, whilst that of Phase 2 was 73 per cent. Therefore the average homepages' indexing for both phases is 83 per cent.

According to the results gathered from Phase 1, five days was the minimum indexing time whilst for Phase 2, 11 days was the minimum. The maximum indexing time for both phases differed by four days as Phase 1's was 33 days and Phase 2's was 29 days. However, the average indexing time for Phase 1 was 15.1 days as compared to 22.7 days for Phase 2. The difference may be due to the four web pages that were not indexed by Google during the Phase 2 experiment.

Conversely both Phases 1 and 2 had an average indexing time of 18.9 days. The minimum indexing time mentioned by interviewees is similar to the one recorded by the experiment.

The authors consider the anomaly (scraping and cloaking done by the Iranian web site) as the reason why Google did not index GLPS1, even though it had the most favourable keyword density.

**Indexing results statistical analysis**

After the collection of the data from Phases 1 and 2 the researchers decided that the best way of statistically analysing the data was by using survival analysis. Survival analysis is based on the time an event takes to occur (indexing in this case). There are occasionally instances when an event does not take place at all for the duration of the study and these cases are labelled “censored”. Applying the concept to this study, the researchers took a case where a web page was not indexed during the period of the study, disregarding the possibility that it might be indexed after the study. The Kaplan-Meier procedure was implemented, which is a method of estimating time-to-time event in the presence of censored cases.

<table>
<thead>
<tr>
<th></th>
<th>GLPS1</th>
<th>GLPS2</th>
<th>GLPS3</th>
<th>GLPS4</th>
<th>GLPS5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>NI</td>
<td>11</td>
<td>NI</td>
<td>NI</td>
<td>NI</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>24</td>
<td>23</td>
<td>23</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Bing</td>
<td>24</td>
<td>23</td>
<td>23</td>
<td>29</td>
<td>19</td>
</tr>
</tbody>
</table>

**Table IV.** Phase 2 website homepage indexing time

**Notes:** GLPS, getlaptops; NI, Not indexed
The SPSS Advanced Statistics manual (SPSS Inc, 2007) describes the Kaplan-Meier model as being founded on estimating conditional probabilities at each time point when an event occurs and using the product limit of those probabilities to estimate the survival rate at each point in time. The Kaplan-Meier Survival Analysis assumes that the probabilities for the event depend only on time after the initial event. The researchers used this model to determine whether the time for a web page to be indexed (e.g. time-to-event) was significantly different between the three SEs. The data had to be transformed into survival data format so that for each situation the number of days it took for the event to happen (SE = Google, keyword count = 13, Phase 1) could be calculated. However, 30 records of data from Phases 1 and 2 were produced (see Table V). The survival analysis was done to compare indexing time between the three SEs.

**Analysis of indexing time comparison**

The indexing times for the three SEs were tabulated and compared (see Table VI).

The data for each of the SEs are ordered by the number of days a web page took to be indexed (time-to-event or survival time). For Google there were five records that

