PROCEEDINGS OF THE 13th ANNUAL CONFERENCE ON WORLD WIDE WEB APPLICATIONS

14-16 September 2011
Johannesburg
South Africa

Editors:

A. Koch
P.A. van Brakel

Publisher:

Cape Peninsula University of Technology
PO Box 652
Cape Town
8000

Proceedings published at
http://www.zaw3.co.za

ISBN: 978-0-620-51918-2

TO WHOM IT MAY CONCERN

The full papers were refereed by a double-blind reviewing process according to South Africa's Department of Higher Education and Training (DHET) refereeing standards. Before accepting a paper, authors were to include the corrections as stated by the peer-reviewers. Of the 59 full papers received, 41 were accepted for the Proceedings (acceptance rate: 69.5%).

Papers were reviewed according to the following criteria:

- Relevancy of the paper to Web-based applications
- Explanation of the research problem & investigative questions
- Quality of the literature analysis
- Appropriateness of the research method(s)
- Adequacy of the evidence (findings) presented in the paper
- Technical (e.g. language editing; reference style).

The following reviewers took part in the process of evaluating the full papers of the 13th Annual Conference on World Wide Web Applications:

Prof S Berman
Department of Computer Science
University of Cape Town
Cape Town

Prof RA Botha
Department of Business Informatics
Nelson Mandela Metropolitan University
Port Elizabeth

Mr AA Buitendag
Department of Business Informatics
Tshwane University of Technology
Pretoria

Prof AJ Bytheway
Faculty of Business
Cape Peninsula University of Technology
Cape Town

Dr A Chigona
E-Learning Support and Innovation Unit
University of the Witwatersrand
Johannesburg

Prof T du Plessis
Department of Information and Knowledge Management
University of Johannesburg
Johannesburg

Dr L Harrison
Educational Technology, CELT
Durban University of Technology
Durban

Prof M Herselman
Meraka Institute, CSIR
Pretoria

Mr B Kalema
Department of Business Informatics
Tshwane University of Technology
Pretoria

Ms F Mohsam
Faculty of Business
Cape Peninsula University of Technology
Cape Town

Dr J Mostert
Centre for Development Support
University of the Free State
Bloemfontein

Prof L Nagel
Department of Education Innovation
University of South Africa
Pretoria

Ms C Muir
Department of Strategic Communication
University of Johannesburg
Johannesburg

Mr R Proske
University Library
Cape Peninsula University of Technology
Cape Town

Mr F Schwenke
Faculty of Informatics and Design
Cape Peninsula University of Technology
Cape Town

Prof A Singh
Business School
University of KwaZulu-Natal
Durban

Prof JS van der Walt
Department of Business Informatics
Tshwane University of Technology
Pretoria

Prof D van Greunen
School of ICT
Nelson Mandela Metropolitan University
Port Elizabeth

Dr SC Warden
Faculty of Informatics and Design
Cape Peninsula University of Technology
Cape Town

Further enquiries:

Prof PA van Brakel
Conference Chair: Annual Conference on WWW Applications
Cape Town
+27 21 469 1015 (landline)
+27 82 966 0789 (mobile)

A comparative analysis of search engine indexing time

H. Zuze
Faculty of Business
Cape Peninsula University of Technology
Cape Town, South Africa
herbertzuze@gmail.com

M. Weideman
Faculty of Informatics & Design
Cape Peninsula University of Technology
Cape Town, South Africa
weidemanm@cput.ac.za

Abstract

Internet usage is increasing dramatically daily, as is Web development which is enhanced by the emergence of new Web technologies. Social networks are continuously dominating our Web interactions through Facebook, Twitter, LinkedIn and MySpace, drawing a considerable number of Internet users. However, the dominance of the Web as the epicentre of both important and useless information has affected competition in industries like e-business. It takes some time for a website to be indexed and appear on the search engine result page. The indexing process involves the reading and recording of the weight-carrying words in a search format to an index file by a crawler. This results in the webpage being discovered by users on a search engine results page. A user can also submit a website manually for indexing. Crawlers should then visit this website, record all the words on the pages, and note links to other sites. The index file is updated regularly, either by human editors or by these crawlers. After an extensive literature survey, empirical evidence indicated that the indexing period for webpages is not fixed. In a recent two phase experiment, five websites were monitored to determine their indexing time for Google, Yahoo! and Bing. The experiment proved that there is a relationship between keyword density and indexing time. Yahoo! and Bing seem to favour sites with high keyword density when indexing. During the first experiment the shortest indexing waiting time was five days and the longest, 33 days. For the second one, the waiting time varied between 19 and 29 days. However, according to this study, a period of approximately 19 days is a reasonable average waiting time. It is possible that the Google sandbox effect plays a role in these experiments, and this was also investigated.

Keywords: Internet, webometrics, indexing, search engines, optimisation, website visibility, website usability

1. Introduction

Visser and Weideman (2011) noted that the "value" of a website can be determined by the page on which it ranks. For this reason Search engine optimization (SEO) becomes a prerequisite.

However, the first step to success is to have a website included in the search engine indices. Research has proven that at least 67% of users will only read the first page of results, while only 9% will read further than the third page (Weideman 2009:32). 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 2008:10).

Due to the increase in use of the Internet as a marketing tool, there is a constant increase in the submission of webpages for indexing by search engines. As a result, website owners have taken an interest in the time it takes before a search engine includes their website in the index. This length of time could have an impact in, for example, a company's profitability, since first-time inclusion in a search engine index is equivalent to being discovered by searchers. It is estimated that nearly 80% of users utilise search engines to locate information on the Internet. This, by implication, places emphasis on the underlying importance of webpages being listed on search engines' indices (Kritzinger & Weideman 2007).

1.1 Research objectives

This research study has the following objectives:

- To determine indexing time of a webpage by Google, Yahoo! and Bing.
- To determine how SEO practitioners view webpage indexing time.
- To determine how search engines index webpages with varying keyword density.

1.2 Research questions

This research is based on the following research questions:

- How long do Google, Yahoo! and Bing take to index a page?
- What is the view of academic experts on the time a search engine takes to index a webpage?
- Does keyword density affect webpage indexing time?

2 Literature review

Different views are presented by scholars in respect of indexing time with the variation based on their own experiments.

