11-1-2005

Search engine exclusion policies : implications on indexing e-commerce websites

Fernie Neo Mbikiwa
Cape Peninsula University of Technology

SEARCH ENGINE EXCLUSION POLICIES: IMPLICATIONS ON INDEXING E-COMMERCE WEBSITES

by

Fernie Neo Mbikiwa

THESIS

Submitted in fulfilment
of the requirements for the degree

MAGISTER TECHNOLOGIAE

in

INFORMATION TECHNOLOGY

in the

FACULTY OF BUSINESS INFORMATICS

at the

CAPE PENINSULA UNIVERSITY OF TECHNOLOGY

Supervisor: Prof M Weideman

November 2005

DECLARATION

I, the undersigned, hereby declare that the work contained in this thesis is my own original work, except where stated otherwise. This thesis has not been submitted before for any degree or assessment at any other university, and all the sources I have used or quoted have been indicated and acknowledged by means of complete references.

STUDENT

F. Mbikiwa: ____________________ Date: ________________

ACKNOWLEDGEMENTS

This project would not have been possible without the constant support from the following:

- My father, Nicholas Mbikiwa, for his love, support, financial assistance and continued understanding.
- My supervisor, Prof. Weideman for his guidance and support.
- Thomas Nyirenda for proofreading this thesis.
- Shaundre Fortuin, for his assistance with the data collection.
- The National Research Foundation and Cape Peninsula University of Technology for financial assistance.
- The Web Factory for assistance with the website analysis.
- Ms Corrie Strümpfer for assistance with the methodology.
- Prof. Watkins for proofreading this thesis.

ABSTRACT

SEARCH ENGINE EXCLUSION POLICIES: IMPLICATIONS ON INDEXING E-COMMERCE WEBSITES

The aim of this research was to determine how search engine exclusion policies and spam affect the indexing of e-Commerce websites. The Internet has brought along new ways of doing business. The unexpected growth of the World Wide Web made it essential for firms to adopt e-commerce as a means of obtaining a competitive edge. The introduction of e-commerce in turn facilitated the breaking down of physical barriers that were evident in traditional business operations.

It is important for e-commerce websites to attract visitors, otherwise the website content is irrelevant. Websites can be accessed through the use of search engines, and it is estimated that 88% of users start with search engines when completing tasks on the web. This has resulted in web designers aiming to have their websites appear in the top ten search engine result list, as a high placement of websites in search engines is one of the strongest contributors to a commercial website's success.

To achieve such high rankings, web designers often adopt Search Engine Optimization (SEO) practices. Some of these practices invariably culminate in undeserving websites achieving top rankings. It is not clear how these SEO practices are viewed by search engines, as some practices that are deemed unacceptable by certain search engines are accepted by others. Furthermore, there are no clear standards for assessing what is considered good or bad SEO practices. This confuses web designers in determining what is spam, resulting in the amount of search engine spam having increased over time, impacting adversely on search engine results.

From the literature reviewed in this thesis, as well as the policies of five top search engines (Google, Yahoo!, AskJeeves, AltaVista, and Ananzi), this author was able to compile a list of what is generally considered as spam. Furthermore, 47 e-commerce websites were analysed to determine if they contain any form of spam. The five major search engines indexed some of these websites. This enabled the author to determine to what extent search engines adhere to their policies. This analysis returned two major findings. A small amount of websites contained spam, and from the pre-compiled list of spam tactics, only two were identified in the websites, namely keyword stuffing and page redirects. Of the total number of websites analysed, it was found that 21.3% of the websites contained spam.

From these findings, the research contained in this thesis concluded that search engines adhere to their own policies, but lack stringent controls for the majority of websites that contained spam, and were still listed by search engines. In this study, the author only analysed e-commerce websites, and cannot therefore generalise the results to other websites outside e-commerce.

RESEARCH OUTPUTS

The following are the research outputs produced by the author of this thesis during the study.

Type	Author	Title	Detail	Status
Journal Article	Weideman, M., Mbikiwa, F.	Implications of SPAM on e-commerce websites: a pilot study	South African Journal of Information Management (SAJIM)	To be submitted in December 2005.
Book	Weideman, M., Kritzinger, W., Mbikiwa, F. & Chambers R (Editors).	ICT research forum	Cape Town, South Africa. 01 May 2005, ISSN:1814-9812	Published
Poster	Mbikiwa, F. & Weideman, M.	A case study on Ananzi: Search engine exclusion policies and implications on indexing e-commerce websites	7th Annual conference on WWW applications	Abstract published in proceedings of WWW 2005. http:www.zaw3.co.za
Poster	Mbikiwa, F. & Weideman, M	Search engine exclusion policies and implications on indexing e-commerce websites	South African Institute of computer Scientists and Information Technologists (SAICSIT)	Abstract published in proceedings of SAICSIT 2005 ISSN:1-59593-258-5 p. 287.

SUPERVISOR:

Prof M Weideman: ______________________ Date: ______________

TABLE OF CONTENTS

CHAPTER 1 – BACKGROUND AND RESEARCH PROBLEM

CHAPTER 2 – LITERATURE REVIEW

CHAPTER 3 – RESEARCH METHODOLOY

CHAPTER 4 – RESULTS AND ANALYSIS

CHAPTER 5 – CONCLUSION

BIBLIOGRAPHY

LIST OF TABLES

LIST OF FIGURES

LIST OF GRAPHS

APPENDICES

GLOSSARY

CHAPTER 1
BACKGROUND AND RESEARCH PROBLEM

1.1 Introduction

The Internet and its associated technologies have introduced new ways of doing business, including electronic payments for goods and services. According to Peng, Trappey and Liu (2005: 476), the development of the Internet has made it essential for firms to adopt the Internet and e-commerce as a means to obtain a competitive advantage over other firms. Other authors have indicated that e-commerce has changed business processes by breaking physical barriers that were evident in traditional business. (Darch & Lucas, 2002: 148).

It is important for e-commerce websites to attract visitors, and one way of accessing these websites is through the use of search engines (Thelwall, 2000a: 150). Ranking high in search engines can be considered a competitive strategy. According to a study done by Nielsen (2004b), 88% of users start with search engines when assessing sites on the web. This has resulted in website owners adopting certain practices to ensure that they gain top rankings in search engine results. Fetterly, Manasse and Najork (2004), are of the opinion that high placement in search engines is one of the strongest contributors to a commercial website's success. According to Henzinger, Motwani and Silverstein (2002), the exclusion of a website from the top ten results of a search engine result list will lead to only a few users actually visiting a site.

Web designers aim to have their website listed in the top ten search results of search engines (Sullivan, 2002c). To achieve these high rankings, web designers use Search Engine Optimization (SEO) practices that are often adopted to assist websites to achieve top rankings (Machill, Neuberger & Schindler, 2003: 54). Henzinger *et al.* (2002) substantiate this argument by stating that some web designers deliberately manipulate their ranking on search engines. It is not clear how search engines view some of these SEO practices. In addition, there are no clear standards for assessing what are considered good or bad SEO practices. Most search engine policies are vague as to what is considered spam (Sullivan, 2001b). According to Perkins (2001), spam is defined as "any attempt to deceive a search engine's relevancy ranking". This type of spam is different from the traditional email spam, which refers to sending a bulk of messages to different email addresses without the consent of the email account holder (Adam, 2002: 91).

Gyöngyi and Garcia-Molina (2005) are of the opinion that the amount of search engine spam has increased. These authors also state that the consequences of spamming include the fact that search engine indices are inflated with useless pages, resulting in an increase in the cost of each processed query.

1.2 Research Background

There is a low level of understanding of how certain SEO practices or spam lead to indexing or exclusion of websites by search engines. It is also unclear what standards various search engines apply in dealing with SEO practices. Most search engine policies are vague as to what is considered spam (Sullivan, 2001b). Some practices that are deemed unacceptable by one search engine will not necessarily be deemed as such by another search engine, which then results in web designers not having clear guidelines as to what is considered to be spam.

According to Thelwall (2001: 114), the higher a website appears in a search engine's ranking, the more exposure and traffic that particular website will receive. Research done by Zhang and Dimitroff, (2004: 666) indicates that most users examine only the top ten results in search engine result lists while only 1% of users look at websites that appear on the third page and beyond of search engine result lists. Website designers want to obtain the highest rank among competitors and often go to great lengths to ensure that their websites receive a top ranking (Alimohammadi, 2004: 222). These designers at times deploy questionable SEO practices, also known as spam, like doorway pages, cloaking, keyword stuffing and page jacking to increase their website's traffic. Some of these practices aim to trick search engines into giving undeserving websites higher rankings. Instead of lifting the rankings, this could result in these websites being excluded from search engine indices (Anon, 2002; Sullivan, 2003a).

As part of the continuing battle against those that apply spam tactics, many search engine designers have tightened their site's eligibility policies (Dahm, 2000). Penalties are often imposed on websites that do not comply with these policies. However, not all search engines are strict about penalties resulting in websites contravening these policies as penalties are not readily imposed by search engines (Anon, 2002). According to Henzinger *et al.* (2002), without penalties, the quality of rankings in search engines suffers and this subsequently leads to users not getting what they want from search engines.

Research done by Sullivan (2001b) has indicated that while some search engines have guidelines against cloaking, others not only allow it but practice it themselves. Sullivan (2001b) further claims that he does not consider cloaking to be spam because it does nothing to satisfy a search engine's algorithm. Sullivan (2003a) goes on to say that cloaking is just a way of delivering targeted content, and unlike a doorway page, it is not spam.

According to Oppenheim, Morris, Mcknight and Lowley (2000: 193), the major problems of evaluating search engines is that the search mechanisms and user interfaces are always changing and developing, making it difficult to have a specific methodology for the evaluation of search engines.

1.3 Research problem

The research problem addressed in this thesis is the fact that there are no standard guidelines for e-commerce website designers with respect to the interpretation of search engine exclusion policies. Furthermore, this study also attempts to explore gaps between the search engine policies and their indexing criteria. The search engine policies that will be examined are four international search engines namely Google, Yahoo! AltaVista and AskJeeves, as well as one local search engine namely Ananzi. The four international search engines are considered to be the top search tools (Sullivan, 2004a), while Ananzi is said to be South Africa's largest search engine (Ananzi, 2005c). The exclusion policies of these search engines will be evaluated and compared.

1.4 Aim of the thesis

Ranking high of websites is of importance to website designers. The aim of this research was to determine how search engine exclusion policies and spam affect the indexing of e-Commerce websites. This study will show the level of adherence of search engines to their own exclusion policies. It will also assist web designers to identify which SEO tactics should be avoided by:

- Reporting on the policies of Google, Yahoo!, AskJeeves, AltaVista and Ananzi.
- Providing a guideline for web designers that will enable them to identify which visibility enhancing factors to avoid when using the above mentioned search engines as their main source/provider.

It is furthermore envisaged that such a publication will help in the policy analysis and establishment of standardisation of search engine policies for e-commerce website designers.

1.5 Research question

The following will serve as the research question in support of the problem statement in this thesis:

- How do the exclusion policies of the five search engines impact the chances of indexing a website which contravenes the accepted SEO practices of the search engines?

1.6 Investigative questions

According to Cooper and Schindler (2003: 75), investigative questions reveal information needed to answer a research question. The following investigative questions were formulated in support of the research question:

- What do search engines consider as spam?
- What are the implications of spam on indexing e-commerce websites?
- How do search engines adhere to/deviate from their own policies?

1.7 Research design and methodology

In this thesis, a qualitative research strategy was deployed. Data was collected, analysed and quantified, as described below.

1.7.1 Data collection and sample size

E-commerce websites were obtained from the Cape Chamber of Commerce website as well as www.onlineshopping.co.za, which is South Africa's online shopping network. More websites were retrieved from Ananzi's shopping and auctions page. The author ended up with a final population of 4985. From this list the author applied judgemental sampling to extract fully functional e-commerce websites. The resultant list contained 136 websites, details of which are contained in Appendix A. A filtering process was applied by making use of keyword verification software provided by Marketleap. This software determined whether the websites were indexed by the search engines. The limitation of this software was that it did not analyse Ananzi and Askjeeves, culminating in the author manually visiting these search engines to determine whether the 136 sample websites were listed or not. Furthermore, random sampling was applied to these websites yielding the final list of 51 (see Appendix B). However, during the analysis of the websites, four were no longer operational (see Appendix D), culminating in the study focussing on five search engines as well as 47 websites.

1.7.2 Data analysis

For each search engine, a checklist of what is considered spam was compiled after evaluation of the search engine policies. More spam tactics were derived from the literature review. The list was then tabulated. However, there were some spam tactics that were listed under certain search engines while not listed in others. To reduce the list of prohibited SEO practices to a manageable size, the author applied judgemental sampling by selecting practices which were detectable by humans, and not only by search engine spiders. The 47 pre-selected e-commerce websites were checked against the prohibited SEO practices. The result of the checklist process was compared to the expected outcome, assuming all search engines adhered to their policies. The implications of the spam techniques were determined by examining whether the websites that have been indexed by the search engines contained any spam that the search engines prohibit (see Appendix C), as well as whether or not the websites that were not indexed contained any spam.

1.7.3 Results and conclusions

The results of this study were obtained from determining whether any of the websites contained what is considered as spam. The analysis of the 47 websites showed two major findings. In the first instance, a small number of the analysed websites contained spam, and in the second instance there were only two types of spam identified, namely keyword stuffing and page redirects. From these findings, it was concluded that to a greater extent the five major search engines comply with their exclusion policies, but lack stricter controls for the minority of websites containing what was identified as spam. A further finding returned that not all search engines registered all the websites that contained spam.

1.8 Limitations of the research

This research was limited to the following:

- Five search engines, namely Google, Yahoo!, AskJeeves, AltaVista and Ananzi.
- Forty-seven fully functional randomly selected e-commerce websites.
- As search tools are continuously changing and developing, this study would only be concluded valid for a period of time.
- Some SEO practices that have been identified by the author as spam were obtained from the identified search engine policies as well as the literature review, and it is possible that certain practices were inadvertently being excluded from the evaluation.
- Spam techniques used in the analysis will be limited to those techniques that can be detected by humans, and not only search engine crawlers.

1.9 Chapter Overview

This thesis comprises of the following chapters:

- **Chapter 1:** This chapter contains a high level background to the research problem as well as the research process to be followed. The research design and methodology is discussed and the research constraints listed.

- **Chapter 2:** Chapter 2 will focus on an in-depth literature review.

- **Chapter 3:** In this chapter, the survey environment will be analysed in detail. The approach to data collection will be explained and various research methods listed, while justifying the chosen research method for this thesis.

- **Chapter 4:** In this chapter, the data that has been collected will be analysed. The results will be mapped to the literature review which was conducted in Chapter 2.

- **Chapter 5:** A conclusion will be drawn based on the analysis of the data in Chapter 4.

1.10 Conclusion

The growth of the Internet has driven companies to adopt e-commerce and trade on line. However, a website that cannot be found by users is practically worthless. With the assistance of search engines, users are able to find these websites on the Internet. E-commerce web designers have realised the commercial potential of appearing in the top result list of search engines. Strategies are sometimes adopted to increase the visibility and rankings of the websites in search engines.

As a rule, search engines have policies that determine which websites will be included in their index. Adherence to these policies has implications as search engines warn against web designers that do not adhere to these policies and thus impose penalties on such websites. Penalties imposed are sometimes as severe as banning websites from search engine results, and subsequently impacting upon the website’s commercial success of being on the Internet.

In this thesis, the author will establish to what extent search engines adhere to their own policies as well as the implications that spam has on indexing websites. A guideline for web designers on which visibility enhancing tactics should be avoided will result from this research.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

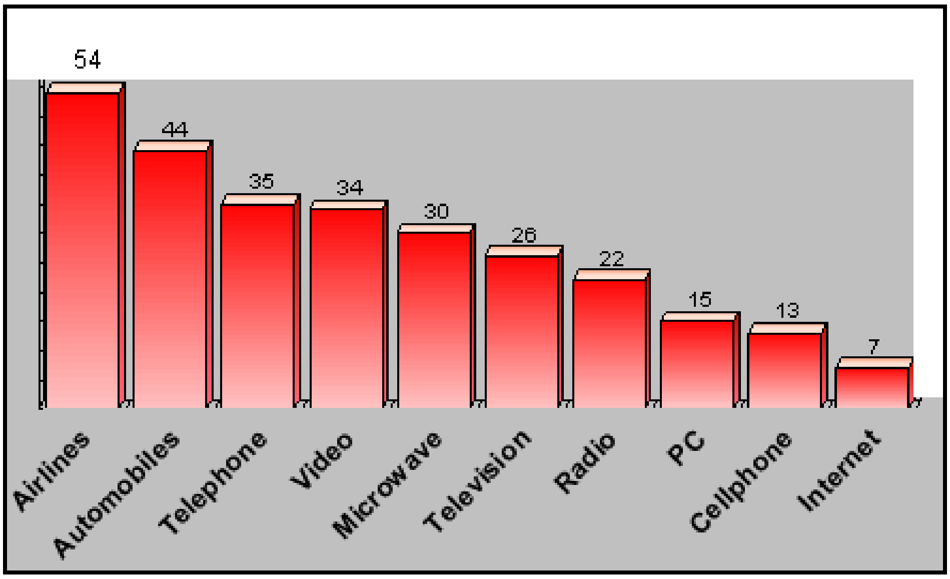
The Internet, which was first developed during the 1960's by the US Department of Defence Research Project Agency, was originally designed to link researchers and defence contractors. Its use spread to academics in the 1970's, to interlink research groups across different universities (Wilson, 2000: 99). The use of the Internet spread beyond academic and military communities only in the late 1980's (Boyes & Irani, 2004: 191). Some authors have defined it as "a vast computer network interconnected globally" (Palumbo & Herbig, 1998: 253).