<table>
<thead>
<tr>
<th>Phase</th>
<th>SE</th>
<th>KWC</th>
<th>Group</th>
<th>Time to be indexed</th>
<th>Max time</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Phase 1</td>
<td>Google</td>
<td>13 keywords</td>
<td>1</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>Phase 1</td>
<td>Google</td>
<td>20 keywords</td>
<td>1</td>
<td>27</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>Phase 1</td>
<td>Google</td>
<td>28 keywords</td>
<td>1</td>
<td>28</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>Phase 1</td>
<td>Google</td>
<td>40 keywords</td>
<td>1</td>
<td>32</td>
<td>67</td>
</tr>
<tr>
<td>5</td>
<td>Phase 1</td>
<td>Google</td>
<td>90 keywords</td>
<td>2</td>
<td>32</td>
<td>67</td>
</tr>
<tr>
<td>6</td>
<td>Phase 1</td>
<td>Bing</td>
<td>13 keywords</td>
<td>1</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>7</td>
<td>Phase 1</td>
<td>Bing</td>
<td>20 keywords</td>
<td>1</td>
<td>8</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>Phase 1</td>
<td>Bing</td>
<td>28 keywords</td>
<td>1</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>9</td>
<td>Phase 1</td>
<td>Bing</td>
<td>40 keywords</td>
<td>1</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>10</td>
<td>Phase 1</td>
<td>Bing</td>
<td>90 keywords</td>
<td>2</td>
<td>4</td>
<td>67</td>
</tr>
<tr>
<td>11</td>
<td>Phase 1</td>
<td>Yahoo!</td>
<td>13 keywords</td>
<td>1</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>12</td>
<td>Phase 1</td>
<td>Yahoo!</td>
<td>20 keywords</td>
<td>1</td>
<td>8</td>
<td>67</td>
</tr>
<tr>
<td>13</td>
<td>Phase 1</td>
<td>Yahoo!</td>
<td>28 keywords</td>
<td>1</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>14</td>
<td>Phase 1</td>
<td>Yahoo!</td>
<td>40 keywords</td>
<td>1</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>15</td>
<td>Phase 1</td>
<td>Yahoo!</td>
<td>90 keywords</td>
<td>2</td>
<td>4</td>
<td>67</td>
</tr>
<tr>
<td>16</td>
<td>Phase 2</td>
<td>Google</td>
<td>100 keywords</td>
<td>2</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>17</td>
<td>Phase 2</td>
<td>Google</td>
<td>132 keywords</td>
<td>2</td>
<td>10</td>
<td>67</td>
</tr>
<tr>
<td>18</td>
<td>Phase 2</td>
<td>Google</td>
<td>170 keywords</td>
<td>2</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>19</td>
<td>Phase 2</td>
<td>Google</td>
<td>232 keywords</td>
<td>2</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>20</td>
<td>Phase 2</td>
<td>Google</td>
<td>321 keywords</td>
<td>2</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>21</td>
<td>Phase 2</td>
<td>Bing</td>
<td>100 keywords</td>
<td>2</td>
<td>23</td>
<td>67</td>
</tr>
<tr>
<td>22</td>
<td>Phase 2</td>
<td>Bing</td>
<td>132 keywords</td>
<td>2</td>
<td>22</td>
<td>67</td>
</tr>
<tr>
<td>23</td>
<td>Phase 2</td>
<td>Bing</td>
<td>170 keywords</td>
<td>2</td>
<td>22</td>
<td>67</td>
</tr>
<tr>
<td>24</td>
<td>Phase 2</td>
<td>Bing</td>
<td>232 keywords</td>
<td>2</td>
<td>19</td>
<td>67</td>
</tr>
<tr>
<td>25</td>
<td>Phase 2</td>
<td>Bing</td>
<td>321 keywords</td>
<td>2</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>26</td>
<td>Phase 2</td>
<td>Yahoo!</td>
<td>100 keywords</td>
<td>2</td>
<td>23</td>
<td>67</td>
</tr>
<tr>
<td>27</td>
<td>Phase 2</td>
<td>Yahoo!</td>
<td>132 keywords</td>
<td>2</td>
<td>22</td>
<td>67</td>
</tr>
<tr>
<td>28</td>
<td>Phase 2</td>
<td>Yahoo!</td>
<td>170 keywords</td>
<td>2</td>
<td>22</td>
<td>67</td>
</tr>
<tr>
<td>29</td>
<td>Phase 2</td>
<td>Yahoo!</td>
<td>232 keywords</td>
<td>2</td>
<td>30</td>
<td>67</td>
</tr>
<tr>
<td>30</td>
<td>Phase 2</td>
<td>Yahoo!</td>
<td>321 keywords</td>
<td>2</td>
<td>15</td>
<td>67</td>
</tr>
</tbody>
</table>

Table V. Transformed survival data
show censored values (e.g. web pages were not indexed for the duration of the study). This did not happen for the other two SEs. The fifth column (Cumulative proportion surviving at the time: estimate) shows that after ten days the cumulative survival value is 0.9. Thus the estimated probability of not being indexed beyond ten days is 90 per cent and beyond 32 days is 50 per cent.

Analysis of Google’s indexing time
Table VII lists mean and median values for the three SEs. The mean values are not the arithmetic average, but an estimated value from the survival curve. The results show that web pages take longer to be indexed by Google than by Yahoo! and Bing.

The distribution of indexing time is significantly different for the three SE populations.