2.1. Search engine scholars, practitioners

The researcher, as well as the interviewees during this research and several other scholars, including Weideman (2009) and Zhang & Dimitroff (2005), found empirical evidence indicating that the indexing period for webpages is not fixed. Borglum (2009:30) stated that after the submission of a website, one has to wait for a period of up to one month to evaluate indexing results.

The author further claimed that a search engine often only updates websites on a monthly basis. When discussing Google's view, Mathews (2011), mentioned that it should take up to two weeks for a site to be indexed.

2.2. Search engines

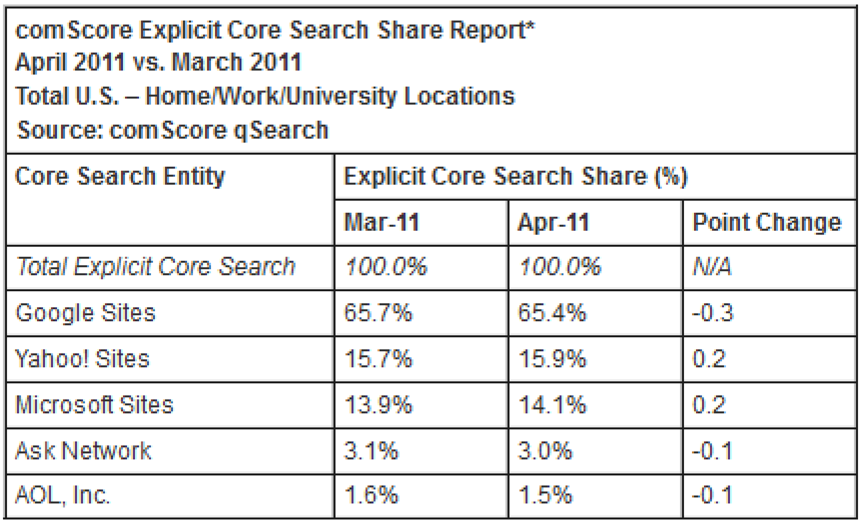
According to Weideman (2004:2-3):

> "A search engine is a program that offers user interaction with the Internet through a front end, where the user can insert a search term or make successive selections from relevant directories. Hereafter, the search engine compares the search term against an index file, which contains information concerning webpages. Matches found are then returned to the user via the front end".

Google, Bing and Yahoo!, as the current search engine giants with the greatest market share (Snack 2011), have a large impact on the way in which Search Engine (SE) practitioners design their websites. They should further adhere to the principles that govern the inclusion of their website in the index of these respective search engines.

Recently, data released by Flosi (2011) shows that Google Sites led the U.S. explicit core search market in April with a 65.4% market share, followed by Yahoo! Sites with 15.9% and Bing with 14.1% (see Table 1)

Table 1: Percentage of U.S. explicit core search (Flosi 2011)



comScore Explicit Core Search Share Report*
April 2011 vs. March 2011
Total U.S. – Home/Work/University Locations
Source: comScore qSearch

Core Search Entity	Explicit Core Search Share (%)		
	Mar-11	Apr-11	Point Change
Total Explicit Core Search	*100.0%*	*100.0%*	*N/A*
Google Sites	65.7%	65.4%	-0.3
Yahoo! Sites	15.7%	15.9%	0.2
Microsoft Sites	13.9%	14.1%	0.2
Ask Network	3.1%	3.0%	-0.1
AOL, Inc.	1.6%	1.5%	-0.1

After submitting a webpage or a website to a search engine, the search engine spider would then index part of or the whole website. Web spiders specialise in downloading Web content, then analysing and indexing this content. These spiders primarily do text indexing and link following, and this means that if a spider fails to find content or links on a site it leaves the site without recording anything (Parhizkar 2010).

2.3. Website visibility

The most basic purpose of a website is to provide relevant, valuable content and enable users to locate it on a Search Engine Result Page (SERP). When a user performs a search for information via a search engine, the interface directs the query to the index where matches are made to the content of the index. The results from the index are presented to the user on a SERP (Weideman 2009:30).

According to Weideman (2009:14), visibility is defined as the ease with which a search engine crawler can find a webpage. After finding the information, it is then defined by the degree of the success the crawler has in indexing the page. Website owners invest substantial resources in order to influence their online visibility (Ron and Zsolt 2011). If the quality of sites corresponds with their estimation for visitors, then SEO aids as a mechanism that improves the ranking by correcting measurement errors.

2.4. Website usability

Usability deals with the feeling of being able to easily and successfully interact with a website. It further measurers the extent with which a visitor can easily and quickly use webpage resources. Usability also includes the following factors:

- easiness of learning,
- subject satisfaction,
- easiness 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, Quarto-vonTivadar, Davis & Crosby 2008:158).

Visser and Weideman (2011) are of the opinion that the inclusion of usability attributes will enhance conversion; therefore, effective website design should incorporate usability as a prerequisite. They further state that there is a need for weighing in terms of importance of usability and SEO towards search engines and visitors, since these two practices occasionally contradict each other.

2.5. Ranking

The more content a website has the more weight-carrying key phrases the website could possibly rank for. Search engines reward both qualitative and quantitative websites with solid, informative and useful content with good rankings for specific search terms or phrases (Visser & Weideman 2011).

Theoretically, the better a particular website ranks, the more traffic that website should receive and the more visitors ought to convert. This is supported by Weideman (2009:32) through the fact that on average 67% of search engine users do not look beyond the first SERP. For commercially-oriented websites whose income depends on their traffic, it is in their interest to be ranked within the top 10 of the SERP for a query relevant to the content of the website. The purpose is to boost those websites' rankings on search engines such as Google when users search for keywords. Higher rankings mean more click-through and often, higher income (Zahorsky 2010:32-33).

2.6. Keywords

Malaga (2009:132-139) claims that keyword density measures the extent to which a certain word or phrase appears on a site or a webpage. Following the user search behaviour notion, the best measurement in terms of the number of words should be established – it is advisable to provide search engines with a rich harvest of keywords, but not too many as this might scare off human readers (Visser & Weideman 2011).

2.7. Indexing

Around 80% of users utilise search engines to locate information on the Internet (Zhang & Dimitroff 2005:665). This fact emphasises the underlying importance of webpage owners being listed with search engines.

Indexing is the process of reading and recording the weight-carrying words in a search format to an index file, with the goal of getting indexed quickly (and hopefully ranked well) by the search engine (Weideman 2009:192). The search engines “discover” a new site when the spiders find a link to that site from other sites (Malaga 2009:132-139). A user can also submit a website manually for indexing.