According to Brinkley and Burke (1995: 3), the Internet was primarily a communications tool, but has grown to become a very important information resource, which has developed and changed exponentially over the years. The Internet, which is also referred to as the web, was built on a foundation of a set of standards that are set by the WWW Consortium (W3C) (Hart & Rolletschek, 2003: 11). Research done by Van der Walt (1998) has classified the Internet as "a means of creating information that can have a global reach". The current estimate of Internet users is now 6 billion world wide, with 56.4% of the users coming from Asia, while Africa carries the lowest percentage of Internet users at just 1.8% of the world population (Anon, 2005b). This implies that the majority of Internet users can be found within the developed countries of the world as illustrated in Table 2.1.

TABLE 2.1: World Internet usage population statistics (**Source:** Anon, 2005b).

WORLD INTERNET USAGE AND POPULATION STATISTICS						
World Regions	**Population (2005 Est.)**	**Population % of World**	**Internet Usage, Latest Data**	**Usage Growth 2000-2005**	**% Population Penetration**	**World Users %**
Africa	896,721,874	14.0 %	16,174,600	258.3 %	1.8 %	1.7 %
Asia	3,622,994,130	56.4 %	323,756,956	183.2 %	8.9 %	34.5 %
Europe	731,018,523	11.4 %	269,036,096	161.0 %	36.8 %	28.7 %
Middle East	260,814,179	4.1 %	21,770,700	311.9 %	8.3 %	2.3 %
North America	328,387,059	5.1 %	223,392,807	106.7 %	68.0 %	23.8 %
Latin America/ Caribbean	546,723,509	8.5 %	68,130,804	277.1 %	12.5 %	7.3 %
Oceania / Australia	33,443,448	0.5 %	16,448,966	115.9 %	49.2 %	1.8 %
WORLD TOTAL	**6,420,102,722**	**100.0 %**	**938,710,929**	**160.0 %**	**14.6 %**	**100.0 %**

According to Vaughan (1999: 89), the Internet has been classified as the most significant development in telecommunications since the invention of the telephone. It reached a market share of 25% in just seven years, making it the fastest growing technology surpassing that of the television, which took 26 years and the telephone, which took 35 years to reach the same market share (Anon, 2005b). Graph 2.1 illustrates the comparative timeframes of several technologies including the Internet, in reaching 25% market share.



GRAPH 2.1: Time technologies took to reach 25% market share (**Source:** Singh, 2002).

Over the years of technology development and use, the term web has often been used to describe the Internet. For the purpose of this research the terms the web and the Internet will be used interchangeably.

According to Green (2000: 124), the web can be divided into two distinct elements, namely the invisible web and the visible web.

2.2 The visible and invisible web

2.2.1 The visible web

Sherman and Price (2002: 55) define the visible web as "webpages that search engines have chosen to index". All the information that is retrievable via search engines forms part of the visible web. Webpages that have been created manually by web designers, also known as static webpages make up the visible web (Green, 2000: 124).

2.2.2 The invisible web

McGuigan (2003: 68), defines the invisible web as "content that exists within the world wide web but cannot be located through the use of search tools", while Ru and Horowitz (2005: 249) define the invisible web as ‘a vast collection of information that is accessible via the world wide web, but is not indexed by search engines’. These authors further state that the invisible web is made up mostly of the content of many specialized databases, and that these databases are only accessible by filling out a form on a webpage and submitting it to the databases. While access to these databases is sometimes restricted, research has shown that about 95% of the invisible web is publicly accessible (Ru & Horowitz, 2005: 250). A study done by BrightPlanet (2005) indicates that the invisible web is actually 500 times bigger than the visible web, while McGuigan (2003: 68) also found that the invisible web is significantly larger that the visible web.

There are a plethora of reasons why search engines do not index some pages, which ultimately culminate in being part of the invisible web. McGuigan (2003: 69) states that some reasons why some webpages are not retrieved by search engine spiders include either programming of the spiders, the formats of the databases or the formats of some scanned images. The same author also cites search engine policy decisions as reasons for the existence of the invisible web, as well as some pages that have been programmed to be invisible by the web designers. Research done by Ru and Horowitz (2005: 250) indicates that another reason for the existence of the invisible web is the fact that search engine spiders do not crawl dynamically generated pages. These authors also state that certain audio/video clips, flash movies and documents in non-standard formats are often not indexed by search engines while Sherman and Price (2002: 59) argue that much of the invisible web is hidden because search engines deliberately choose to exclude certain formats of web content. Despite all the reasons why search engines do not index some webpages, it is possible that some of these pages contain potentially essential information. (Sullivan, 2003b) substantiates this argument by stating that by indexing more webpages, search engines are likely to retrieve more relevant results.

As indicated by research done by Van der Westhuizen (2001), Adobe Portable Document Format (PDF) files were previously part of the pages that were not indexed by search engines until Google introduced a feature that enabled searchers to find information contained in PDF format.

The fact that the invisible web exists indicates to an extent the indexing capabilities of search engines. It is therefore essential for web designers to ensure that they are aware of these capabilities, and design websites that will form part of the visible web. Dynamically generated pages are one of the few reasons for the existence of the invisible web and therefore should be avoided by web designers. Due to the size of the invisible web, it is possible that search engines are excluding relevant information from their indices (BrightPlanet, 2005). Studies have shown that in some instances the search interfaces of invisible websites are indexed, but Ru and Horwitz (2005: 261) argue that this is not sufficient as there may be content within the site that users are unable to access.

2.3 Search tools

2.3.1 Background and History

The average user of the Internet is faced with the daunting task of having to find relevant information on the Internet (Machill *et al.*, 2003: 52). Search engines have made this task much easier. Before the existence of search engines, the task was almost impossible. This transposed into the requirement to know the exact Uniform Resource Locator (URL). A URL is described as the address of a file on the Internet (Thurow, 2003: 14). Search engines help assist the user with retrieving information from the Internet (Machill *et al.,* 2003: 52).

One of the earliest search engines to be developed was Archie. This search engine allowed keyword searching of file names from a database that was accessible via a File Transfer Protocol (FTP). FTP refers to the protocol that governs the transfer of files across the Internet. This database was also accessible via a network of Archie servers that offered local access to a copy of the Archie database. Searching via Archie retrieved references to files that were stored on many different locations. The nearest copy of this file would then be retrieved via FTP. Another early search engine to be developed was the Wide Area Information Server (WAIS). This search engine allowed keyword searching of files like documents as well as mailing lists. Unlike current search engines, WAIS did not allow Boolean searching for documents, but instead relied on best-match search techniques. These techniques presented results in relevance-ranking order (Poulter, 1997: 131-133).

Before the existence of the web, a protocol called Gopher was used to publish information on the Internet. Veronica, a search engine primarily designed for Gopher was then developed. Unlike WAIS, this search engine supported Boolean searching and searched Gopher menu item descriptions that were located in databases around the world. Multiple search queries that did not contain any Boolean operators automatically defaulted to the AND operator. Another functionality of the Veronica search engine was that searchers had the option of limiting the resources retrieved (Poulter, 1997: 133). Gophers were in existence for just a couple of years when they were superseded by the development of the web, which ultimately led to the demise of the Veronica search engine. Even though the web was successor to the Gopher, it still retained some of Gopher functionalities. Poulter (1997: 134) is of the opinion that the search features of the web browser clients had the same functionality as Gopher. However, this author also states that one new feature of the web was that it could be accessed by a client running under a Graphical User Interface (GUI), while webpages could display text, graphics as well as other multimedia resources. Unlike Gopher which had only one dedicated search engine designed for it, the development of the web resulted in the development of a multitude of search engines (Poulter, 1997:

134). This sequence of events in development of search engines is illustrated in Table 2.2.

TABLE 2.2: A timeline of Internet Search technologies (**Source:** Sherman & Price, 2002: 15).

Year	Search Service
1945	Vannevar Bush Proposes “MEMEX”
1965	Hypertext Coined by Ted Nelson
1972	Dialog – First Commercial Proprietary System
1986	OWL Guide Hypermedia Browser
1990	Archie & the Web
1991	Gopher
1993	ALIWEB, WWWWander, JumpStation, WWWWORM
1994	ELNet Galaxy, WebCrawler, Lycos, Yahoo!
1995	Infoseek, SavvySearch, AltaVista, MetaCrawler, Excite
1996	HotBot, LookSmart
1997	NorthernLight
1998	Google, InvisibleWeb.com
1999	FAST
2000+	Hundreds of search tools

2.3.2 Types

Search engines and directories are considered as two of the most important tools for locating/retrieving information (Thelwall, 2002b: 101) and are also seen as the primary tools for users to find websites (Sullivan, 2002c). The following search engines have been classified as top choices for users when choosing search tools.

- Google.
- Yahoo!

- AskJeeves.
- AltaVista.
- AOL.
- MSN (Thelwall, 2002b: 101; Nielsen, 2004a; Sullivan, 2004c).

The tools that receive the most traffic according to Nielsen/Netratings (2005) are reflected in Table 2.3.

TABLE 2.3: Top search tools in 2005 (**Source:** Nielsen/Netratings, 2005).

Position	Provider	June 05	July 05
1	Google Search	47.0%	46.2%
2	Yahoo! Search	22.3%	22.5%
3	MSN Search	12.5%	12.6%
4	AOL Search	5.5%	5.4%
5	My Way Search	1.8%	2.2%
6	AskJeeves Search	1.8%	1.6%
7	Netscape Search	0.9%	1.6%
8	Dogpile.com Search	0.8%	0.9%
9	iWon Search	1.0%	0.9%
10	EarthLink Search	0.8%	0.8%

According to Sullivan (2002d), search tools can be categorised as either crawler-based search engines, directories or meta search engines.

2.3.2.1 Crawler-based search engines

Crawler-based search engines create their listings automatically (Sullivan, 2002b). To compile their databases, search engines rely on computer programs called robots or spiders which crawl the web by following links and indexing each site they visit (Shenton, 2001). Search engines have two primary functions namely, to index as many websites as possible, as well as to retrieve the most appropriate websites and pages requested by a user (Sekhar, 2002: 8). Furthermore, they are used to locate information on the web, whether relevant or not (Alimohammadi, 2003: 238).

Crawler-based search engines follow a set of rules known as algorithms to determine relevancy of a search query to a webpage. Each search engine uses its own criteria to decide what to include in its database. Some search engines index each individual page in a website, while others index only the main page of a website (Anon, 2001). Different search engines have different indexing strategies and ranking algorithms (Zhang & Dimitroff, 2004: 666).

Crawler-based search engines have three major elements, namely the search engine spider, the index and the search engine software (Sullivan, 2002b).

- **Search engine spider**

The spider, also referred to as a crawler or robot, is an automated program which visits a webpage. It then reads the webpage and follows links from that webpage to other pages (Sullivan, 2002b). According to Thurow (2003: 15), search engine spiders are continuously crawling the web, resulting in their indices being constantly updated. If a webpage has no in-bound links pointing to itself, it is highly likely that spiders will not find it (Sullivan, 2002b). Furthermore, the only way a new page that has no external links to it can get into a search engine index is through a manual request for inclusion. The URL of the page is sent to the search engine companies with a request for it to be included.

- **Index**

According to Sullivan (2002d), the detail collected by the spider is deposited into the second part of a search engine known as the index. It contains a copy of every webpage that the spider finds. The search index in addition contains full-text indices of webpages. Green (2000: 126) refers to the index as the main element of any search engine. This author further states that the index is what a user interrogates when searching for webpages on the web. When a user performs a search query to a search engine, the user is actually searching the full-text index of the retrieved pages by the spider and not the web itself (Thurow, 2003: 15). Green (2000: 126) expresses the same opinion by stating that it is impossible to search the web directly and that all search engines do is search their compiled databases of indexed websites. Some webpages that have been crawled by the spider take a while to be added to the index and until the webpage is indexed, it is not available to search engine users (Sullivan, 2002b).

- **Search engine software**

The search engine software is the component that matches a search query and retrieves pages that it believes are the most relevant (Sullivan, 2002b). This software matches the words that have been typed as search keywords with a webpage that is most likely to contain the information that the user is looking for. The search engine software is also known as a query processor (Thurow, 2003:16). The ranking of the matching websites is determined by the search engine's algorithm, and these differ from search engine to search engine (Green, 2000:126).

According to Thurow (2003:15), search engine spiders are constantly crawling the web. The size of search engines indices cannot therefore remain constant. The crawling methodology of webpages by search engine spiders is depicted in Figure 2.1.

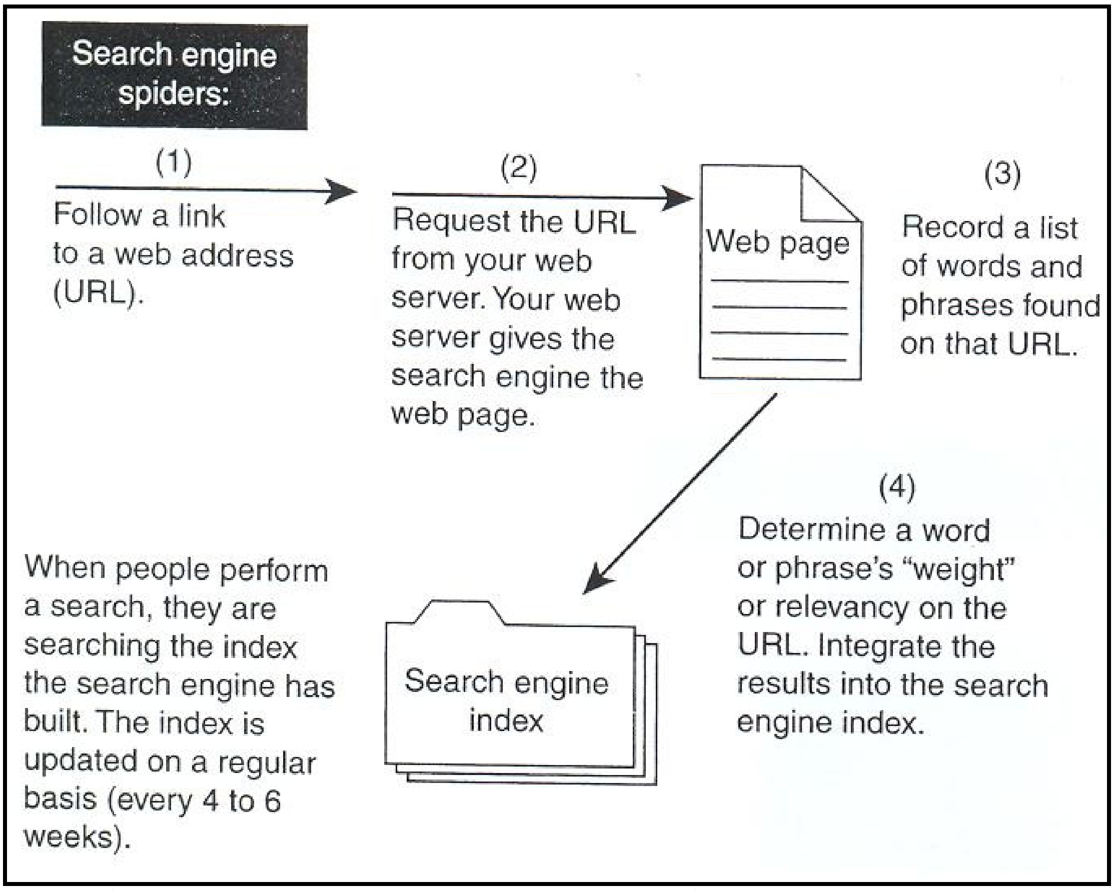

Search engine spiders:
(1)
Follow a link to a web address (URL).
(2)
Request the URL from your web server. Your web server gives the search engine the web page.
Web page
(3)
Record a list of words and phrases found on that URL.
(4)
Determine a word or phrase's "weight" or relevancy on the URL. Integrate the results into the search engine index.
Search engine index
When people perform a search, they are searching the index the search engine has built. The index is updated on a regular basis (every 4 to 6 weeks).