Survival functions for Google, Yahoo! and Bing
Figure 4 shows the cumulative survival function over time. There is a more rapid drop-off in the cumulative survival function for Bing and Yahoo! than for Google; there are no censored values for either Bing or Yahoo!

The cumulative hazard (Figure 5) plot reflects the same trends as the survival plot. It indicates that the “risk” of being indexed increases more rapidly over time for Bing and Yahoo! than for Google.

There are many parameters employed by SEs in order to index a web page; these parameters differ from one SE to another depending on their crawler’s visits. However, this study has determined that Google took a minimum of 11 days, Yahoo! five days and Bing five days to index the web pages in the experiment.

Conclusion
The results gathered from the academic literature, the interviews and the SE guidelines were triangulated against results gathered from the experiment. The Phase 1 experiment showed Bing and Yahoo! indexing all five web sites, whilst Google indexed four. Google

<table>
<thead>
<tr>
<th>SE</th>
<th>Total n</th>
<th>No. of events (indexed)</th>
<th>Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>10</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Bing</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Overall</td>
<td>30</td>
<td>25</td>
<td>5</td>
</tr>
</tbody>
</table>

Table VI.
Indexing time comparison

<table>
<thead>
<tr>
<th>SE</th>
<th>Mean</th>
<th>95% confidence interval</th>
<th>Median</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SE</td>
<td>Lower bound</td>
<td>Upper bound</td>
<td>SE</td>
</tr>
<tr>
<td>Google</td>
<td>46.400</td>
<td>33.141</td>
<td>59.659</td>
<td>32.000</td>
</tr>
<tr>
<td>Overall</td>
<td>25.167</td>
<td>17.875</td>
<td>32.458</td>
<td>22.000</td>
</tr>
</tbody>
</table>

Note: Estimation is limited to the largest survival time if it is censored.
did not index GLPS1, which was predicted by all the interviewees to be indexed first instead of GLPS4 and GLPS5, since it had the most “natural” keyword density.

A Phase 2 experiment was conducted with a similar spread of keyword densities, except the values were higher. GLPS5 had a keyword density of more than 97 per cent, which is an extreme level. Likewise Bing and Yahoo! indexed all five web sites; however, Google surprisingly indexed GLPS2, omitting the other four web sites. There were no notifications from SEs to inform the researchers about the indexing status of the four web sites that were not indexed by Google. However, the researchers are of the opinion that GLPS1 was not indexed due to the damage caused by content scraping and cloaking from an Iranian web site, although no evidence was found to support this claim. The text displayed by Google on the SERP was exactly the same as the one for GLPS1; however, if the user visited the Iranian site it presented sales information for books and DVDs.

GLPS1 was not indexed by Google, even though it had the lowest and most favourable keyword density of 3.94 per cent. However, the Google algorithm was able
to accept a web page keyword density of 40 per cent. This lead to the assumption that Google has banned GLPS1 from its index, not as a result of keyword density, but rather from its contents having been scraped and cloaked.

Both experimental phases recorded a maximum indexing time of 33 days for the 134 days of the experimental period; i.e. the maximum indexing time according to this study was 33 days and the minimum indexing period was five days. However, many scholars differ in terms of indexing time since the time is relative to crawlers’ visits. It is best to use the long tail string search, Webmaster Tools and the site search methods in order to determine whether any of the web pages have been indexed. Using at least two of the search methods will provide a clear picture of the status of the web site indexing, but Webmaster Tools appear to be the most reliable check as it provided analytical data regarding the indexing of the web site pages.

All the interviewees acknowledged an indexing period of one day to three months if appropriate procedures are used during the design and submission of the web pages. However, according to this study, the average indexing time for the Phase 1 experiment was 15.1 days whilst Phase 2 was 22.7 days. Together Phase 1 and Phase 2 showed a period of 18.9 days as the reasonable average waiting time for indexing.

The research has further proven that it is not easy for a web developer to know whether their web page content was interpreted as SE spamdexing. This is due to the fact that the web sites that were expected to be blacklisted in this research and considered spam, were also indexed. There was no form of notification from the SEs indicating the status of the highly populated keyword-dense web pages.