In its process of determining the most relevant webpages, a SE selects a set of candidates’ pages that comprises of some or all of the query terms and calculates a score for every webpage. Lastly, a list of webpages are sorted by their respective scores and returned to the end-user (Egele, Kolbitsch & Platzer 2009:51-62).

Spiders constantly crawl the Web, returning with new and updated pages to be indexed and stored. According to Benczur, Erdelyi, Masanes and Siklosia (2009), when sharing knowledge across different domains, the linkage and the crawl strategies in use differ. For example, Erdelyi, Garzo and Benczur (2011) found that recent results and crawlers’ visitations have concentrated on the definition of new features, hence ignoring other important factors in the domain of machine learning techniques that affect SEO results.

2.8 Conclusion

Indexing time is not fixed but varies with search engine crawler visitation. Also, researchers differ on the ability of a webpage to meet the correct design and submission procedures. This has become evident through the work of Borglum, Weideman and Zhang & Dimitroff. These scholars found further similarities in the amount of time taken by search engines to index a webpage - they identified periods of one to 90 days as the minimum and maximum waiting time. These opinions motivated this study to be done in two phases with the maximum waiting time exceeding 90 days in order to precisely test the validity of the literature.

3 Methodology

An analysis of the webpage indexing time was done based on triangulation. The data was gathered through a combination of a literature review, personal interviews and an empirical experiment.

Five new websites were created, all with similar but different content. New domains were registered where these websites were hosted, to remove the possibility of some domains already having been indexed prior to the experiment. The three main search engines were chosen, and the five domains submitted for the first time within an hour of each other on the same day. This was the start of the monitoring period.

All three search engines were then checked daily from this point onwards, to note if any one or more of the domains have been indexed. This was done using specific operator searching, and by searching for specific content known to be present on the websites.

The dates of all these checks were recorded, from which the data listed elsewhere was drawn.

3.1. Interviews

Five participants representing five well-known companies in Cape Town were involved in the interviews. A structured questionnaire was used during the interview and the researcher summarised the responses. The researcher drew this sample to represent the entire population of the body of SEO because of:

- their ability to conduct business in a trusted manner and
- their experience with the application of SEO principles.

3.2. Experimental research

According to Fox and Bayat (2007:10), an experimental research approach aims at forecasting what may occur or, otherwise, intends to bring together changes or new approaches within the prevailing situation in order to determine the outcome. The researcher divided the experiments into two phases - Phase 1 and Phase 2. During the Phase 1 experiment, five experimental websites were designed in simple HTML (Hyper Text Mark-up Language) with no Flash files, frames, JavaScript or excessive graphics so that the crawlers would easily visit them. Consistent, relevant, different but similar content was placed on every website to meet all the appropriate white hat regulations with the exception of keyword density. The five websites are commercial sites that provide information about various second hand laptops sales as well as laptop accessories. The research is centred on the homepage whilst the other pages were intended to provide increased site content that would help during crawler visitation. The website consisted of three pages, namely:

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

These websites can be found at, respectively:

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

The catalogue page and the contact page had similar but different content and were designed to increase the website informational value that would assist in terms of importance of the site to the user. This would also improve the webpage's chances of being indexed and favoured by crawlers. The home page for each website was the page with varying keyword densities and distribution, whereas the content was similar and the word count identical.

After collecting the experimental results, certain anomalies regarding the theoretical claims on webpage indexing were noted. As a result, the researcher 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).

The following were changed from Phase 1 experiment to Phase 2 experiment to investigate the response of the search engine crawlers:

- the keyword density of all the websites, with the fifth website carrying close to the maximum keyword density with “laptops” as the keyword and
- a fourth page per website was added to attract crawlers' attention to the sites.

Currently the content of the five domains as listed above reflect Phase 2.

3.3. Website submission

The websites were submitted to Google, Bing and Yahoo! Within one hour from each other on the same day, and the researcher investigated how these search engines indexed each site. A daily check was then done to establish when, if at all, these sites were indexed. The submission was done using the standard search engine submission consoles.

4 Results and analysis

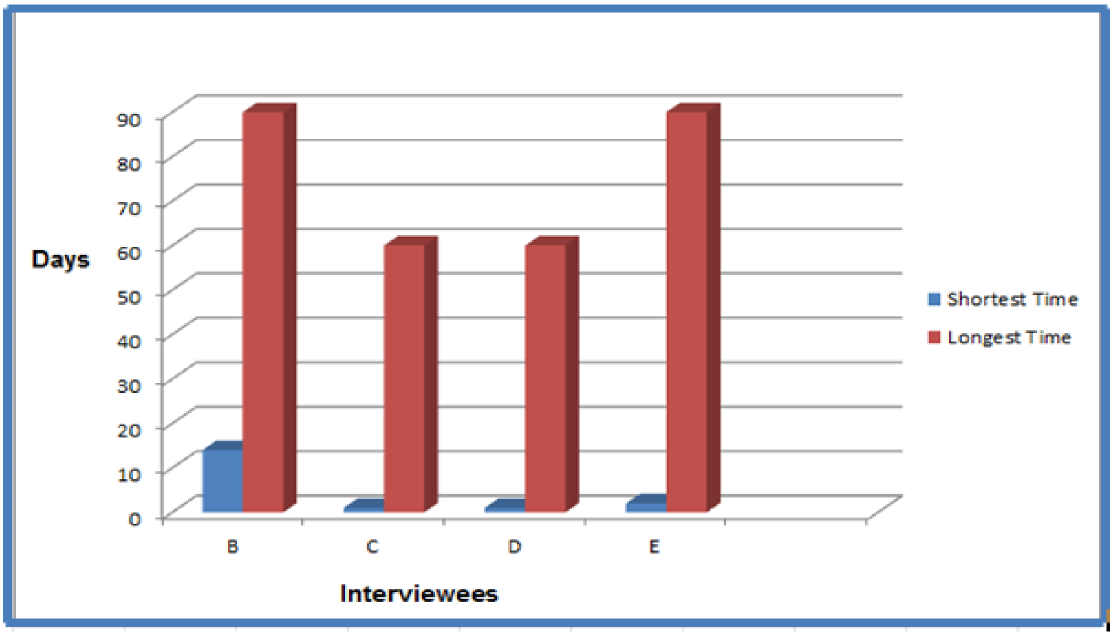
In this section, data will be analysed according to the categories specified below:

- interviews with the SEO practitioners,
- literature and
- data collected from the five experimental websites.