FIGURE 2.1: How search engines crawl web pages (**Source:** Thurow, 2003: 15)

2.3.2.2 Directories

According to Sullivan (2002d), directories are often equated to search engines, yet they are completely different. Unlike search engines, directories use human editors which review and index websites. Green (2000: 125) has described a directory as 'a predefined list of websites compiled by human editors and categorized according to a certain topic or subject'. The same author claims that websites that are listed in a directory are likely to stay there longer than websites that have been indexed by a search engine, due to the manual process involved in compiling a directory.

One of the earliest directories to be developed was the World Wide Web Virtual Library. This directory presented an alphabetical index of subjects. Each of these subjects had links to other sites and also contained brief descriptions of the contents of the sites they were linked to (Poulter, 1997: 136).

When a site is submitted to a directory, the site is reviewed by a human editor, who also determines whether the website should be included in the directory. Websites that are listed in directories are sorted per category. (Thurow, 2003: 31).

Sullivan (2002d) argues that even though web directories tend to have smaller indices, they tend to be 'cleaner'. Another author states that the smaller indices have resulted in search results of a directory being supplemented with additional results from a search engine partner, also known as fall through results. Furthermore, unlike search engines that list individual webpages, directories list the entire website (Thurow, 2003: 27-28).

Although directories perform the same functions, they are subject to variations. According to Poulter (1997: 136), directories vary in the subjects they cover, in the way the directory structures are organized, and in terms of the content descriptions that they use. According to another author, there are several main categories in a directory listing with each main category having a sub category and subsequent sub categories (Anon, 2001). Table 2.3 lists some differences between search engines and directories

TABLE 2.4: Differences between search engines and directories (**Sources:** Anon, 2001; Sherman and Price, 2002: 36).

Directory	Search engine
Edited by a human reviewer	Crawled by a robot 'spider'
Meta-tags are not considered	Meta- and title-tags considered
HTML code not very important	HTML code extremely important
Most allow paid submission	Few allow paid submission
Quality of site very important	Quality of site not very important
Small in size	No size restrictions
Often points to a website home page and not deeper	Typically indexes the full text of pages

According to Thurow (2003: 29), top directory listings are based on the following criteria:

- The directory category.
- The website's title.
- The website's description.

Research done by Thurow (2003: 32-33) has produced the following as some characteristics that directory editors consider to determine whether or not to include a website in the directory.

- **Unique content**

A website should contain some unique content as directory editors do not want to place sites with identical information in the same category. The content of the website should add value to the directory's category. A web designer can prove that the content is unique by either using the description section or the extra comments field in the submission form.

- **Most appropriate category**

A website's content must accurately reflect the category that the site is to be listed under, and it should be similar, and unique to the other sites listed in the same category.

- **Legitimate organization/company**

Editors want legitimate companies listed in their commercial categories. Thurow (2003: 32) states that having a virtual domain, e.g. (www.companyname.com) is usually an indication that the company is legitimate. However, directory editors also have extra requirements for e-commerce websites. They require that these sites should have items such as secure credit card processing, return policies or money back guarantee and a physical address.

- **Accurate description**

The description that is submitted to directory editors should accurately reflect the content of the website that is being submitted.

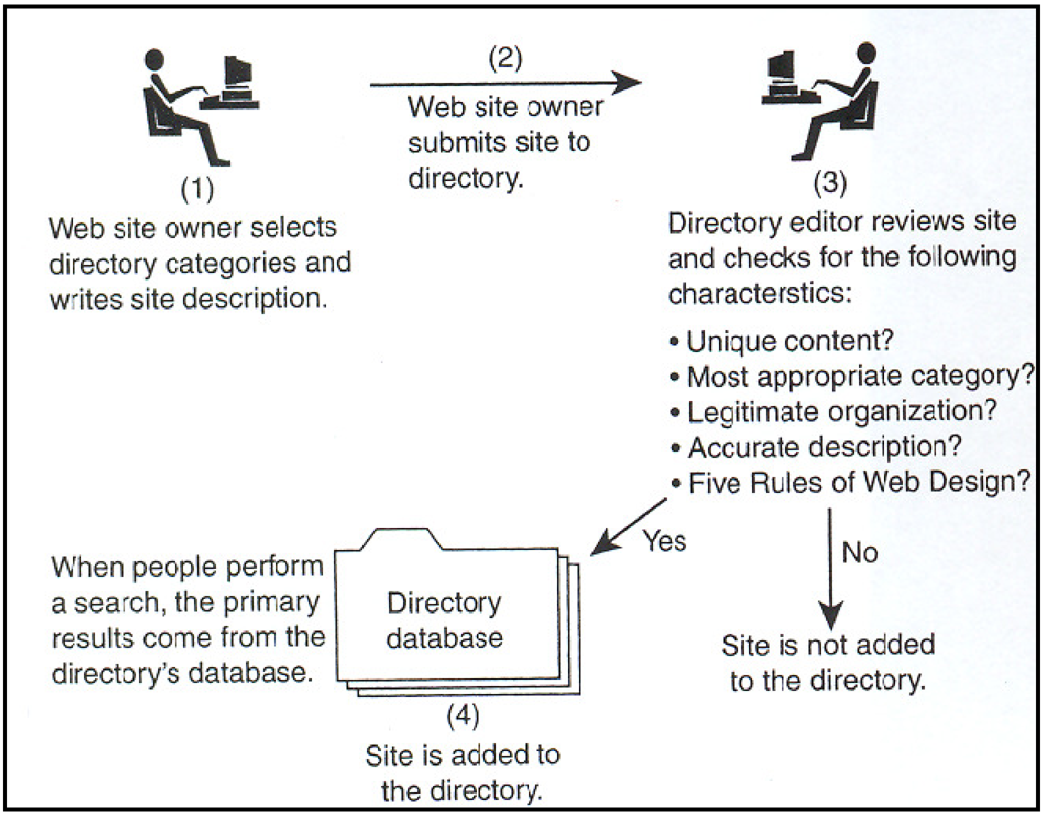

(2)
Web site owner submits site to directory.
(1)
Web site owner selects directory categories and writes site description.
(3)
Directory editor reviews site and checks for the following characterstics:
• Unique content?
• Most appropriate category?
• Legitimate organization?
• Accurate description?
• Five Rules of Web Design?
Yes
No
When people perform a search, the primary results come from the directory's database.
Directory database
(4)
Site is added to the directory.
Site is not added to the directory.

FIGURE 2.2: Directory Submission (**Source:** Thurow, 2003: 31).

Figure 2.2 depicts the process of submitting a website to a directory (Thurow, 2003: 31).

2.3.3 Search engine submission

2.3.3.1 Search engines

According to Sullivan (2004c), Google is one of the most important crawler-based search engines. Sullivan points to the fact that the best way to get a website listed in crawler-based search engines is to build links to that particular website. Furthermore, crawlers follow links, so the more relevant links there are pointing to a website, the more chances there are that the crawlers will find it. However web designers need to exercise caution as too many links can be passed off as spam by some search engines (Sullivan 2004c).

2.3.3.2 Directories

Shenton (2001) argues that unlike search engines that look at a website's code or other elements in the likes of link popularity, directory editors are more concerned with the actual content of the website. The same author states that when submitting to a directory, the website needs to be fully functional, well designed and content rich for it to be listed (Shenton, 2001). The same author lists the following as tactics that can get a website rejected by a directory, and should be avoided:

- Temporary sites.
- Sites that are still under construction.
- Dead links.
- Sites with little content or consisting of only lists of links.

Furthermore, Shenton (2001) has identified the following as strategies that will increase the chance of a website being accepted by a directory:

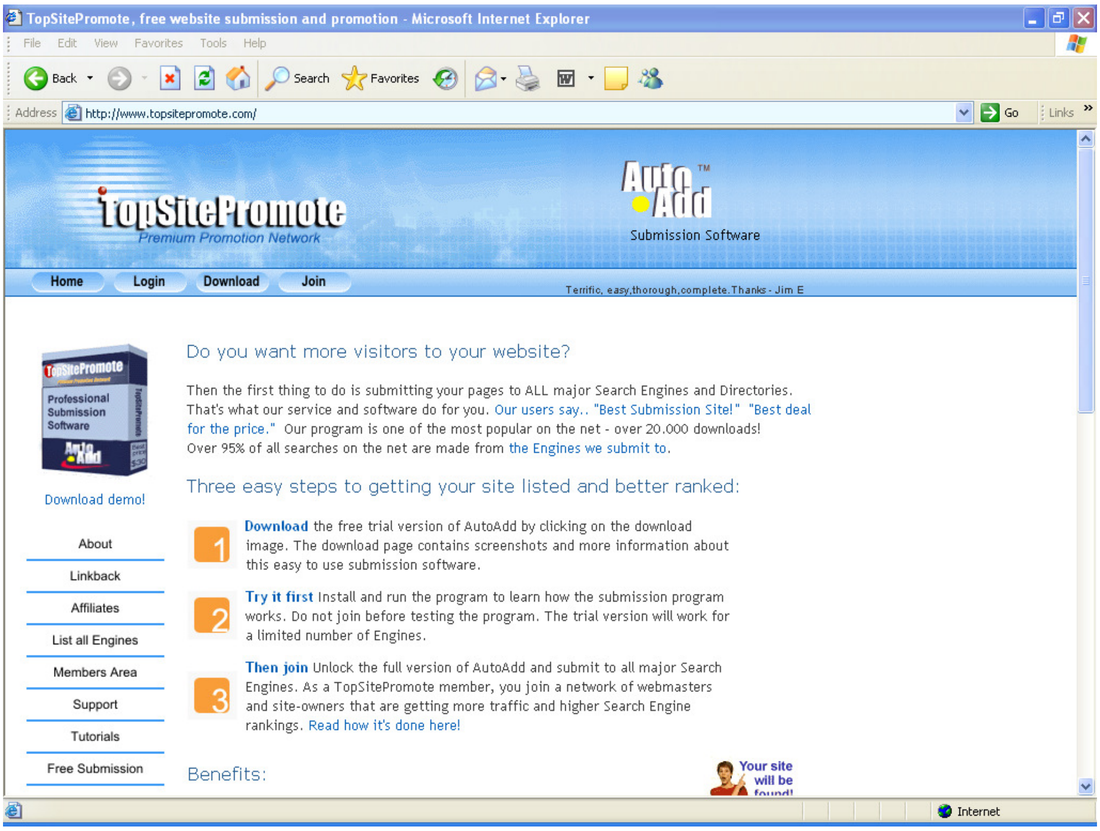
- Fast-loading well designed pages.
- Useful content.
- Full functionality of the website.
- Interactivity of the website.
- An appropriate category and description.

2.3.3.3 Automatic submission

According to Dunn (2004), automatic submission is a process of using software to automatically submit websites, sometimes submitting the same website repeatedly. This author also points out that the repeated submission of a website maybe classified as spam, while another author has stated the following;

> "Auto-submission software is (and always have been) a violation of the submission procedure. Site submitted automatically are flagged and deleted after the submission is accepted without notification. Persistent automatic submission may force us to ban you form dmoz site, so we can provide resources to real human beings" (Anon, 2004b).

Figure 2.3 represents an automatic submission website.


TopSitePromote, free website submission and promotion - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.topsitepromote.com/
Go
Links
TopSitePromote
Premium Promotion Network
AutoAdd™
Submission Software
Home
Login
Download
Join
Terrific, easy,thorough,complete.Thanks - Jim E
TopSitePromote
Professional Submission Software
Download demo!
About
Linkback
Affiliates
List all Engines
Members Area
Support
Tutorials
Free Submission
Your site will be found!
Internet

Do you want more visitors to your website?

Then the first thing to do is submitting your pages to ALL major Search Engines and Directories. That's what our service and software do for you. Our users say.. "Best Submission Site!" "Best deal for the price." Our program is one of the most popular on the net - over 20.000 downloads! Over 95% of all searches on the net are made from the Engines we submit to.

Three easy steps to getting your site listed and better ranked:

1 **Download** the free trial version of AutoAdd by clicking on the download image. The download page contains screenshots and more information about this easy to use submission software.

2 **Try it first** Install and run the program to learn how the submission program works. Do not join before testing the program. The trial version will work for a limited number of Engines.

3 **Then join** Unlock the full version of AutoAdd and submit to all major Search Engines. As a TopSitePromote member, you join a network of webmasters and site-owners that are getting more traffic and higher Search Engine rankings. Read how it's done here!

Benefits:

FIGURE 2.3: Automatic submission website (**Source:** Anon, 2000).

2.3.4 Indexing websites

Indexing of websites is important to ensure that websites appear in search engine result lists. Wilson (2002) is of the opinion that indexing is an important element of information retrieval because it enables users to search for information from websites that have been indexed, by using keywords that are relevant to a particular website. The same author states that indexing involves the process of making documents retrievable. Garofalakis, Kappos and Makris (2002: 43) have indicated that search engines index websites using their own techniques, also known as algorithms.

According to Thurow (2003: 15), a search engine index contains full-text indices of webpages. This index is compiled as a result of spiders that retrieve pages to include in the index. Research done by Thelwall (2002a: 124) indicates that search engines index only a fraction of the web. Thelwall further states that it is important to consider the algorithm of search engines, because it represents the criterion that is used to determine whether to index websites or not (Thelwall, 2002a: 125). According to Thurow (2003: 15), search engines update their indices about every four to six weeks while Sullivan (2004g) is of the opinion that search engines automatically visit webpages to compile their listings.

Wilson (2002) has classified indexing under the following categories.

- Automatic indexing

  Under this category, webpages are indexed by search engine spiders. These spiders crawl the web by following links and indexing websites that they visit. Although automatic indexing has less room for error than manual indexing, it is more complex, and therefore it is up to programmers to ensure that the software generates an effective index. The high cost and impracticality of manual indexing has led to automatic indexing becoming the most common indexing method (Wilson, 2002; Sullivan, 2002b).

- Manual indexing

  According to Wilson (2002), manual indexing is performed by humans, who identify important keywords in a document and match these keywords to a standardized index vocabulary. The human indexer predicts which keywords users are likely to search for. Wilson further states that manual indexing is time consuming and costly, while automatic indexing is cheaper and more consistent (Wilson, 2002).

2.3.5 Payment systems

Initially, usage of search engines was free, but advertising on these tools was implemented to compensate for the cost of maintaining them (Poulter, 1997: 139). According to Oppenheim *et al.* (2000: 190), search engines are generally funded by advertising revenue. Some websites have turned to paid advertising in an attempt to increase traffic to their sites (Duffy, 2005: 162). Research done by Goh and Ang (2003: 88) indicates that search engines adopt payment systems to increase their revenue and maintain profitability. The following are some advertising practices that are carried out by both search engines and directories.

2.3.5.1 Paid Placement

This model, also known as pay for placement or Pay Per Click (PPC) guarantees top rankings for websites. Website owners bid for keywords related to their websites, and undertake to pay a fixed amount per resultant click (Weideman, 2004: 907; Sullivan, 2004b). Every time a searcher clicks a link from search results to a participating website, the site's account is charged. In situations where there is no bid for certain search terms, searches from these terms produce fall through results from search engine partners (Thurow, 2003: 18).

Some authors have deemed this model the most common advertising model for search engines (Moxley, Blake & Maze, 2004: 61). According to Weideman (2004: 907), the main advantage of paid placement is that web designers do not have to be concerned about optimizing their websites for higher rankings because there is a high ranking guarantee, on condition that the bidding price is paid.

2.3.5.2 Paid submission

Under paid submission, Web designers pay a certain fee to have directory editors review their website. This model does not guarantee high ranking, but only guarantees that editors will review websites more quickly. The main advantage of this model is that a website will be evaluated faster as opposed to being submitted via the directory's submit site option. However, a site still needs to be optimized for higher rankings, if indexed (Thurow, 2003: 183; Sullivan, 2004e).

2.3.5.3 Paid Inclusion

Paid inclusion, also known as pay-for-inclusion involves a process where a webpage is included in a search engine's index in exchange for payment. According to Thurow (2003: 17), the advantages of this model include that web designers will have guarantees that their websites will not be dropped from a search engine index for a set period, and any changes to webpages will be reflected quickly as the website will be revisited more often. However, unlike the paid placement model, paid inclusion does not guarantee top rankings for websites. Most search engines have difficulty in indexing dynamic webpages, but with this model, all pages, dynamic or static will be indexed (Sullivan, 2004e).

2.3.6 Meta search engines

According to Green (2000: 127), when a searcher performs a search query on a meta search engine, the searcher is not searching the database of the meta search engine but is actually searching across several search engines and web directories. Barker (2005) has described how meta search engines work by explaining that once a searcher submits a keyword in a meta search engine search box, the search query is transmitted simultaneously to several search engines and their databases. This author also states that meta search engines do not own any database of webpages but rather send the search terms to the databases maintained by search engine companies (Barker, 2005). Garofalakis, Kappos and Makris (2002: 43-44) indicate that search results of a meta search engine are derived from combining the ranked results of popular search engines and then sorting in terms of relevance.