**Contributions and significance of the study**
The study has made the following contributions to the body of knowledge:

- a recent examination of possible minimum and maximum indexing times for Google, Yahoo! and Bing;
- a definition of a minimum indexing failure rate as web pages are censored before submission to avoid indexing delays or failures; and
- proof that web site designers can take advantage of higher keyword density strategies for quick indexing by Yahoo! and Bing.

The data collected from the interviews and literature as well as the experimental study on keyword density, have shown a significant difference between the perceptions of SEO practitioners and scholars. The triangulation method enabled the researchers to clearly identify areas of divergence and uncertainties displayed by SEO practitioners and web site designers in understanding keyword density and keyword stuffing.

It is believed that if scholars, SEO practitioners and web site designers incorporate the results of this study the following may be expected:

- higher user satisfaction due to adherence to end-user specifications;
- minimum bounce rate as web pages will be more content-centred than SE-centred;
- scholars adopting the new keyword evolution displayed by SEs; and
- increased conversion rate due to strategies and methodology changes.

Some implications of this research now become evident. The waiting time for indexing appears to be slightly shorter than the expected number of days, meaning the use of
paid indexing services loses some value. It could make use of these services unnecessary, except in cases where a web site is of such a nature that immediate exposure could mean the difference between success and failure. In these cases paid placement service such as Google’s Adwords should be used.

Second the keyword density results are surprising. It appears as if higher keyword densities are not as frowned upon by the SEs as the earlier literature seems to suggest. The implication is that web site designers can now load more keywords onto web pages, thereby increasing the keyword count and association crawlers make with searching concepts. However, unnaturally high keyword densities should be avoided, since it would probably deter human visitors. This is directly in opposition to the reason web site owners desire high rankings: to draw visitors to web pages, and keep them there.

Finally the last three research questions have been addressed through this study. SE practitioners disagree on the definition of keyword stuffing and the definition of keyword richness. Web developers have no easy way of knowing how a SE has interpreted their web site, short of waiting and noting that it might disappear from the index.

Recommendations
Web site owners should make use of programs such as Webmaster Tools and other open source and commercial software for monitoring web page indexing. These tools also offer detailed information regarding the status of the submitted web pages e.g. if the page contains some errors. After around 33 days one should analyse the submitted web page to see whether there are errors preventing indexing, otherwise the web page should be resubmitted.

Keyword stuffing should not be used as a deceptive technique for web site ranking; instead keyword density should be leveraged in ways that increase a site's visibility, usability, conversion and accessibility. SEO tools should be utilised regularly as they provide a basic structural architecture of how the web site is built and how it can be accessed on the web. They further offer a range of recommendations that assist in amendments and other strategic decisions that have to be made. A company’s methodology or strategy must be reviewed constantly in accordance with the SE algorithm changes, as each algorithm change could possibly have a positive but more likely negative impact on the company’s strategy. A relevant example is the Google Panda “Farmer” algorithm update that was released at the time of writing.

It would be more effective to spend resources on making web page content interesting, relevant and engaging, rather than compromising web page relevance by keyword stuffing. The cost of retaining a client is lower than winning a new one and clients’ intent should be well understood.

Research limitations and further research
This study was limited to three SEs only: Google, Yahoo! and Bing. Furthermore the study considered keyword density only and all other SEO techniques are assumed to be held constant. The study was limited to 67 days for each phase; future studies could be done for a longer time, e.g. 90 days.

Future studies could also be undertaken in areas that include other black hat techniques such as content scraping and cloaking, as well as the relevancy of information displayed on the SERP.
References


**About the authors**

Herbert Zuze holds an honours degree in Computer Science from the Bindura University of Zimbabwe. He recently graduated with a Master’s Degree of Technology (BIS) from CPUT, and has experience as IT administrator and PC support. He is currently pursuing his doctoral degree.

Melius Weideman is a Professor heading the Research Development Department (Faculty of Informatics and Design) at the Cape Peninsula University of Technology in Cape Town. He graduated with a PhD in Information Science from UCT in 2001, and focuses his research on website visibility and usability. Melius Weideman is the corresponding author and can be contacted at: weidemanm@cput.ac.za

To purchase reprints of this article please e-mail: reprints@emeraldinsight.com
Or visit our web site for further details: www.emeraldinsight.com/reprints