4.1 Interviews indexing analysis

There were different views in respect of the time it could take for a website to be indexed. This was characterised by time ranging from a day to three months if the correct procedure of submission was followed. The research shows that 80% of the interviewees indicated that they have waited for up to a period of 90 days to have a website indexed; various reasons, which were beyond their control, were stated. The main reason was based on crawler visitation that is not known by anyone. Figure 1 indicates the opinions of the five experts on the time it takes to get a website indexed.

Figure 1: Minimum and maximum indexing time as stated by each interviewee



According to the interviewees, the shortest indexing time experienced was a single day whilst the longest indexing time was 90 days. However, the average shortest indexing time was 6 days and the average longest time was 72 days.

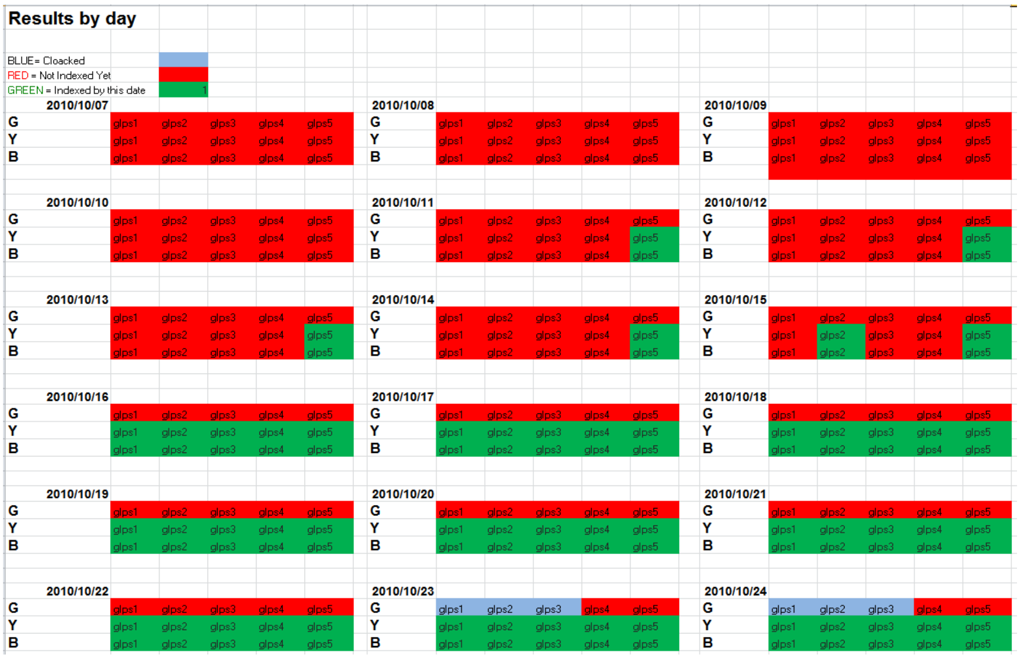
4.1.1 Experimental results analysis

After the submission of five websites to Google, Yahoo! and Bing, the researcher recorded each day's indexing results (see Figure 2 and Figure 3 for indexing results recording). They were checked using the following methods:

- a string search,
- a site search and
- the Webmaster tools (search engine analysis for each registered website).

Figure 2 shows the results recorded over the experimental period. Three colour coding was used to represent indexing status for each day. Red colour represented "not indexed" whilst green represented "indexed".

Figure 2: Phase 1 indexing results recording



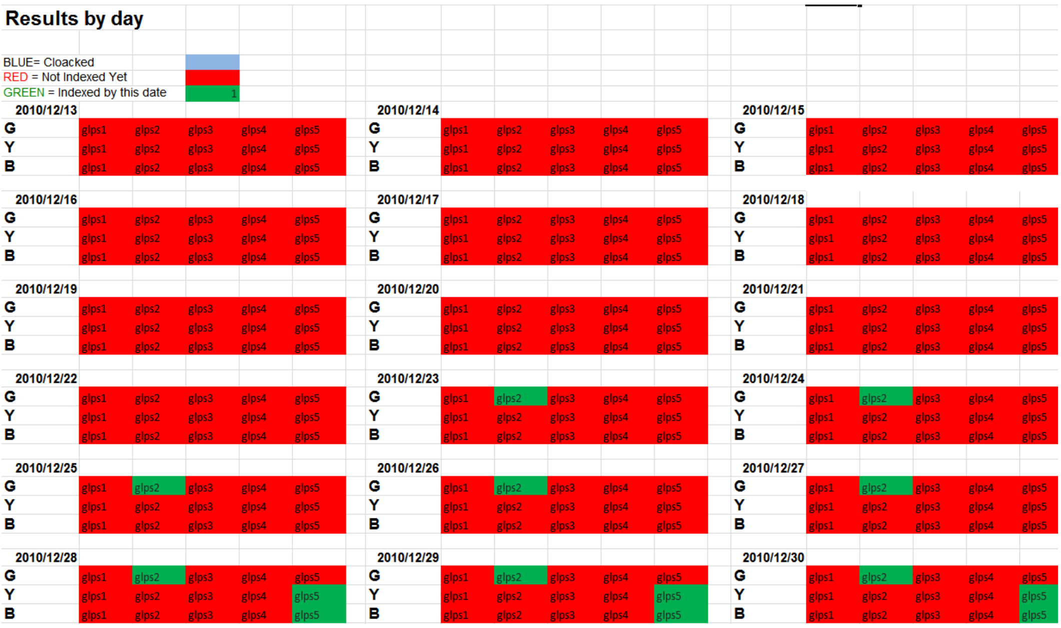
Results by day

BLUE= Cloacked 🟦
RED = Not Indexed Yet 🟥
GREEN = Indexed by this date 🟩 1

2010/10/07						2010/10/08						2010/10/09					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/10/10						2010/10/11						2010/10/12					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5	Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5	B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5
2010/10/13						2010/10/14						2010/10/15					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5	Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5	Y	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟩 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5	B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5	B	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟩 glps5
2010/10/16						2010/10/17						2010/10/18					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5
B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5
2010/10/19						2010/10/20						2010/10/21					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5
B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5
2010/10/22						2010/10/23						2010/10/24					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5	G	🟦 glps1	🟦 glps2	🟦 glps3	🟥 glps4	🟥 glps5	G	🟦 glps1	🟦 glps2	🟦 glps3	🟥 glps4	🟥 glps5
Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	Y	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5
B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5	B	🟩 glps1	🟩 glps2	🟩 glps3	🟩 glps4	🟩 glps5

The recordings of indexing results for Phase 2 are in Figure 3. The same colour coding in Phase 1 experiment were also used in recording indexing results for Phase 2 experiment.