A study done by Oppenheim *et al*. (2000: 192) also indicates that the functionality of meta search engines depend on the performance of the participating search engines. The same authors are of the opinion that when using meta search engines, users do not have to re-enter their search query on a number of search engines as being the major advantage of using a meta search engine. Table 2.5 lists examples of two meta search engines.

TABLE 2.5: Example of meta search engines (**Source:** Barker, 2005).

Meta-Search Tool	What's Searched	Complex Search Ability	Results Display
Clusty clusty.com	Currently searches a number of **free,** search engines and directories, not Google or Yahoo!.	Accepts and "translates" complex searches with Boolean operators and field limiting.	Results accompanied with subject subdivisions based on words in search results, giving usually the major themes (Vivisimo Clustering Engine™). Click on these to search within results on each theme.
Dogpile www.dogpile.com	Searches Google, Yahoo!, LookSmart, AskJeeves/ Teoma, Google ADS, MSN search. Sites that have purchased ranking and inclusion are blended in. Watch for **Sponsored by...** links below search results.	Accepts Boolean logic, especially in advanced search modes.	Dogpile allows you to see each search engine's results separately in a **useful list for comparison**. Click the search engine icons by "Best of Breed."

According to Zhang and Cheung (2003: 433), it is almost impossible for a single search engine to index the entire web. The authors state that using multiple search engines can retrieve a broader scope of information, as in the case of meta search engines. Meta search engines enable users to save time and effort by using a single interface to conduct multiple searches (Zhang & Cheung, 2003: 433). The same authors have indicated the following as advantages of using meta search engines:

- The ability to access a group of search engines simultaneously.
- The capability to removing duplicate records from different search engines.
- The capability of ranking results against different criteria.

The following have been classified as disadvantages of using meta search engines (Zhang & Cheung, 2003: 434):

- Some meta search engines do not support advanced searching techniques.
- Some meta search engines do not conduct exhaustive searches.

TABLE 2.6: Shared visits to search engines and directories (**Source:** Sullivan, 2005).

Rank	Name	Domain	May	June	July
1	Google	www.google.com	38.3%	39.0%	39.4%
2	Yahoo!	Search.yahoo.com	18.4%	18.3%	18.2%
3	MSN Search	Seach.msn.com	15.6%	15.5%	15.4%
4	Google images	Images.google.com	4.5%	4.2%	4.0%
5	AskJeeves	www.askjeeves.com	2.4%	2.0%	2.0%
6	Yahoo! Images	Images.search.yahoo.com	1.6%	1.5%	1.5%
7	AOL Search	www.aolsearch.com	0.6%	0.8%	1.0%
8	My Web Search	www.mywebsearch.com	1.1%	0.9%	0.9%
9	Dogpile	www.dogpile	0.8%	0.8	0.8%
10	My Search	www.mysearch.com	0.7%	0.7%	0.7%

Table 2.6 provides a list of shared visits made to both search engines and directories in the US, over the period of May to July 2005 (Sullivan, 2005). Despite the fact that search engines perform the same tasks, they often produce different results. According to Machill *et al.* (2003: 52), the selection and listing criteria of search engines differ and are often hidden from a user. This is mostly due to the fact that search engines all have different methods of measuring relevancy of a search query that a user enters (Cooper, 2000: 13). According to (Cooper, 2000: 13-14), search results from different search engines can vary in the following ways:

- Speed of response.
- Total number of hits.
- Number of relevant hits.
- Position of relevant hits.
- Presentation of the hits.

The size of the web is increasing exponentially, resulting in the retrieval process of webpages becoming more and more complex. This in turn has resulted in search engines increasing their retrieval methods to keep up with the increase in size of the web (Moxley *et al.*, 2004: 61). The amount of information available on the web keeps on growing. According to Sullivan (2004b), Google currently has an index of over 8 billion webpages, while Zhang and Dimitroff (2004: 665) state that over 1 million new websites are being added annually. Search engines provide access to this complex information resource.

Before registering pages with search engines, there are a number of design issues that need to be addressed. Some of these design issues include flashing text, poor quality images as well as unnecessary moving images. These designs are likely to be passed off as spam by some search engines and need to be addressed depending on the search engine that the web designer is planning to submit to (Thelwall, 2000b: 152).

Search engines need to be programmed to determine how to rank websites. There are different criteria that search engines need to apply to decide which website to index (Thelwall, 2000a: 151). Research done by Courtois and Berry (1999) indicate that the basic principle of relevancy searching is that results are sorted or ranked according to certain criteria. These criteria represent a set of rules that search engines follow to determine which websites to index. It is also known as the ranking algorithm and varies among search engines.

The database size of all search engines varies significantly. According to Poulter (1997: 138), database size of a search engine can be measured in the following ways:

- The number of pages retrieved.
- The number of unique URL's.
- The number of total URL's.

2.3.7 Search engine results

Many search engines provide their listings from a variety of sources. The following table reflects where each search engine sources its main results (Sullivan, 2004d).

TABLE 2.7: Sources of search engine results (**Source:** Sullivan, 2004d).

Search Engine	Type Of Main Results	Provider Of Main Results	Paid Results	Directory Results
AllTheWeb	Crawler	Yahoo!	Overture	None
AltaVista	Crawler	Yahoo!	Overture	Open Directory
AOL Search	Crawler	Google	Google	Open Directory
AskJeeves	Crawler	Teoma	Google	None
Gigablast	Crawler	Gigablast	None	None
Google	Crawler	Google	Google	Open Directory
MSN Search	Crawler	Yahoo!	Overture	None
Netscape	Crawler	Google	Google	Open Directory
Teoma	Crawler	Teoma	Google	None
Yahoo!	Crawler	Yahoo!	Overture	Yahoo!

Research done by Nielsen (2004a) indicates that traffic from a search engine depends on the following factors:

- The search engine's raw traffic.
- The interests of the users using a particular search engine.
- The reputation of the search engine.

According to Goh and Ang (2003: 87), the search results of search engines are sorted according to particular algorithms. The same authors also state that the search engine ranking algorithms vary from search engine to search engine and are more often than not based on the following document characteristics:

- Number and frequency of matching terms.

- Location of terms within the document.
- Link structure.

2.3.8 Search engine optimization

SEO aims at improving the ranking of a website on search engine result lists (Wikipedia.com, 2005b). Commercialisation of the Internet has resulted in website designers adopting strategies to direct as much traffic to their website as possible. The aim is to attract as many users as possible to their site via their search engine rankings. This has culminated in many web designers implementing search engine optimizers in an attempt to heighten the retrievability of their websites by search engines (Machill *et al.,* 2003: 53).

Designers of webpages often attempt to influence the results of any web engine search (Oppenheim *et al.,* 2000: 194). Whenever a searcher enters a search query, search engine algorithms attempt to return the most relevant results. Search engines sometimes return millions of results, and if a website is not well optimized, chances of that website being part of the search results are minimal. Research done by one author indicates that search engine companies want to include as many websites on the Internet in their index as possible. This author has also indicated that despite search engines wanting to index as many websites as possible, the search results need to be as accurate and relevant as possible, and to contain the best quality websites. The retrieval of relevant results by search engines is critical for search engines because it makes them more popular among searchers. Being more popular among search engines is likely to increase advertising revenue, which is the primary source of profit for search engines (Anon, 2002).

According to Zhang and Dimitroff (2004: 666) search engine optimization aims at achieving easy access to the search engines for webpages, high visibility in search engine results and an improvement of the chances that webpages are retrieved. The authors are further of the opinion that SEO has become a very complex and sophisticated practice that requires constant research, practice and re-evaluation to be effective, while Google (2005d) argues that certain dubious SEO practitioners have participated in unethical marketing practices that have the sole purpose of influencing search engine results.

Fetterly *et al.* (2004) have pointed out that there are some SEO firms that guarantee high placements for websites in search engine result lists. The websites are loaded with pages that contain irrelevant keywords in an attempt to appear relevant to search engines. This is classified as spam. Some SEO practices have in the past raised some ethical issues. According to Weideman (2004: 910), there are two main ethical questions that arise regarding some SEO practices:

- A given webpage is excluded from search engine rankings while it deserves to be there due to high quality or relevance of content.
- A given webpage is included in rankings resulting from payment by the owner, regardless of the quality or relevance of its content.

Machill *et al*. (2003: 57-58) have categorised the ways of manipulating search engine results. The categorization reads as follows:

- External manipulation of search engines.

  In this category, website designers exert influence on search engine results without the involvement of search engine operators. The website designers add as many search terms as possible in the URL. This involves selecting a corresponding long domain name, incorporating frequently used search terms in the URL and the directory structure of the web server. Meta-tags are often the subject of searches, although they often do not accurately characterize the page contents. Website designers usually take heed of this and enter keywords in the meta-tags. Other ways of manipulation that fall within the ambit of this category include the creation of doorway pages. Doorways play the primary role of misleading search engines into giving undeserving websites top rankings. Another practice, referred to as cloaking, also falls in this category. It involves a process where a different webpage is passed to the user than that which is passed to a search engine crawler. Another form of manipulation by website designers involves the design of a network of webpages which have all been optimised for a certain search term. This is an attempt to influence the ranking which search engines that use link popularity as a measure of relevancy, allocated to a webpage.

- Internal manipulation of search engines.

  When attempting to internally manipulate search engine results, website providers turn to advertising practices deployed by search engines. Paid placement, paid inclusions and paid submissions fall into this category (Machill *et al.*, 2003: 58).

2.3.9 Challenges faced by search engines

Search engines have simplified ways of finding information on the Internet, unlike before where users had to know the exact URL of what they were looking for. Despite the benefits these tools have presented to Internet users (including free information retrieval), search engines have also received a lot of criticism. Some of this criticism is directed at the time it takes for search engines to respond to queries, the tendency to retrieve duplicate records as well as the amount of irrelevant information retrieved (Oppenheim *et al.*, 2000: 190-191). This argument was supported by Machill *et.al.* (2003: 52) who stated that the performance of search engines in general is disappointing. Research conducted by these authors indicate that one of the reasons that search engines are receiving so much criticism is the fact that it is unclear how they derive their results. According to Machill *et al.* (2003: 52):

> “Search engines lack any transparency to clarity how search results are found and how they are connected to the search term…”

However, Google (2005d) argues that the reason for hiding the algorithm is to maintain integrity in the search engine results, while Thelwall and Vaughan (2004: 25) state that it is not possible to be exact about search engine algorithms because only the broadest details of the algorithms are usually supplied by search engines. Another author has stated the fact that some search engines do not order results by relevance, but rather by how much money has been paid by the particular website (see Paragraph 2.3.5.1) (Wikipedia.com, 2005a). However, Drott (2002: 209) has supported search engines by stating that the enormous amount of information available on the web impacts adversely on new advances in automated web searching and algorithmic indexing. As indicated by Kline (2002: 253) managing the results of information retrieval is a problem. Fetterly *et al*. (2004) have stated that this is due to the congestion of search engines by websites that contain spam and are of no benefit to users. The number of searching transactions done per day on certain search engines is listed in Table 2.8.

TABLE 2.8: Daily searching transactions (**Source:** Sullivan, 2003c).

Service	Searches Per Day	As Of
Google	250 million	February 2003
Overture	167 million	February 2003
Inktomi	80 million	February 2003
LookSmart	45 million	February 2003
FindWhat	33 million	January 2003
AskJeeves	20 million	February 2003
AltaVista	18 million	February 2003
FAST	12 million	February 2003

2.4 Search engine spam

2.4.1 Introduction

According to Wilkinson (2004), search engine spam can be defined as anything that constitutes unethical practice within SEO, and this includes manipulating search engine spiders and redirecting users to inappropriate content. Anon (2002) has stated that the function of SEO firms is to achieve high search engine rankings for client websites. This author further argues that these SEO firms and web designers sometimes use dishonest tricks or unethical methodologies to improve the rankings of usually low quality websites in the search engine results. The artificial boosting of website rankings reduces the quality of search results resulting in the quality of these search results becoming questionable (Anon, 2002). However, Sullivan (2001b) argues that there is no set definition for what is considered spam. This author states that search engines are independent entities, and each one controls what it considers spam. Sullivan cites an example of cloaking, which is considered as spam by Google, but in contrast both AltaVista and Inktomi allow it in certain circumstances. Thurow (2003: 218) regards search engine spamming as a waste of time. This author states that the amount of money and time spent on using techniques that are considered as spam can be better spend optimizing websites using the correct techniques that will please both the user and search engines.

There are numerous opportunities of manipulating search engine results by web designers, search engine optimisation companies as well as search engines. Machill *et al.* (2003: 54) indicate that the use of questionable methods can be take place as follows:

- During the indexing of websites.
- In the actual search and ranking mechanisms.
- During efforts to optimise websites.
- Displaying preference of clients websites by both search engine operators and search engine optimisation companies.

The possibilities of manipulating search engine results are graphically depicted in Figure 2.4.

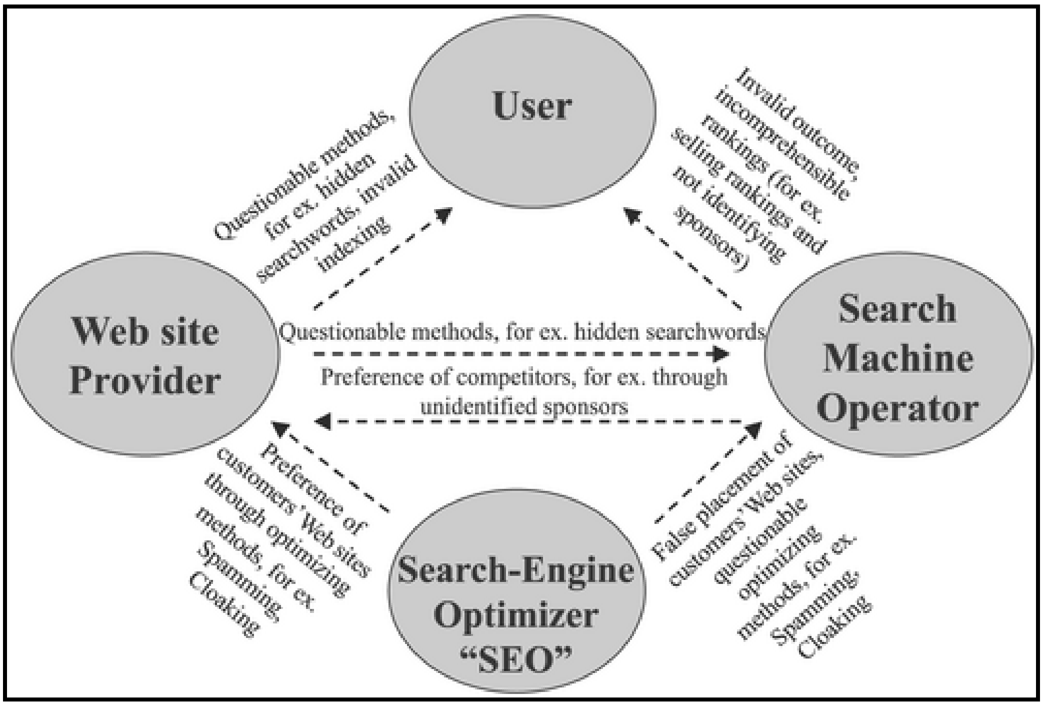

User
Web site Provider
Search Machine Operator
Search-Engine Optimizer "SEO"
Questionable methods, for ex. hidden searchwords, invalid indexing
Invalid outcome, incomprehensible rankings (for ex. selling rankings and not identifying sponsors)
Questionable methods, for ex. hidden searchwords
Preference of competitors, for ex. through unidentified sponsors
Preference of customers' Web sites through optimizing methods, for ex. Spamming, Cloaking
False placement of customers' Web sites, questionable optimizing methods, for ex. Spamming, Cloaking

FIGURE 2.4: Possibilities of manipulating search engine results (**Source:** Machill *et al.,* 2003: 54).

2.4.2 Classification

There is currently uncertainty over what is considered as spam. Furthermore, not all search engines are equally strict about spam and techniques which are acceptable for one search engine may be considered as spam by another search engine (Anon, 2002). According to Thurow (2003: 218), a number of website designers spend a lot of time making use of spamming techniques to gain top search engine rankings. Search engines on the one hand have been attempting to develop and improve techniques to detect spam, while on the other hand web designers are developing new spam techniques (Henzinger, *et al.,* 2002). As part of the continuing battle against spammers, many search engine designers have reviewed their site's eligibility policies (Dahm, 2000).

Research conducted by different authors has identified the following practices as spam:

2.4.2.1 Cloaking

Rowlett (2003) has defined cloaking as a process where visitors to a website are shown a completely different page to what search engine spiders see. There has been a lot of controversy surrounding cloaking. According to Sullivan (2003a), very few issues have been as controversial as cloaking. This author argues that even though most search engines have guidelines against cloaking, some still allow it, while others go as far as practicing it themselves. Anon (2002) explains that with cloaking, when a searcher requests a page a well-designed webpage is delivered while plain keyword-stuffed webpages are delivered to search engine spiders. While Thurow (2003: 227) considers cloaking to be spam, Sullivan (2003a) argues that cloaking should not be considered as spam because it does nothing to satisfy a search engine's algorithm, and its just a way of ensuring that targeted content is delivered.

Research done by Sullivan (2003a) also indicates that web designers deploy cloaking because some search engines cannot read flash content, resulting in an extra page being built for the search engines to index. Another reason cited is that some search engines do not index dynamic pages, also resulting in web designers designing an extra static page for spiders to index.

Despite all the controversy regarding cloaking, Thurow (2003: 227) maintains that all major search engines regard cloaking as a form of spam. Research done by the same author states that search engines will only accept cloaking if it is delivered through a trusted feed program, raising the debate of how such a practice could be acceptable if it is considered to be spam by some search engines. According to Chambers and Weideman (2005), cloaking is more popular on websites containing intense multimedia content. Research done by Henzinger *et al*. (2002) points to the fact that one way of detecting cloaking is by crawling a website twice using an HTTP client that the cloaker believes is a search engine and from a client that the cloaker does not believe is a search engine.

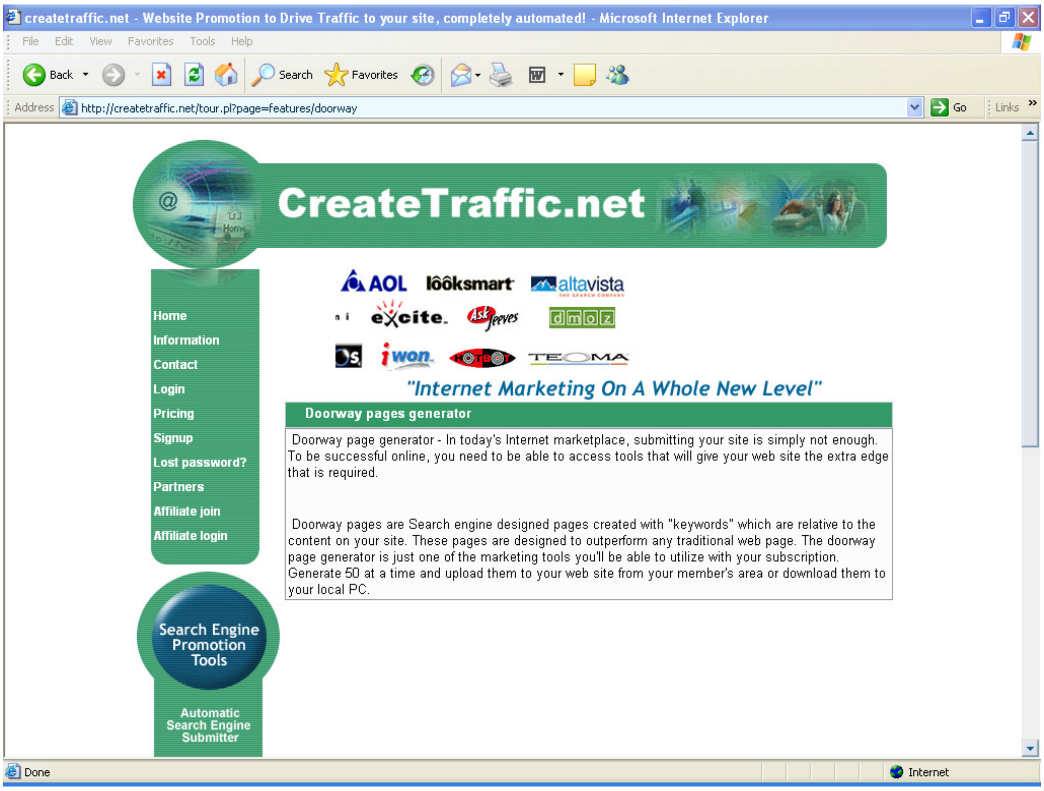
2.4.2.2 Doorway pages

Thurow (2003:227) argues that the primary reason why web designers create doorway pages is to obtain high rankings with search engines. A doorway page has been defined as webpages that have been designed to deceive search engines into ranking them higher for one or more particular keywords. Human visitors are redirected to a different, more human friendly webpage.

Thurow (2004a) states that:

> "Many people do not understand how doorway page companies work. They create thousands of pages for a single keyword or keyword phrase. All of these pages are fed to the search engines, polluting their indices with unnecessary information. They are not pretty, and they often contain so much gibberish they must be [hidden]. End users would not continue visiting a website if they viewed these pages".

Gikandi (1999) is of the opinion that doorway pages are not necessarily good or bad. This author argues that doorway pages are effective web promotion tools and goes on to state that they compensate search engine weakness and ultimately assist searchers find what they are searching for, while Thurow (2003: 227) deems doorway pages as filling search engine indices with webpages containing junk. Anon (2002) concurs by stating that doorway pages do not contain any valuable content and are just stuffed with keywords. Research conducted by Sullivan (2003a) indicates that doorway pages and cloaking often go hand in hand. This author differentiates between the two by stating that doorway pages merely attempt to satisfy a search engine's algorithm, while cloaking is a way of retrieving and delivering targeted content (Sullivan, 2003a). Nobles and O'Neil (2000: 166) advise against having too many doorway pages, causing more confusing on whether doorway pages can be viewed as spam or not. Thurow (2004b) has indicated that some doorway pages are very difficult to identify because they have graphic images and navigation schemes, and therefore look exactly like a normal webpage. A study conducted by Dunn (2004) found that the use of doorway pages was very popular among webmasters until 2000, when these pages became one of the most obvious forms of spam. See Figure 2.5 for an example of a doorway page generator site.


createtraffic.net - Website Promotion to Drive Traffic to your site, completely automated! - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://createtraffic.net/tour.pl?page=features/doorway
Go
Links
CreateTraffic.net
Home
Home
Information
Contact
Login
Pricing
Signup
Lost password?
Partners
Affiliate join
Affiliate login
AOL
lòòksmart
altavista
excite
Ask Jeeves
dmoz
iwon
HOTBOT
TEOMA
"Internet Marketing On A Whole New Level"
Search Engine Promotion Tools
Automatic Search Engine Submitter
Done
Internet

Doorway pages generator

Doorway page generator - In today's Internet marketplace, submitting your site is simply not enough. To be successful online, you need to be able to access tools that will give your web site the extra edge that is required.

Doorway pages are Search engine designed pages created with "keywords" which are relative to the content on your site. These pages are designed to outperform any traditional web page. The doorway page generator is just one of the marketing tools you'll be able to utilize with your subscription. Generate 50 at a time and upload them to your web site from your member's area or download them to your local PC.

FIGURE 2.5: Doorway page generator site (**Source:** CreateTraffic.net, 2001).

2.4.2.3 Invisible text

For websites to receive high rankings, keywords contained in search queries must be present in a webpage. Some web designers place keywords that are the same colour as the background of a webpage, on the webpage itself, resulting in the keywords not changing the site design. These keywords are intended to be visible to spiders, and not to users that visit the site (Anon, 2002; Thurow, 2003: 222). Other implementation methods of invisible text include hiding text behind layers and placing text at the bottom of oversized pages (Dunn, 2004). An example of how black text would be hidden in a black background is shown in Figure 2.6 (Collins, 2004):

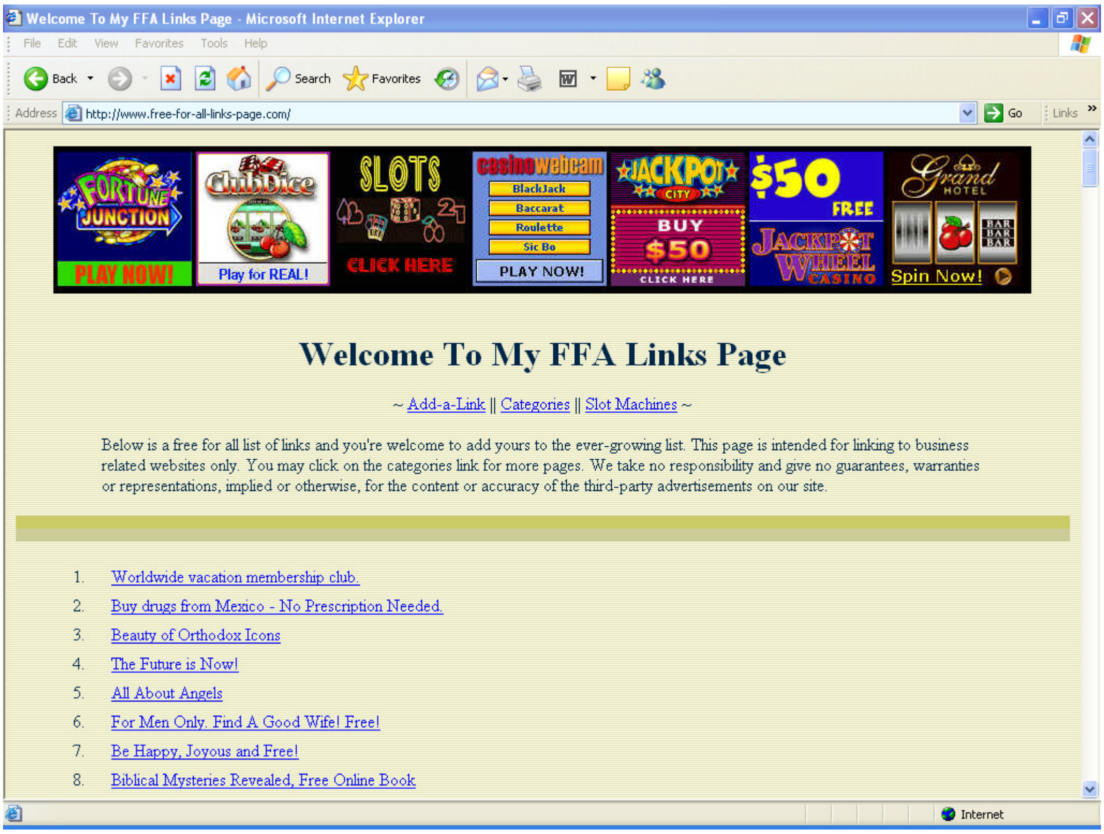
```
<body bgcolor="#000000">
<table width="14%" border="0" cellpadding="6" cellspacing="0" bgcolor="#FFFFFF">
 <tr>
   <td background="black.gif"><font color="#000000">invisible text</font></td>
 </tr>
</table>
<div id="Layer1" style="position:absolute; width:200px; height:115px; z-index:1; left: 5px; top:
8px; background-image: url(black.gif); layer-background-image: url(black.gif); border: 1px
none #000000;"></div>
</body>
```

FIGURE 2.6: Hidden text code (**Source:** Collins, 2004).

2.4.2.4 Artificial link farms

Search engines rely on link analysis to determine relevancy of websites to search engines. A study conducted by Sullivan (2002e) indicates that what websites link to, represents a major component of how that particular website is ranked. Sullivan has also stated that the best way to get listed in crawlers like Google is to build links to the website, as crawlers follow links and index new pages (Sullivan, 2004g). This has resulted in web designers attempting to artificially increase link popularity by creating multiple websites with the intention of linking the sites to one another, a practice commonly referred to as artificial link farms. However, Thurow (2003: 224) points out that search engines consider link farming to be spam, and websites that implement artificial link farms are likely to be penalized by search engines. Henzinger *et al*. (2002) have stated that the main purpose of artificial link farms is to manipulate systems that use the number of incoming links to a website to determine relevancy.

An example of artificial links is a Free-for-all link page. According to Rowlett (2003), these pages are not designed to benefit any specific area and should be avoided at all times. Figure 2.7 serves as a graphical depiction of a Free-for-all link page.



Welcome To My FFA Links Page

~ Add-a-Link || Categories || Slot Machines ~

Below is a free for all list of links and you're welcome to add yours to the ever-growing list. This page is intended for linking to business related websites only. You may click on the categories link for more pages. We take no responsibility and give no guarantees, warranties or representations, implied or otherwise, for the content or accuracy of the third-party advertisements on our site.

1. Worldwide vacation membership club.
2. Buy drugs from Mexico - No Prescription Needed.
3. Beauty of Orthodox Icons
4. The Future is Now!
5. All About Angels
6. For Men Only. Find A Good Wife! Free!
7. Be Happy, Joyous and Free!
8. Biblical Mysteries Revealed, Free Online Book

FIGURE 2.7: A home page of a Free-For-All links site (**Source:** The Endless Links Page Company, 2005).

2.4.2.5 Keyword stacking/stuffing

Thurow (2003: 221) defines keyword stacking as the repeated use of a keyword or keyword phrase to artificially boost a webpage's relevancy in search engines. This author further states that keyword stacking and keyword stuffing often have the same meaning and therefore the two terms will be used interchangeably in this research. According to Wilkinson (2004), even though keyword stacking may help a website's keyword relevancy attain a higher rank, it is likely that the high ranking will not last long and also likely that the website may get penalized by search engines. The same author states that keyword stuffing also refers to loading the page with content that extends so far down the page that it is unlikely anyone will continue to scroll down. This technique is likely to be used by designers who rely on splash pages for their index pages (Wilkinson, 2004).

The following is an example of keyword stacking:

'hair clips hair clips hair clips hair clips hair clips hair clips
hair clips
clips clips clips clips clips clips clips clips clips clips clips'

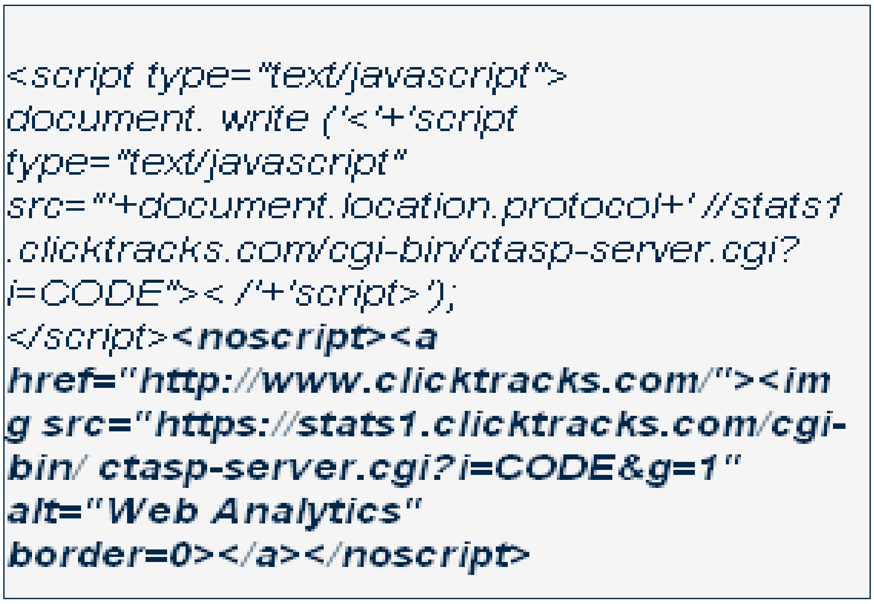
These keywords are then placed within any HTML tag. They can also be placed at the bottom of the webpage so that users cannot view them easily.

2.4.2.6 Hidden links

Some sites contain hidden links to other pages. Google, which is one the top search engines (Nielsen/Netratings, 2005), places emphasis on the number of inbound amount of links when determining relevancy (Wikipedia, 2005a). This has resulted in web designers creating links that are only visible to search engine spiders, and not to users, known as hidden links (Thurow, 2003: 223). According to Anon (2005a), hidden links are commonly used to increase the link count of webpages. Thurow (2003: 223) has identified the following as ways of hiding links:

- Using the same font colour for a hypertext link as regular text.
- Hiding hypertext links in punctuation marks.
- Hiding hyperlinks in transparent images.
- Hiding hyperlinks in invisible layers.

- Hiding many links inside a small graphic image (See Figure 2.8).



```
<script type="text/javascript">
document. write ('<'+'script
type="text/javascript"
src=""+document.location.protocol+' //stats1
.clicktracks.com/cgi-bin/ctasp-server.cgi?
i=CODE"></'+'script>');
</script><noscript><a
href="http://www.clicktracks.com/"><im
g src="https://stats1.clicktracks.com/cgi-
bin/ ctasp-server.cgi?i=CODE&g=1"
alt="Web Analytics"
border=0></a></noscript>
```

FIGURE 2.8: Hidden link code (**Source:** Anon, 2005a).