Figure 3: Phase 2 indexing results recordings



Results by day

BLUE= Cloacked 🟦
RED = Not Indexed Yet 🟥
GREEN = Indexed by this date 🟩 1

2010/12/13					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/14					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/15					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/16					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/17					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/18					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/19					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/20					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/21					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/22					
G	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/23					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/24					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/25					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/26					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/27					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟥 glps5
2010/12/28					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5
2010/12/29					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5
2010/12/30					
G	🟥 glps1	🟩 glps2	🟥 glps3	🟥 glps4	🟥 glps5
Y	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5
B	🟥 glps1	🟥 glps2	🟥 glps3	🟥 glps4	🟩 glps5

4.1.2 Phase 1 indexing results analysis

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 website pages were successfully indexed with the exception of Getlaptops1, which was not indexed by Google. The research showed the fifth website as being the most favoured one and was indexed first by Yahoo! and Bing four days after submission. After five more days Yahoo! and Bing registered the remainder of the websites, including even Getlaptops5, having the highest keyword density of 27.30%.

Table 2 shows the homepage indexing time in days, recorded over a period of 67 days.

Table 2: Phase 1 website homepage indexing time

	GLPS1	GLPS 2	GLPS 3	GLPS 4	GLPS 5
GOOGLE	NI	28	29	33	33
YAHOO!	10	9	10	10	5
BING	10	9	10	10	5

Key

GLPS - Getlaptops
NI - Not indexed

According to Castle (2011), the website age has a considerable impact on its rankings, where the age includes the age of the websites that link to the website. The writer further acknowledged that Google checks each time a website links to a site and the rate at which the site's inbound links grow over time. This means that new domains can find it difficult to rank as it takes them long to get trusted by Google (Nade 2010). This is termed the "Google sandbox effect".

Google took longer to index all five test websites than the other two SE's; the researchers consider that this was due to the Google "Sandbox Effect". The "sandbox" denotes the fact that Google applies an aging filter to its index; simply put, it prefers older to newer sites.

4.1.3 Phase 2 indexing results analysis

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 highest keyword density of 97.3%. Google only indexed one of the five sites, with a keyword density of 40%. After 67 days, the experiment was officially closed and results were recorded.

Table 3 shows the homepage indexing time in days recorded over a period of 67 days.

Table 3: Phase 2 website homepage indexing time

	GLPS1	GLPS 2	GLPS 3	GLPS 4	GLPS 5
GOOGLE	NI	11	NI	NI	NI
YAHOO!	24	23	23	20	19
BING	24	23	23	29	19

Key

GLPS - Getlaptops
NI - Not indexed

4.2 Phase 1 and Phase 2 indexing analysis

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

According to the results gathered from Phase 1 experiment, five days was the minimum indexing time whilst Phase 2 showed 11 days as minimum. The maximum indexing time for both Phase 1 and Phase 2 differed by four days as Phase 1 had 33 days and Phase 2 had 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 webpages that were not indexed by Google during Phase 2 experiment.

Conversely, both Phase 1 and Phase 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 experimental study.

4.3 Indexing results statistical analysis

After the collection of the data from Phase 1 and Phase 2 the researcher found 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. 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 researcher took a case where a webpage did not get indexed during the period of the study, disregarding the event that it might be indexed after the study. The researcher implemented the Kaplan-Meier procedure, which is a method of estimating time-to-time-event in the presence of censored cases.

The SPSS Manual (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 researcher used this model to determine if the time for a webpage to be indexed (e.g. Time to event) was significantly different between the three search engines. The data had to be transformed into survival format data 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 Phase 1 and Phase 2 were produced (see Table 4). The survival analysis was done on comparing indexing time between the three search engines.

Table 4: Transformed survival data

	Phase	SE	KWC	Group	Time to be indexed	Max Time	Status
1	Phase 1	Google	13 keywords	1	67	67	Censored
2	Phase 1	Google	20 keywords	1	27	67	Event
3	Phase 1	Google	28 keywords	1	28	67	Event
4	Phase 1	Google	40 keywords	1	32	67	Event
5	Phase 1	Google	90 keywords	2	32	67	Event

6	Phase 1	Bing	13 keywords	1	9	67	Event
7	Phase 1	Bing	20 keywords	1	8	67	Event
8	Phase 1	Bing	28 keywords	1	9	67	Event
9	Phase 1	Bing	40 keywords	1	9	67	Event
10	Phase 1	Bing	90 keywords	2	4	67	Event
11	Phase 1	Yahoo!	13 keywords	1	9	67	Event
12	Phase 1	Yahoo!	20 keywords	1	8	67	Event
13	Phase 1	Yahoo!	28 keywords	1	9	67	Event
14	Phase 1	Yahoo!	40 keywords	1	9	67	Event
15	Phase 1	Yahoo!	90 keywords	2	4	67	Event
16	Phase 2	Google	100 keywords	2	67	67	Censored
17	Phase 2	Google	132 keywords	2	10	67	Event
18	Phase 2	Google	170 keywords	2	67	67	Censored
19	Phase 2	Google	232 keywords	2	67	67	Censored
20	Phase 2	Google	321 keywords	2	67	67	Censored
21	Phase 2	Bing	100 keywords	2	23	67	Event
22	Phase 2	Bing	132 keywords	2	22	67	Event
23	Phase 2	Bing	170 keywords	2	22	67	Event
24	Phase 2	Bing	232 keywords	2	19	67	Event
25	Phase 2	Bing	321 keywords	2	15	67	Event
26	Phase 2	Yahoo!	100 keywords	2	23	67	Event
27	Phase 2	Yahoo!	132 keywords	2	22	67	Event
28	Phase 2	Yahoo!	170 keywords	2	22	67	Event
29	Phase 2	Yahoo!	232 keywords	2	30	67	Event

30	Phase 2	Yahoo!	321 keywords	2	15	67	Event

4.3.1 Analysis 1: comparing indexing time between Google, Yahoo! and Bing

The indexing time for the three search engines was tabulated and compared, the indexing data recorded was summarised as shown in the case processing summary table (see Table 5).

Table 5: Case processing summary

			Censored	
SE	Total N	N of Events (Indexed)	N	%
Google	10	5	5	50.0%
Bing	10	10	0	.0%
Yahoo!	10	10	0	.0%
Overall	30	25	5	16.7%

The data for each of the search engines is ordered by the number of days a webpage took to be indexed (time-to-event or survival time). For the search engine Google, there were five records that show censored values (e.g. webpages were not indexed for the duration of the study). This did not happen for the other two search engines. The fifth column ("Cumulative Proportion Surviving at the Time: Estimate") shows that after 10 days the cumulative survival value is 0.9. Thus, the estimated probability of not being indexed beyond 10 days is 90.0% and beyond 32 days is 50%.

4.3.2 Analysis of Google indexing time

The mean values in Table 6 are not the arithmetic average, but an estimate value from the survival curve. The results showed webpages taking longer to be indexed with Google than with Yahoo! and Bing.

Table 6: Means and medians for survival time on webpage indexing

	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
SE	Mean	Std. Error	Lower Bound	Upper Bound	Median	Std. Error	Lower Bound	Upper Bound

Google	46.400	6.765	33.141	59.659	32.000	.	.	.
Bing	14.000	2.226	9.637	18.363	9.000	2.767	3.577	14.423
Yahoo!	15.100	2.718	9.773	20.427	9.000	2.767	3.577	14.423
Overall	25.167	3.720	17.875	32.458	22.000	3.757	14.637	29.363

a. Estimation is limited to the largest survival time if it is censored.

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

4.3.3 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 as the survival plot. It indicates that the "risk" of being indexed increases more rapidly over time for Bing and Yahoo! than for Google.

Figure 4: Webpage indexing survival function

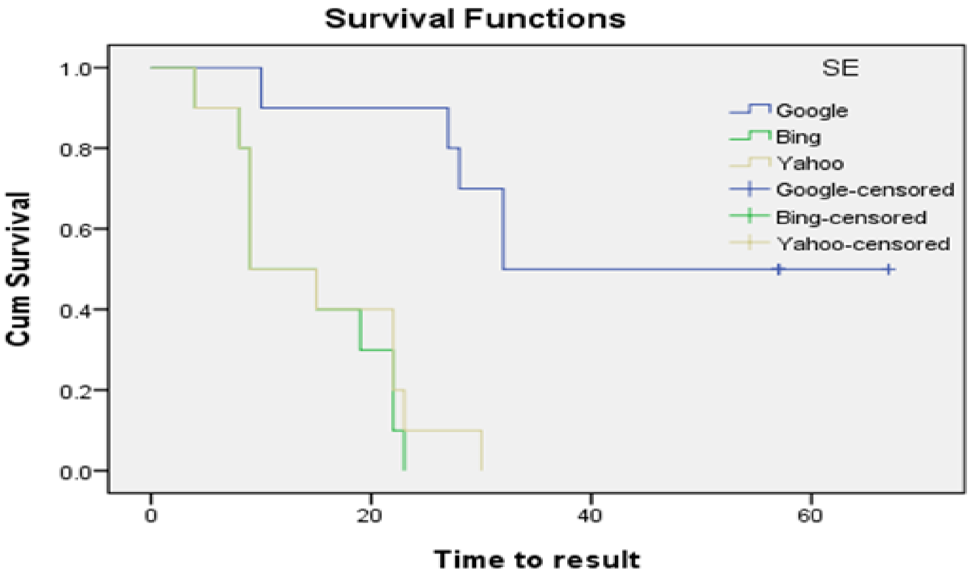
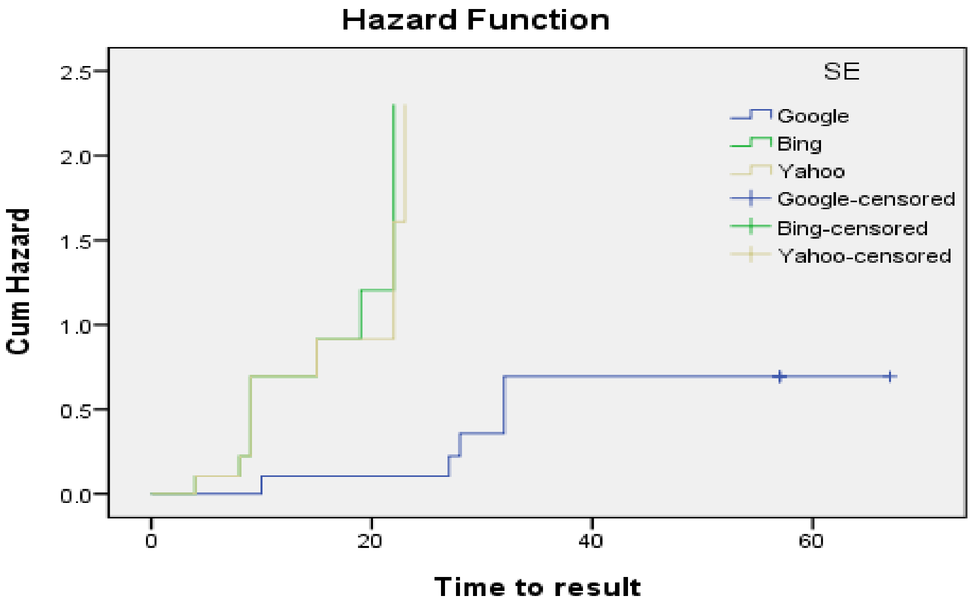


Figure 5: Webpage indexing hazard function



4.4 Contribution of the study

In this study, a combination of the opinions/results of the following has been done: the literature, SE practitioners, website designers and empirical experiments. Consequently the study has made the following contributions to the body of knowledge:

- A recent understanding of a possible minimum and maximum indexing time as being relevant to Google, Yahoo! and Bing.
- A definition of a minimum indexing failure rate as webpages are censored before submission to avoid indexing delays or failures.
- Enabling website designers to take advantage of a higher keyword density strategy for quick indexing by Yahoo! and Bing.

4.5 Conclusion

The results gathered from academic literature, the interviews and the search engine guidelines were triangulated against results gathered from the experiment conducted. Both Phase 1 and Phase 2's experiment results were monitored for 67 days. The Phase 1 experiment showed Bing and Yahoo! indexing all five websites, whilst Google indexed four. Google did not register Getlaptops1, which was predicted by all the interviewees to be indexed first instead of Getlaptops4 and Getlaptops5.