2.4.2.7 Page redirects

A page redirect involves the placing of HTML code that will redirect a site visitor from a page that is designed only for a search engine, to another page designed for searchers (Anon, 2002; Thurow, 2003: 225). According to Thurow (2003: 225), spammers create webpages that are optimized with certain keyword phrases. These pages are then submitted to search engines, but whenever a searcher clicks on this webpage they are automatically redirected to another page, known as a destination page.

Thurow (2003: 225) goes on to state that in an attempt to battle this type of spam, search engines do not index any pages with redirects except the HTTP 301 (permanent) redirect. In cases where websites that contain redirects are indexed, search engines list the destination page. Dunn (2004) also states that not all redirects are spam. According to Anon (2004a), page redirects are most commonly brought in to compliment doorway pages.

2.4.2.8 Duplicate Pages

This process involves spammers slightly modifying webpages and resubmitting them to search engines. This often results in the same webpages with different title tags appearing in search results (Anon, 2002; Thurow, 2003: 226). According to Dunn (2004), when duplicating content, webmasters create a website, and then duplicate webpages of that particular site are created. Each page is often optimized differently than the other in order to get different placements in search engines. Dunn (2004) is of the opinion that this practice decreases the bandwidth of search engines.

2.4.2.9 Domain Spam

Domain spam refers to the purchasing of several domain names and creating websites with identical content. This is done for the sole purpose of getting multiple listings in directories with the intention of achieving link popularity and more traffic. Web designers anticipate that the link popularity will improve search engine rankings (Thurow, 2003: 227).

2.4.2.10 Automatic Submission

This process involves automatically submitting websites to search tools, using software. This process is often done repeatedly, causing the search engines to be flooded with similar websites (Anon, 2004b; Dunn, 2004). Anon (2004b) also states that automatic submission is done by visiting an ‘auto-submit URL’ site, which then submits the URL to many search engines that are in their database. Dunn (2004) has pointed out that automatic submission floods the bandwidth of search engines as websites are often submitted repeatedly within the same search engine.

2.4.2.11 Meta-tag stuffing

Meta-tag stuffing refers to the practice of repeating keywords in the meta-tags and using keywords which are unrelated to the site’s content (Wikipedia.com, 2005c). Thurow, (2004c) has stated that search engines initially used meta-tags to determine relevancy of websites to search queries entered by users. Furthermore, that due to meta-tag stuffing, search engines do not place further emphasis on meta-tags. A study done by Wallace (2003) indicates that some unethical web designers place high traffic keywords that are not related to the webpage in meta-tags in order to generate more traffic. Alimohammadi (2003: 240) has stated that the biggest problem with keyword meta-tags is web designers repeating keywords that have no relevance to the website.

2.4.2.12 URL spam

According to Gyöngyi and Garcia-Molina (2005), some spammers create long ULR's that include sequences of spam terms, also known as URL spam. The same authors cite the following example of typical URL spam:

buy-canon-rebel-20d-lens-case.camerasx.com,
buy-nikon-d100-d70-lens-case.camerasx.com…

2.4.3 Penalties

Search engines have different penalties that they impose on websites that are suspected of search engine spamming. According to Anon (2002), search engines are continuously changing their algorithms in an attempt to prevent spammers from flooding search engine result pages with websites that contain irrelevant content.

According to Konia (2002: 311), if a website uses questionable SEO practices, the website can be penalised in one of the following ways:

- The page is red flagged for closer inspection by a human reviewer.
- The page's ranking is reduced.
- The offending page is dropped from the index.
- The entire site is banned from the engine.

Marckini (2000) states that search engine spam penalties can be as severe as having the domain name, IP address and all pages registered under the website's Internic handle banned. The same authors state that other penalties include the search engine checking domain registrations to prevent known spammers from registering new domains and getting back onto the index of that search engine. Other penalties imposed by search engines include refusing to index pages believed to contain spam, giving the sites lower rankings, or even banning the whole site (Anon, 2002). Google (2005c) states that websites that are suspected of cloaking can be permanently banned from the Google index.

Website authors want their websites to achieve top rankings in search engines in order to attract as many visitors as possible. However, it is common for websites to lose their rankings. Anon (2002) has highlighted the following as reasons for the loss of rankings in search engine results:

- Search engines keep changing and modifying their ranking algorithm, which can result in websites losing their ranking.
- The arrival of new websites entering a search engine index can affect a website's current ranking.
- In a case where a search engine changes its spam guidelines, a website designer that used SEO practices that were not previously regarded as spam can receive a lower ranking or even result in the entire website being banned from the search engine, should that particular practice is subsequently regarded as spam by the search engine.
- Changes in the server technology of a website may cause the search engine ranking of a website to fall.

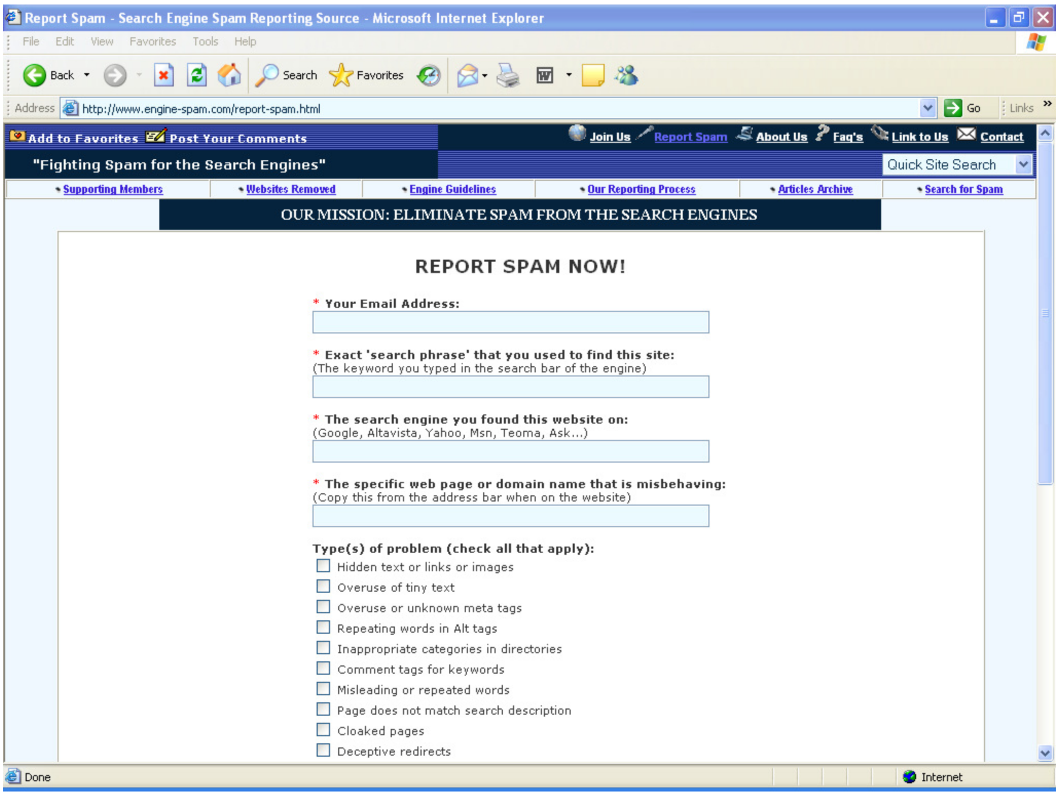
According to (Kirkpatrick, 2002; Wikipedia.com, 2005b), the following have been identified as acceptable SEO practices:

- Using a robots.txt file to grant permission to spiders to access files in the site.

- Using a short and relevant page title to name pages.
- Using a reasonably sized description meta-tag without excessive use of keywords.
- Developing links through accepted methods and not via hidden links or artificial link farms (refer to Paragraph 2.4.2.4).
- Having a site index to ensure that the entire site is indexed.

2.4.4 Spam detection

According to Fetterly *et al.* (2004), search engines should aim to remove spam pages so as to ensure that the search experiences of users are improved. These authors also state that the identification of spam pages is valuable as it enables search engines to develop more sophisticated algorithms to detect spam. Figure 2.9 depicts the home page of a site that promotes the reporting of spam by web users.


Report Spam - Search Engine Spam Reporting Source - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.engine-spam.com/report-spam.html
Go
Links
Add to Favorites
Post Your Comments
Join Us
Report Spam
About Us
Faq's
Link to Us
Contact
"Fighting Spam for the Search Engines"
Quick Site Search
Supporting Members
Websites Removed
Engine Guidelines
Our Reporting Process
Articles Archive
Search for Spam
OUR MISSION: ELIMINATE SPAM FROM THE SEARCH ENGINES
Done
Internet

REPORT SPAM NOW!

* Your Email Address:

* Exact 'search phrase' that you used to find this site:
(The keyword you typed in the search bar of the engine)

* The search engine you found this website on:
(Google, Altavista, Yahoo, Msn, Teoma, Ask...)

* The specific web page or domain name that is misbehaving:
(Copy this from the address bar when on the website)

Type(s) of problem (check all that apply):

□ Hidden text or links or images
□ Overuse of tiny text
□ Overuse or unknown meta tags
□ Repeating words in Alt tags
□ Inappropriate categories in directories
□ Comment tags for keywords
□ Misleading or repeated words
□ Page does not match search description
□ Cloaked pages
□ Deceptive redirects

FIGURE 2.9: Spam reporting site (**Source:** Anon, 2004c).

2.5 Search engine exclusion policies

Search engines have policies that govern the use of their services. Search engines may penalize pages or exclude them from their index, if search engine spamming is detected. Search engines detect common spamming methods in a variety of ways, including following intervention by owners resulting from complaints from users (Sullivan, 2003b).

Search engine exclusion policies can assist web designers in terms of which SEO tactics to avoid when optimizing their website. By following the guidelines of search engines, web designers can avoid having their websites banned. Web designers will refrain from practices that are seen as tricking search engines into giving undeserving sites top rankings.

According to Konia (2002: 312-316) the following tactics need to be avoided when optimizing websites:

- Serve duplicate or near-duplicate pages.
- Misrepresent a site by listing keywords that have no relevance to the site.
- Hide keywords.
- Cloak.
- Stuff too many keywords into a page.
- Redirect.
- Page jack.
- Build bad doorway pages.
- Fail to cross link.
- Build junk.

2.5.1 Google

Google has been rated the world's largest search engine (Joint, 2005; Sullivan, 2004c). Google is a crawler based search engine that uses a system called PageRank to rank webpages. Although it is not the only factor that Google uses to determine ranking, it is an important one. The PageRank of a webpage is represented by a numeric value between zero and ten, and is calculated after analyzing the inbound links of a site (Wikipedia, 2005a).

Google has an index of over 8 billion webpages. It is an automated search engine which relies on spiders to crawl the web on a monthly basis and find pages that match the criteria of Google's submission policy (Google, 2005b). Research done by Sherman (2002) states that Google was the first major search engine to index non-HTML web content when it started listing PDF files. The same author indicates that Google's index contains more than 22 million PDF files (Sherman, 2002).

Google was the first search engine to introduce a feature which searches exclusively for scholarly information (Jasco, 2005: 208). With regard to SEO, Google's basic principle for a quality website is for web designers to design webpages for users and not for search engines, as is the case with spammers (Google, 2005d). The aim, according to Google (2005d) is:

> "Trying to deceive (spam) our web crawler by means of hidden text, deceptive cloaking or doorway pages compromises the quality of our results and degrades the search experience for everyone. We think that's a bad thing".

To index a website can take between six and eight weeks. Google states that to determine relevancy, more than 100 factors are considered (Google 2005c).

TABLE 2.9: A comparison of Google and Yahoo! search transactions (**Source:** Sullivan, 2004b).

Country	Google	Yahoo!
Germany	80.5%	5.6%
UK	65.6%	10.8%
China	72.6%	12.7%

Table 2.9 indicates that Google is the biggest search engine both in the US and outside (Sullivan, 2004b). The Google home page is depicted in Figure 2.10.

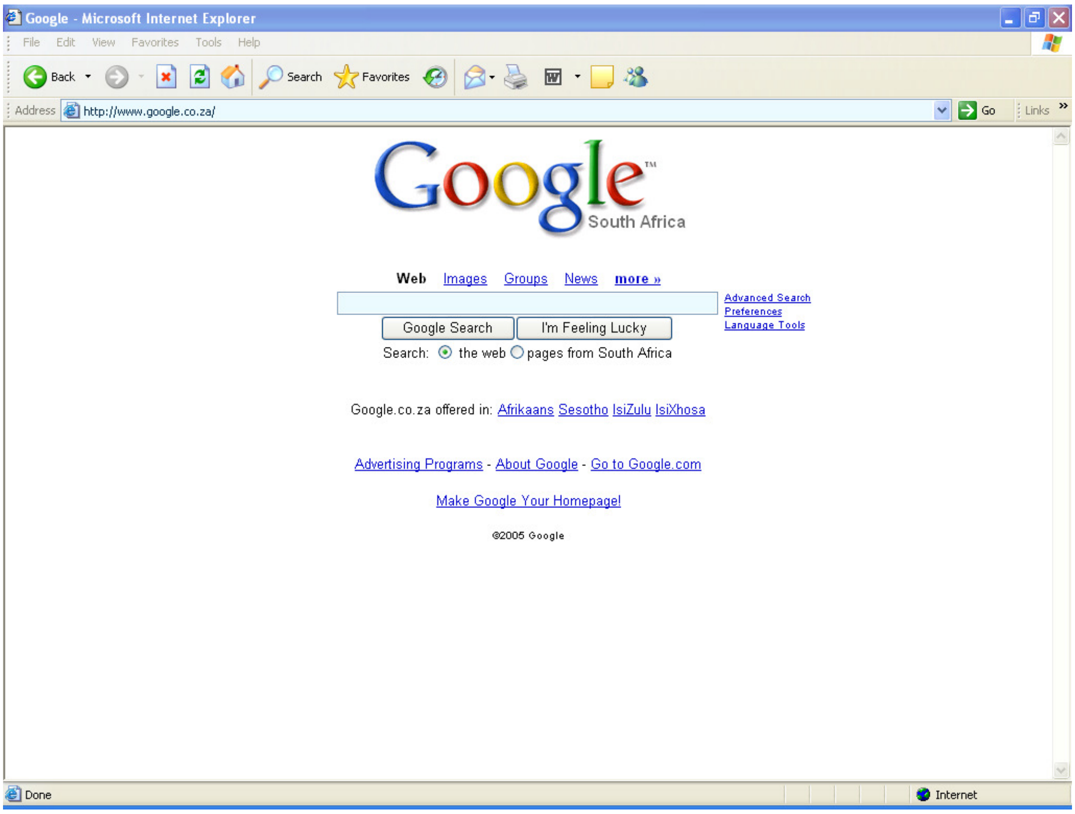
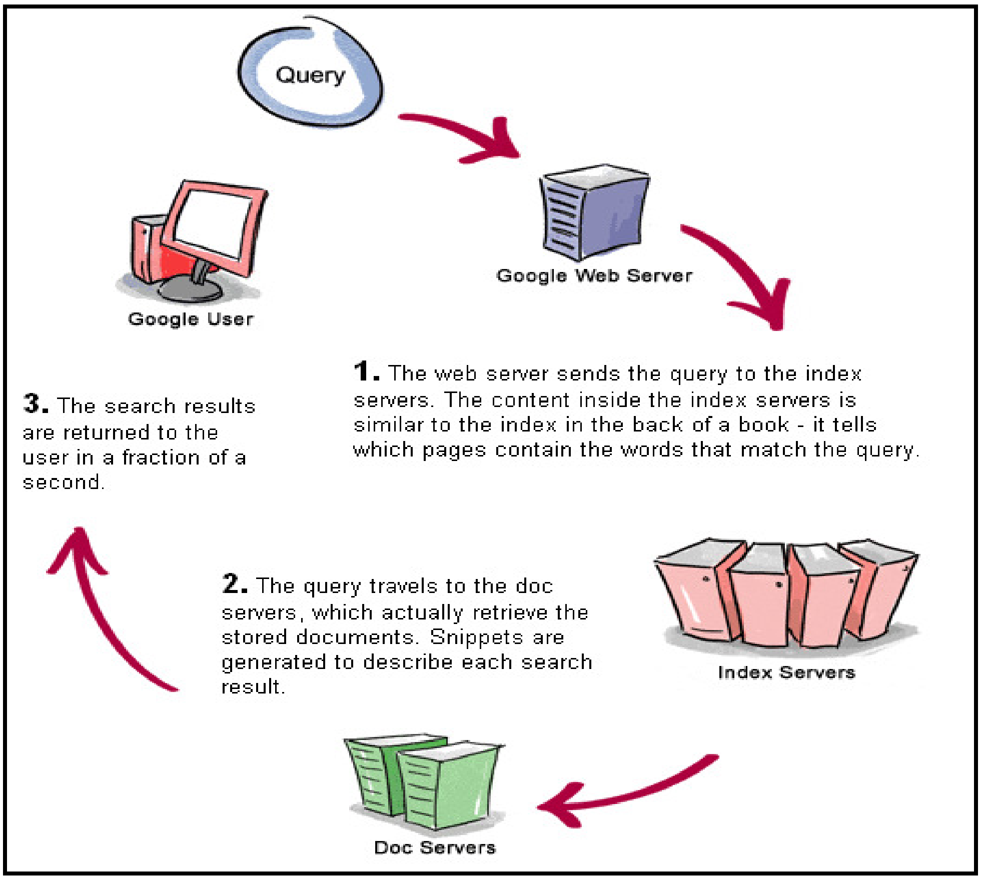

Google - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.google.co.za/
Go
Links
Google™
South Africa
Web Images Groups News more »
Advanced Search
Preferences
Language Tools
Google Search
I'm Feeling Lucky
Search: ● the web ○ pages from South Africa
Google.co.za offered in: Afrikaans Sesotho IsiZulu IsiXhosa
Advertising Programs - About Google - Go to Google.com
Make Google Your Homepage!
©2005 Google
Done
Internet

FIGURE 2.10: Google home page (**Source:** Google, 2005a)

There are a plethora of reasons Google will not index some webpages. Some of these reasons include dynamically generated webpages and the use of frames on a webpage. Google (2005d) argues that frames often cause problems with search engine crawlers. One problem identified is that frames do not fit the conceptual model of the web. Other reasons why some pages may not be indexed include the fact that the sites may be unreachable and subject to technical hitches (2005d). According to Sullivan (2002a), some website designers attempt to influence Google search engine results by controlling where they link to, and which websites in turn link to theirs.