A Phase 2 experiment was conducted with the fifth website having a keyword density of more than 97% which is an extreme excess. Likewise, Bing and Yahoo! indexed all five websites; however, Google surprisingly indexed Getlaptops2, omitting the other four websites. There were no notifications from search engines to inform the researcher about the indexing status of the four websites that were not indexed by Google. However, the researcher is of the opinion that Getlaptops1 was not indexed due to the damage caused by content scraping and cloaking from an Iranian website, but no evidence was found to support this claim. The text displayed by Google was exactly the same as the one for Getlaptops1; however, if the user visited the Iranian site it presented sales information for books and DVDs.

Getlaptops1 was not indexed by Google even though it had the lowest and most favourable keyword density of 3.94%, supported by the interviewees and scholars. However, the Google algorithm was able to accept a webpage keyword density of 40%. Figure 6 and Figure 7 depicts the Phase 1 and Phase 2 indexing time recorded during the experiment.

Figure 6: Phase 1 website indexing time recorded over a period of 35 days

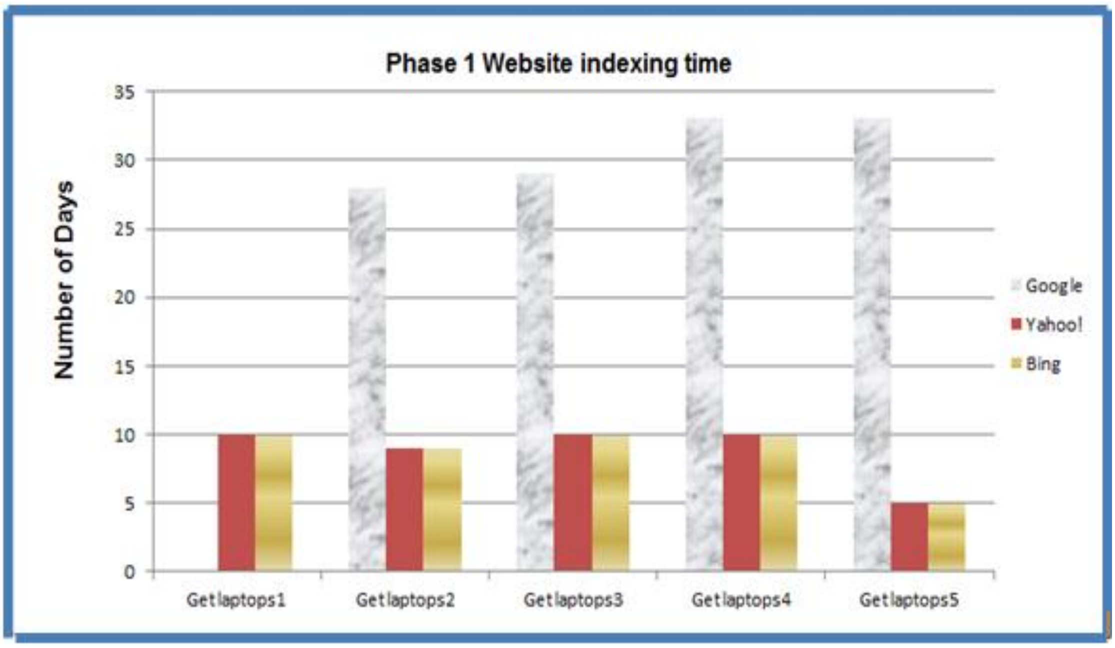
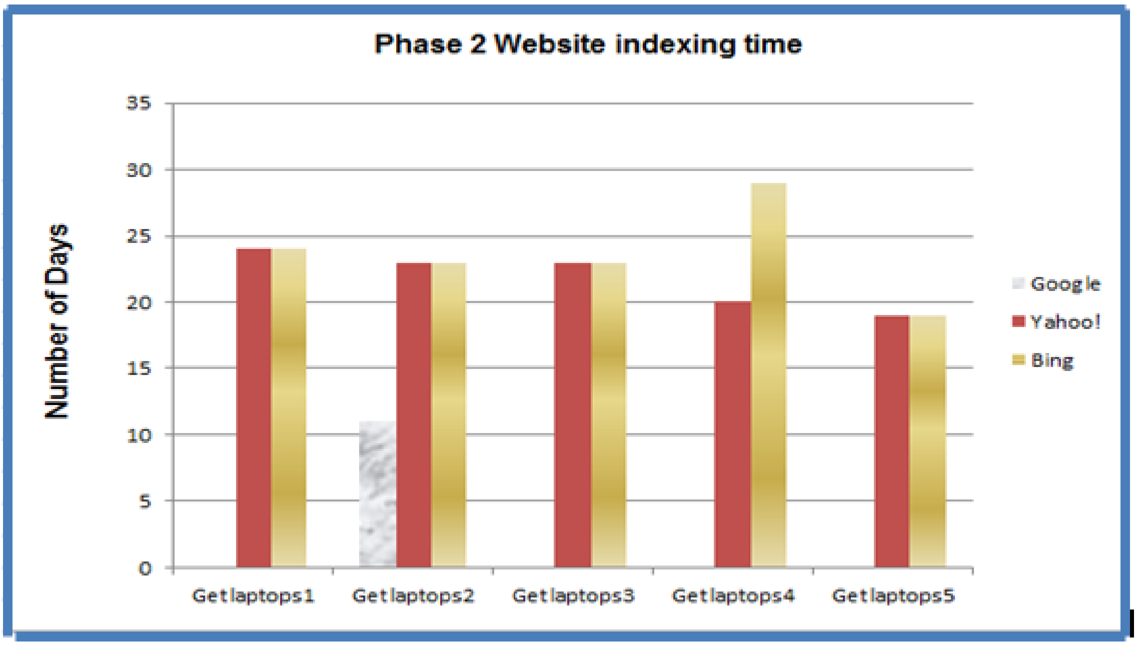


Figure 7: Phase 2 website indexing time recorded over a period of 35 days



Both experimental phases recorded a maximum indexing time of 33 days for the 67 days of the experimental period; that is to say 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' visitation. Thus a site can even be indexed in one day. It is ideal to use the long tail string search, Webmaster Tools and the site search methods in order to determine if the homepage or any of the subpages have been indexed. Using at least two of the search methods will provide a clear picture of the status of the website indexing, but above all Webmaster Tools form the standard check as it provided analytical data regarding the indexing of the website pages.