Information about Google's search engine policy indicates that Google strongly discourages the use of spam. This is supported by the following general recommendations when submitting to Google:

- Hidden text or hidden links should be avoided at all times.
- Cloaking and certain redirects should not be deployed.
- Automated queries should not be sent to Google.
- Multiple pages, subdomains or domains that contain duplicate content should not be created.
- Doorway pages should be avoided.



Query
Google Web Server
Google User
1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book - it tells which pages contain the words that match the query.
3. The search results are returned to the user in a fraction of a second.
2. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.
Index Servers
Doc Servers

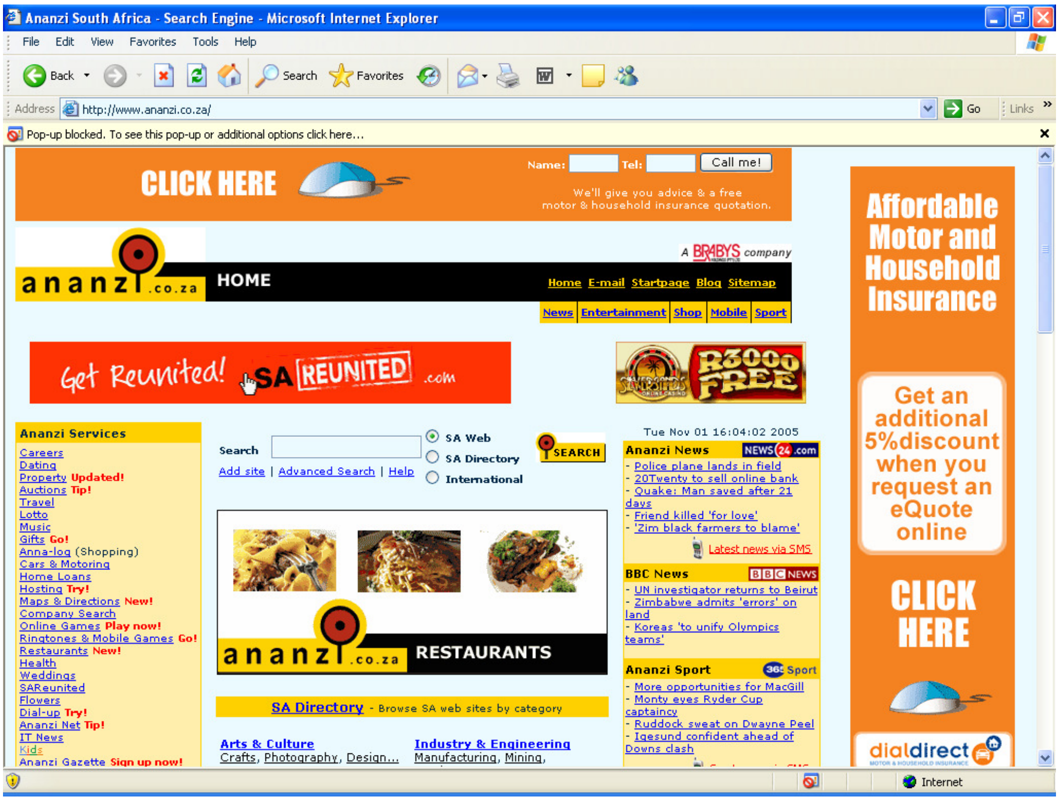
FIGURE 2.11: Steps taken to execute a query (**Source:** Google, 2004).

Query execution by Google is shown in Figure 2.11.

2.5.2 Ananzi

Ananzi has grown to become South Africa's most popular search engine, with over 11 million page impressions per month (Ananzi, 2005c). Adding a website to Ananzi is free, however only websites with Southern African based content are included in their index.

Ananzi currently indexes over 300 000 webpages within South Africa, and the number is growing exponentially. Alongside the search engine, Ananzi's SA site Directory is a hand picked category-based list of the best sites Southern Africa has to offer (Ananzi, 2005c).


Ananzi South Africa - Search Engine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
CLICK HERE
Call me!
ananzi.co.za
A BRABYS company
Get Reunited! SA REUNITED .com
SILVERSANDS ONLINE CASINO R3000 FREE
SEARCH
ananzi.co.za RESTAURANTS
Affordable Motor and Household Insurance
Get an additional 5%discount when you request an eQuote online
CLICK HERE
dialdirect MOTOR & HOUSEHOLD INSURANCE
Internet

http://www.ananzi.co.za/

Pop-up blocked. To see this pop-up or additional options click here...

Name: Tel:

We'll give you advice & a free motor & household insurance quotation.

HOME

Home E-mail Startpage Blog Sitemap

News | Entertainment | Shop | Mobile | Sport

Tue Nov 01 16:04:02 2005

Ananzi Services

Careers
Dating
Property Updated!
Auctions Tip!
Travel
Lotto
Music
Gifts Go!
Anna-log (Shopping)
Cars & Motoring
Home Loans
Hosting Try!
Maps & Directions New!
Company Search
Online Games Play now!
Ringtones & Mobile Games Go!
Restaurants New!
Health
Weddings
SAReunited
Flowers
Dial-up Try!
Ananzi Net Tip!
IT News
Kids
Ananzi Gazette Sign up now!

Search
● SA Web
○ SA Directory
○ International

Add site | Advanced Search | Help

SA Directory - Browse SA web sites by category

Arts & Culture
Crafts, Photography, Design...

Industry & Engineering
Manufacturing, Mining,

Ananzi News NEWS24.com

- Police plane lands in field
- 20Twenty to sell online bank
- Quake: Man saved after 21 days
- Friend killed 'for love'
- 'Zim black farmers to blame'

Latest news via SMS

BBC News BBC NEWS

- UN investigator returns to Beirut
- Zimbabwe admits 'errors' on land
- Koreas 'to unify Olympics teams'

Ananzi Sport 365 Sport

- More opportunities for MacGill
- Monty eyes Ryder Cup captaincy
- Ruddock sweat on Dwayne Peel
- Igesund confident ahead of Downs clash

FIGURE 2.12: Ananzi home page (**Source:** Ananzi, 2005b).

Ananzi consists of two sections, namely the SA site directory and the Search Engine (SA web). In order to get a site onto the Ananzi search engine, it has to be submitted to the SA Site Directory from where it will be indexed and then added to the search engine should it meet Ananzi's acceptance criteria. Figure 2.13 provides a brief graphical display of how Ananzi operates.

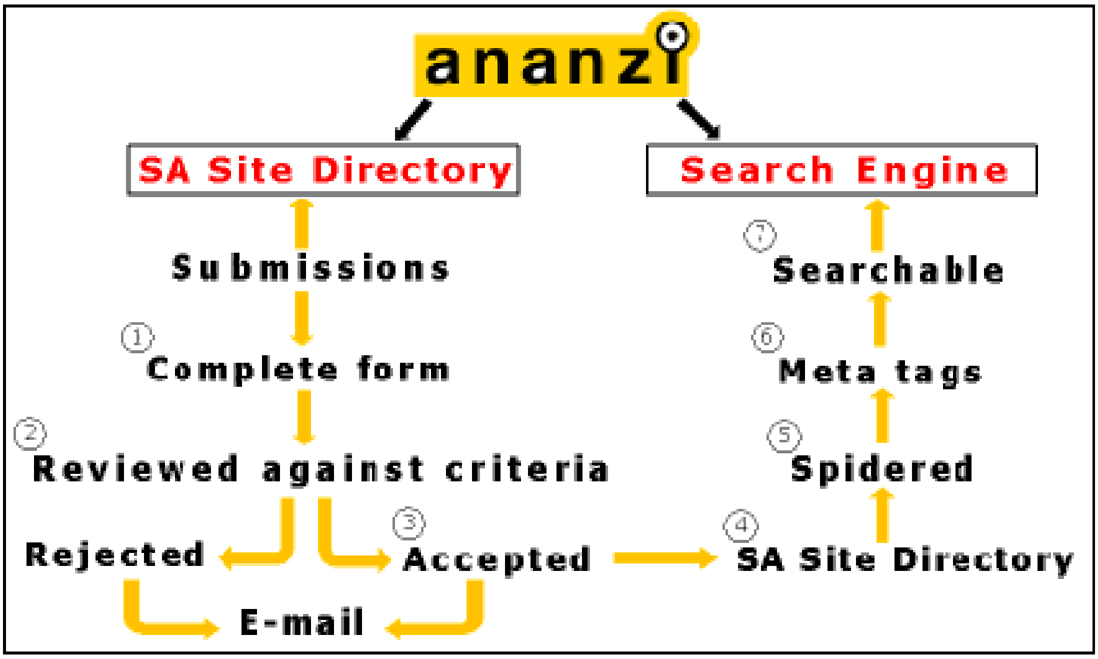


FIGURE 2.13: Ananzi operational diagram (**Source:** Ananzi, 2005a).

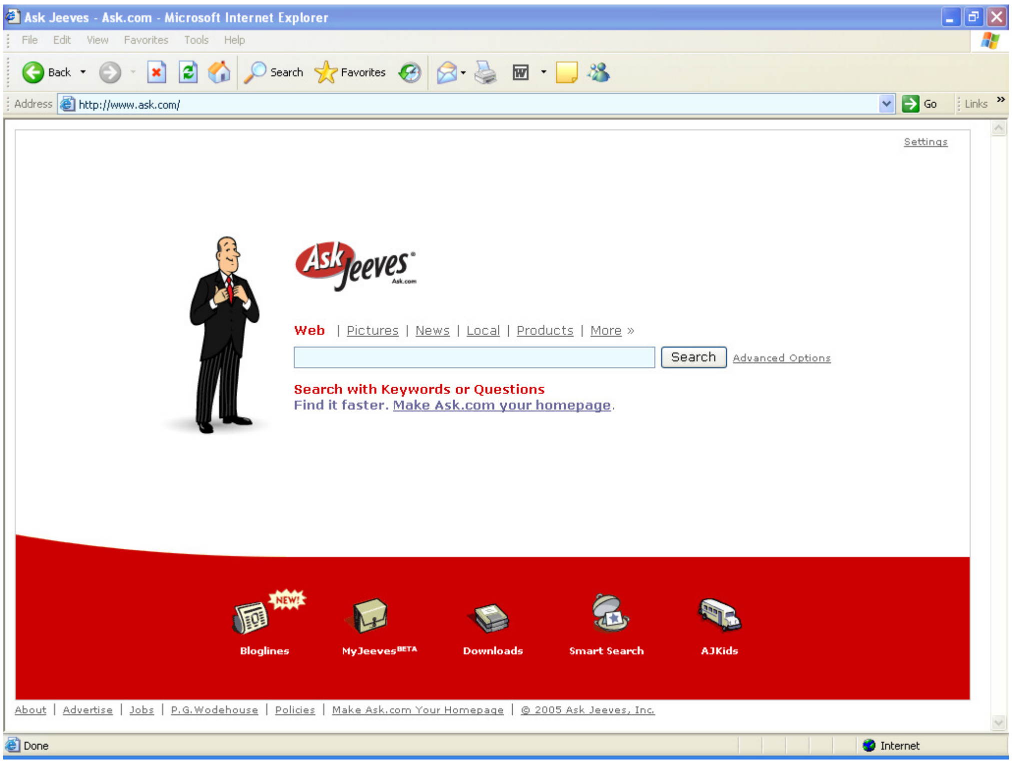
The Ananzi search engine regulates itself by imposing the following rules:

- Sites without meta-tags will be rejected.
- Ananzi will reject sites whose content is essentially repetitive in nature.
- Ananzi does not accept doorway pages.
- Words and meta-tags used to describe sites must accurately represent their content.
- Sites which have no bearing on Southern Africa will be rejected.
- No submission will be accepted if the full entry is in capital letters.
- Excessive use of punctuation marks and symbols in the title to boost site listings will not be allowed.
- Duplicate entries will only be allowed under special circumstances.

Ananzi's ranking algorithm is dependant on a website's meta-tags as well as other factors, while like Google, frames on a webpage can lower a website's ranking (Ananzi, 2005a; Google, 2005b).

2.5.3 AskJeeves

AskJeeves was founded in 1996 and has grown to become one of the top search engines (Sullivan, 2004c). AskJeeves operates a range of websites, portals and downloadable applications. Additionally, AskJeeves owns the differentiated search technology Teoma, as well as natural language processing, portal and ad-serving technologies (AskJeeves, 2005).


Ask Jeeves - Ask.com - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Go
Links
Settings
Ask Jeeves
Web | Pictures | News | Local | Products | More »
Search
Advanced Options
NEW!
Bloglines
MyJeeves BETA
Downloads
Smart Search
AJKids
About | Advertise | Jobs | P.G.Wodehouse | Policies | Make Ask.com Your Homepage | © 2005 Ask Jeeves, Inc.
Done
Internet

Address http://www.ask.com/

Search with Keywords or Questions
Find it faster. Make Ask.com your homepage.

FIGURE 2.14: AskJeeves home page (**Source:** AskJeeves, 2005).

The Teoma crawler serves as AskJeeves' Web-indexing robot. Teoma is different from any other search technology due to the fact that it analyzes the web as it actually exists - in subject-specific communities. To do this, a comprehensive and high-quality index is created. Web crawling is an essential tool in this process, and it ensures the most up-to-date search results (AskJeeves, 2005).

Examples of techniques which are listed under the AskJeeves (2005) site that are considered to be spamming include, but are not limited to, the following:

- Webpages containing deceptive text.
- Webpages with intentionally misleading links.
- Webpages in Groups with deceptive self linking referencing patterns.
- Webpages with off-topic or excessive keywords.
- Webpages with duplicate content.
- Webpages that show different content than the spidered pages.
- Fabricated pages designed to lead users to other webpages.
- Metadata that does not accurately describe the content of a webpage.
- Webpages that abuse affiliate or referral programs.

2.5.4 Yahoo!

Yahoo!, which was launched in 1994, originally served as a directory. In 2002, Yahoo! changed to include crawler based results. There are numerous factors that can affect the listing of a website under Yahoo! and the Yahoo! directory (Sullivan, 2004c).

According to Sherman (2002), the Yahoo! Search index captures the full text of webpages, up to a 500kb limit. A study done by Sullivan (2004g), indicates that Yahoo! has two options of submitting a website namely the standard submission which is free, and the Yahoo! express option, which involves a submission fee (Sullivan, 2004h).

Yahoo! search ranks results according to their relevance to a particular query by analyzing the webpage text, title and description accuracy as well as its source, associated links, and other unique document characteristics (Yahoo!, 2005b).

Yahoo! (2005b) has listed the following directives as the correct ways of ensuring that a website is indexed by search engines.