All the interviewees acknowledged a period of one day to three months if appropriate procedures are engaged during the design and submission of the webpages. 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. Conclusively Phase 1 and Phase 2 showed a period of 18.9 days as the reasonable average waiting time.

4.6 Research implications

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 website is of such a nature that immediate exposure could mean the difference between success and failure. Secondly, the keyword density results are surprising. It appears as if higher keyword densities are not frowned upon by the search engines as earlier literature seems to suggest. The implication is that website designers can now load more keywords onto webpages, thereby increasing the keyword count and association crawlers make with searching concepts.

4.7 Recommendations

Website owners should make use of tools such as Webmaster Tools and other open source and commercial software for checking if a webpage is indexed. These tools also offer detailed information regarding the status of the submitted webpages e.g. if the page contains some errors. After an indexing time of 33 days one should analyse the submitted webpage to see if there are no errors preventing indexing, otherwise the webpage should be resubmitted.

4.8 Research limitations and further research

This study was limited to three search engines only - Google, Yahoo! and Bing. Other, albeit less used search engines could be included in a further study. Furthermore, the study was run for a limited time to allow indexing. Further studies could allow longer waiting times (more than three months, for example) to allow crawlers more visitation time.

This study paved the way for future studies to be executed in areas that include content scraping, as well as the relevancy of information displayed on the SERP. In addition, research needs to be done on the effects of keyword stuffing to webpage indexing.

References

Benczur, A.A., Erdelyi, M., Masanes, J., Siklosia, D. 2009. Web Spam Challenge Proposal for Filtering in Archives. In: *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb2009),* April 2009. Spain: Madrid: 438-444.

Borglum, K. 2009. Getting your website to show up in search ranking (Practice Management Q&A). *Medical Economics, GALE*, 86(14):30.

Castle, R. 2011. Google Sandbox Effect Sucks? Overcome the Google Sandbox and Earn TrustRank! Available WWW: http://www.roncastle.com/google-sandbox-aging-delay.htm (accessed 31 May 2011).

Chen, L. 2010. Using a two-stage technique to design a keyword suggestion system. *Information Research,* 15(1). Available WWW: http://informationr.net/ir/15-1/paper425.html (accessed 10 April 2011).

Egele, M., Kolbitsch, C., Platzer, C. 2009. Removing web spam links from search engine results. *Journal in Computer Virology,* 7(1):51-62.
Eisenberg, B., Quarto-vonTivadar, J., Davis, L.T., Crosby, B. 2008. *Always be testing: The complete guide to Google website optimizer*. Sybex: Indianapolis.
Erdelyi, M., Garzo, A., Benczur, A.A. 2011. Web spam classification: a few features worth more. In: *Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2011) In conjunction with the 20th International World Wide Web Conference,* March 2011. Hyderabad: India: 27-35.

Flosi, L.S. 2011. ComScore Releases April 2011 U.S. Search Engine Rankings. Available WWW: http://www.comscore.com/Press_Events/Press_Releases/2011/5/comScore_Releases_April_2011_U.S._Search_Engine_Rankings (accessed 26 May 2011).
Fox, W., Bayat, S.M. 2007. *A guide to Managing Research*. Cape Town: Juta & Co Ltd.
Kritzinger, W.T., Weideman, M. 2007. Key word placing in Webpage body text to increase visibility to search engines. *South African Journal of Information Management,* 9(1). Available WWW: http://www.sajim.co.za (16 August 2010).

Malaga, R.A. 2009. Web 2.0 Techniques for search engine optimization: Two case studies. *Review of Business Research,* 9(1):132-139.

Mathews, J. 2011. Get on Google front page: 2011 SEO tips. Jason Mathews. Available WWW: http://books.google.co.za/books?id=6n70oyVmgAQC&pg=PT7&dq=how+long+does+it+take+to+index+a+website&hl=en&ei=f1R3TYymIpC38QPhto2gDA&sa=X&oi=book_result&ct=result&resnum=8&ved=0CE8Q6AEwBw#v=onepage&q&f=false (accessed 09 March 2011).

Nade, J. 2010. Evidence of Google Sandbox Effect & Does a Drop from Pagerank 3 to Pagerank 0 Equal a Google Penalty? Available WWW: http://www.google.com/support/forum/p/Webmasters/thread?tid=0eb1ec75370dfdf9&hl=en (accessed 31 May 2011).

Parhizkar, M. 2010. Critical Analysis of Web Crawlers' Algorithms. Available WWW: http://skincarefreesamples.info/critical-analysis-of-web-crawlers-algorithms/ (accessed 26 May 2011).

Ron, B., Zsolt, K. 2011. The Role of Search Engine Optimization in Search Marketing. *Social Science Research Network*. Available WWW: http://ssrn.com/abstract=1745644 (accessed 15 April 2011).

Snack, K. 2011. Search Engine Market Share (April 2011). Available WWW: http://www.karmasnack.com/about/search-engine-market-share/ (accessed 07 April 2011).

SPSS Manual. 2007. SPSS Advanced Statistics 17.0: *Kaplan-Meier Survival Analysis.* Available WWW: http://www.hks.harvard.edu/fs/pnorris/Classes/A%20SPSS%20Manuals/SPSS%20Advanced%20Statistics%2017.0.pdf (accessed 18 March 2011).

Visser, E.B., Weideman, M. 2011. An empirical study on website usability elements and how they affect search engine optimisation. *South African Journal of Information Management* 13(1). Available WWW: http://www.sajim.co.za (accessed 07 April 2011).

Weideman, M. 2009. *Website Visibility: The theory and practice of improving rankings*. Oxford: Woodhead Publishing Limited.

Weideman, M. 2008. Internet Searching and other Research Challenges: Publish or Perish. Cape Town: Cape Peninsula University of Technology. Inaugural Speech.

Weideman, M. 2004. Empirical evaluation of one of the relationships between the user, search engines, metadata and websites in three-letter .com websites. *South African Journal of Information Management, 6*(3). Available WWW: http://www.sajim.co.za (accessed 16 June 2010).

Zahorsky, R. M. 2010. A Web trick catches a venerable law directory. *ABA Journal*. 96(2):32-33.

Zhang, J., Dimitroff, A. 2005. The impact of webpage content characteristics on webpage visibility in search engine results (Part I). *Information Processing and Management,* 41:665-690.