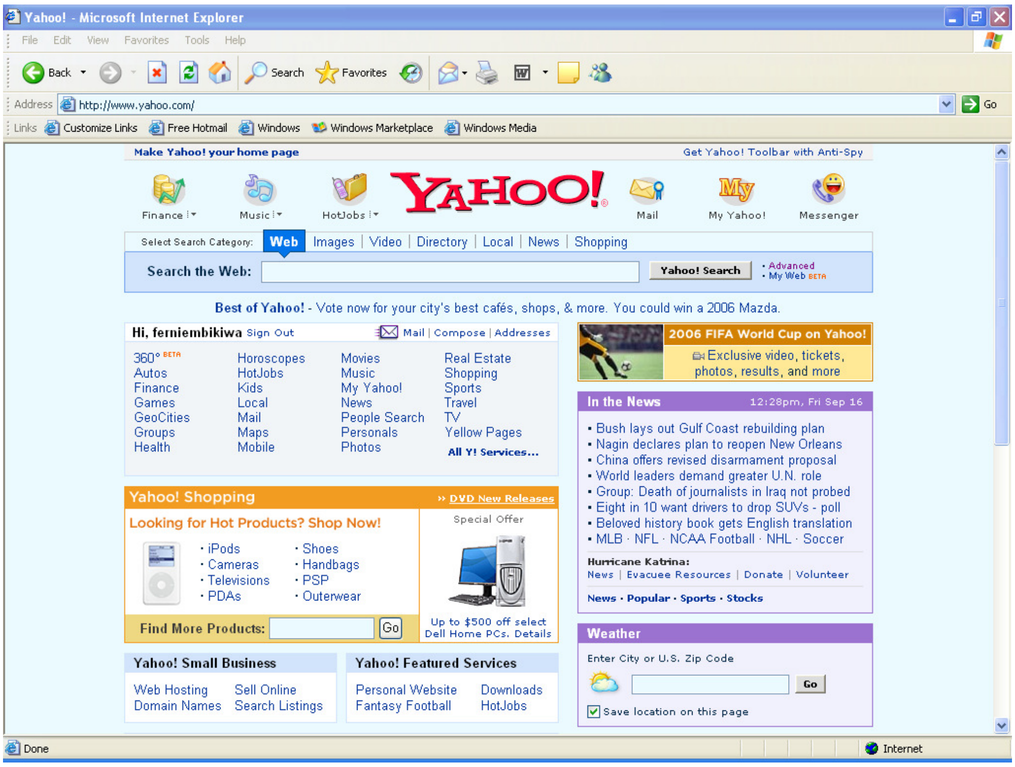
- Keywords that users are likely to search on should be thought of carefully. These keywords should closely represent the content of the website.
- Title names should match the contents of the website because users are more likely to choose link on URL's that closely matches their search.
- The 'description' meta-tag should be used with the description written accurately and cautiously. The document title and description should attract the interest of the user as well as fit the content on your site.
- The 'keyword' meta-tag should be used to list relevant keywords for the document.
- Avoid keeping relevant text and links in graphics or image maps because is likely that some search engine crawlers cannot follow links to your site's other pages. If possible, use an HTML site map to increase the chances of your site getting indexed.
- Use ALT text for graphics as it helps improve the text content of your page for search purposes.
- Build rich linkages between related pages with other webmasters. Using link farms violates Yahoo!'s Site Guidelines and will not improve a webpage ranking.

As in the case of most search engines, Yahoo! has policies that regulate indexing of websites. Yahoo! considers the following as webpages that are unlikely to be indexed:

- Pages that are intended to interfere with the accuracy of search results.
- Redirects.
- Repetitive content.
- Sites with unnecessary host names.
- Automatically generated webpages.

- Hidden text.
- Cloaking.
- Doorway pages.
- Multiple sites displaying same content.

Figure 2.15 is a depiction of Yahoo! Home page.



http://www.yahoo.com/

Best of Yahoo! - Vote now for your city's best cafés, shops, & more. You could win a 2006 Mazda.

Hi, ferniembikiwa

- Bush lays out Gulf Coast rebuilding plan
- Nagin declares plan to reopen New Orleans
- China offers revised disarmament proposal
- World leaders demand greater U.N. role
- Group: Death of journalists in Iraq not probed
- Eight in 10 want drivers to drop SUVs - poll
- Beloved history book gets English translation
- MLB · NFL · NCAA Football · NHL · Soccer

FIGURE 2.15: Yahoo! home page (**Source:** Yahoo!, 2005a).

2.5.5 AltaVista

AltaVista is a crawler based search engine. Yahoo! powers the search results of AltaVista. AltaVista has an interface with tabs above the search box which allows users to go beyond web search to find images, MP3/Audio, Video, human category listings and news results that has its own technology to spider and rank webpages (see Figure 2.16). AltaVista also makes use of other types of listings in the likes of human edited directory, paid listing section as well as a news events section (AltaVista, 2005).

There are two major ways of getting listed on AltaVista. The AltaVista spider can find a site while crawling the web. The spider generally finds sites which have not been submitted to AltaVista by following links from other sites that have been submitted. Another way of getting listed on AltaVista is by explicitly informing the spider about the existence of a website. This can be done by either submitting a website via the free Add URL feature, or through paid inclusion (AltaVista, 2005).

AltaVista strictly opposes techniques that manipulate search results. Penalty of spamming by AltaVista include placing the website at a low rank or exclusion of the site from the entire index.

The following tactics are treated as spam by AltaVista (AltaVista, 2005):

- Over-repetition of the keywords within a page.
- Using small a font size in order to hide keywords.
- Creating artificial links to boost link popularity.
- Using software programs for page submission.
- Irrelevant keywords.
- Creating software-generated webpages which are very similar to each other.

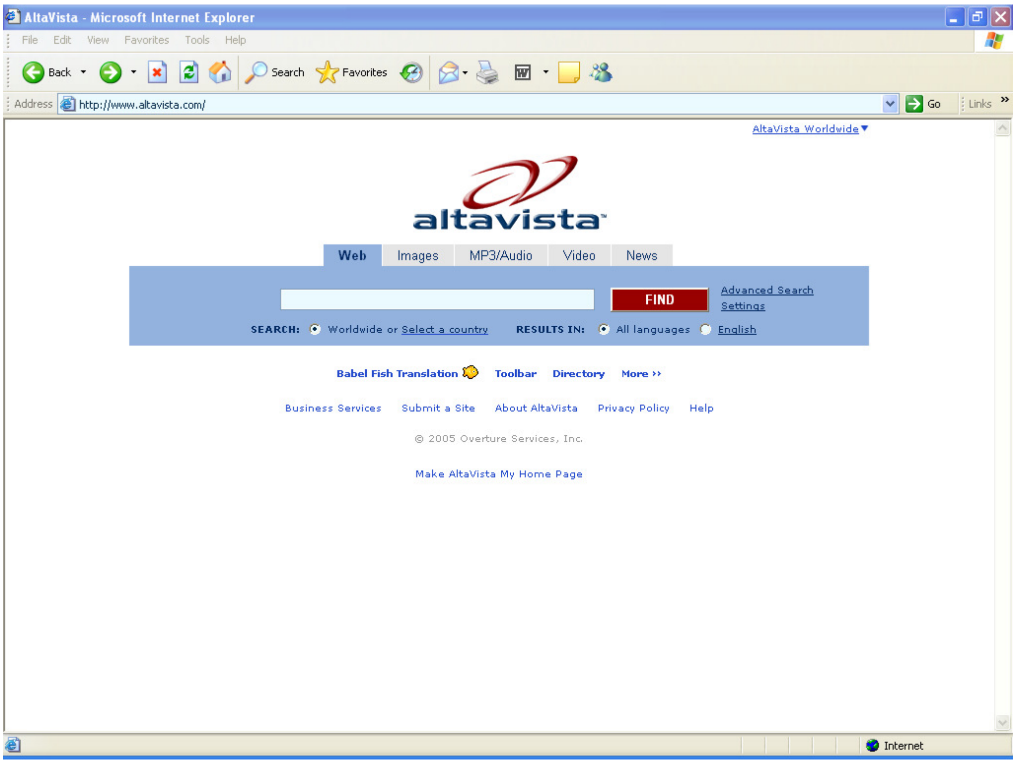

AltaVista - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.altavista.com/
Go
Links
AltaVista Worldwide
altavista
Web
Images
MP3/Audio
Video
News
FIND
Advanced Search
Settings
SEARCH: ● Worldwide or Select a country
RESULTS IN: ● All languages ○ English
Babel Fish Translation
Toolbar
Directory
More ››
Business Services
Submit a Site
About AltaVista
Privacy Policy
Help
© 2005 Overture Services, Inc.
Make AltaVista My Home Page
Internet

FIGURE 2.16: AltaVista home page (**Source:** AltaVista, 2005)

2.6 E-commerce websites

Darch and Lucas (2002: 148) have defined e-commerce as 'the process of doing business electronically where the Internet and its related technologies is the enabler of business processes'.

According to Kim, Shaw and Schneider (2003: 17), the web has become the primary infrastructure for e-commerce. These authors further state that e-commerce has grown to become a significant factor for commercial marketing strategies in the world today, while Podesta, (2000: 73) predicts that companies that do not embrace e-commerce could be destroyed by it. Hsieh and Lin (1998: 113) have stated that the growth of the Internet has led to an influx of companies conducting business online. A report by World Wide Worx (2002) indicates that the number of retail websites in South Africa increased from 215 to 719 during a period of two years. Further research by Davidrajuh refers to e-commerce as a quicker, cheaper global and secure way of gaining better customer value (Davidrajuh, 2003: 434), while Epstein (2005: 23) has classified e-commerce as one of the most essential value added activities any business can adopt.

Cox and Dale (2002: 862) stress the importance of a good e-commerce website by stating that a potentially important customer can be lost if unable to access a website, or if the whole experience of accessing the website proves inadequate. In this respect, Podesta (2000: 73) is of the opinion that:

> "Industry members should not feel threatened by the Internet, but should instead recognize the possibilities the web offers for improved customer knowledge, greater efficiency and faster delivery".

Many customers are demanding efficient services and one of the ways of meeting customer demand efficiently, is through Internet use. Even though the Internet acts as a faster way for doing business, it has also brought along a greater need for customer satisfaction (Podesta, 2000: 74; Cox & Dale, 2002: 862). According to Barnes and Vidgen, (2002: 114), a key challenge in e-commerce adoption is understanding customer requirements and designing websites that will serve the needs of the customer.

Too much attention is paid to the interface of the website, which ends up looking good, but causing frustration during navigation, resulting in customers having difficulty in finding what they are searching for (Cox & Dale, 2002: 862). A website that is difficult to use usually projects a poor image of the organization it represents and is likely to result in loss of revenue as potential customers will not visit it again (Barnes & Vidgen, 2002: 114).

When a customer accesses a website, the perception of how the transaction will be executed as well as the company's image, is determined by the appearance of the website. This indicates that a poorly designed website can lead to a loss of revenue. A poorly designed website can be avoided by following certain guidelines or criteria of website design (Kim *et al.,* 2003: 18). According to research done by Kim *et al.* (2003: 19), website users are more concerned with the information and content that is available on a webpage. Thelwall (2000a: 152) has stated the following criteria for a quality website:

- Site visibility in search engines.
- Ease of use.
- Design quality.
- Ease of site maintenance and updating.

Cox and Dale (2002: 863-871) have classified key quality factors of a website as the following:

- Clarity of purpose.
- Design.
- Accessibility and speed.
- Content.
- Customer service.
- Customer relationships.

Kim *et al.* (2003: 19-20) classify attractiveness as an important criteria for measuring quality of a website. These authors's further state that a website should be clear and contain relevant information. Website designers should always keep in mind that if a user is not satisfied with a website, that user is likely not to visit the site again. According to Yates (2005: 182), accessibility and usability of a website are the key attributes of a good website. Thelwall (2001: 114) is of the opinion that the effectiveness of a site is often linked to the amount of potential customers that actually visit the site.

As with most initiatives, there are guidelines for implementing successful e-commerce. According to Epstein, successful e-commerce initiatives require company strength in leadership, strategy, structures and systems. The same author also states that another key area for successful e-commerce lies in the implementation, and could culminate into failure if not executed correctly (Epstein, 2005: 24-25). Kim *et al.* *(*2003: 17) state that businesses that acknowledge the importance of well designed websites as a critical success factor for e-commerce often reap the benefits of a successful e-commerce initiative.

E-commerce facilitates the ability for customers to conduct business with companies that were previously not available to them. As one of the fastest growing components of the Internet, e-commerce has brought along ways of conducting business that companies cannot afford to ignore, as it has the potential of reaching global customers. All the physical barriers associated with traditional business have been broken down, because of the existence of e-commerce (Wen, Chen & Hwang, 2001: 5). E-commerce has also ensured that both consumers and companies have a more flexible, less costly and faster way of doing business. The variety of products and information available via e-commerce also allow for faster decision making (Simeon, 1999: 297).

With the expansion of e-commerce, the importance of a well designed website also increases. This is due to the fact that e-commerce websites are the main interface between a business and a customer, making the need for a well designed website of vital importance.

2.7 Conclusion

Trading online (e-commerce) has proven to be a commercial benefit for many firms. More companies are adopting e-commerce in an attempt to stay competitive. In South Africa, the number of businesses trading online increased by 234.4% within a period of two years (World Wide Worx, 2002). This growth indicates that more businesses are realizing the potential benefit of trading online. However, it is essential for users to visit these websites;

otherwise the benefits of e-commerce are not achieved. As more web designers realized the benefit of appearing in the top results of search engines (see Table 2.3), more SEO techniques come into practice. From the conducted literature review, it is evident that there are many SEO practices that are often adopted by web designers in an attempt to increase website ranking.

It has also become evident that some practices that are classified as unethical SEO practices by certain search engines are not viewed as such by others. Search engines have policies that state what is considered as spam. However, search engine policies have proven to be inconsistent as some SEO unacceptable practices are acceptable to others. One practice which has been the focus of controversy is cloaking (see section 2.4.2.1). It is generally considered as spam, and yet some of the policies do not mention it. However, information on Google states that assumptions should not be made about spam techniques that are not listed on search engine policies (Google, 2005d).

Sullivan, (2004f) provides the following as a solution to spam:

> “The search engines should agree to publish lists of companies they’ve banned. That would help consumers seeking SEM firms to understand which to avoid. If they do use a banned firm, at least they were warned of the consequences of going with a rule breaker”.

This approach however, would only help with companies that use SEM firms, and not web designers that design their own sites. In one study Alimohammadi, 2003: 240) argues that the biggest problem pertaining to keywords meta-tags is spam, while in another study the same author states that meta-tags can enable the precise and efficient analysis of websites by search engines (Alimohammadi, 2004: 220). However Garofalakis *et al.*

(2002: 44) argue that stuffing many keywords in meta-tags is a trick that is exploited by many web designers to achieve high rankings, while Ananzi states that sites without meta-tags will be rejected. Another SEO practice that has no set standard is page redirects. While some authors view it as spam because the user is redirected to another site without their intervention, Google states that some page redirects will not be accepted, thus causing confusion. The Ananzi policy does not mention page redirects, leaving the unanswered question of whether it considers page redirects as spam. AskJeeves states that it considers pages designed to lead users to other sites as spam, and a redirect leads users to other sites. However, research done by Sullivan (2004f) indicates that even though search engine rules differ, they are now becoming more similar.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, the research methodology deployed to investigate the exclusion policies of search engines will be expanded upon. In the sections that follow, different aspects of research methodology in the likes of research design, research environment, study population, sample size and sampling will be explained. Design methods that have been chosen will be elaborated on, with justification for choice.

According to Cooper and Schindler (2003: 24), research has been defined as "any organized inquiry carried out to provide information for solving problems". The same authors also indicate that good research is characterised by a number of elements including the honest and complete reporting of procedures. There are various research methods that exist. Furthermore, it is also important for researchers to have clear understanding of the methodologies that are available, as well as the chosen one (Hines, 2000: 7).

The main aim of a researcher is to understand and interpret the collected data. According to Cooper and Schindler (2003: 5,14), studying research methods equips a researcher with the necessary skills to solve problems while good research will generate data that is dependable and reliable.

3.2 Research design

Cooper and Schindler (2003: 170) have defined research design as "the strategy for a study and the plan by which the strategy is carried out". There are different research design approaches which are either qualitative or quantitative.

3.2.1 Qualitative research

According to Struwig and Stead (2001: 11), qualitative research is associated with a lot of research methods and therefore cannot be easily defined. However, other authors have made a distinction between qualitative and quantitative research by stating that qualitative studies usually aim for depth of understanding instead of quantity of understanding, as is the case with quantitative research (Henning, Van Rensburg & Smit, 2004:3). The following have been identified as approaches pertaining to qualitative research:

- Document analysis.
- In-depth interviewing.
- Participant observation.
- Films, photographs and video tape.
- Projective techniques.
- Case studies.
- Elite or expert interviewing
- Street ethnography.
- Other observational techniques.

3.2.2 Quantitative research

Quantitative research, which has been differentiated from qualitative research by the fact that it requires the research data to be expressed in numbers, has been defined as "conclusive research involving large representative samples and structured data collection procedures" (Struwig & Stead, 2001: 4). Another author indicates that data in quantitative research is usually collected by means of a survey, an experiment or through observation (Hines, 2000: 8).

3.3 Research methods

3.3.1 Exploratory research

According to Struwig and Stead (2001: 7), this type of approach involves researching problems where little research has been done and involves the collection of a large quantity of information from a small sample. Cooper and Schindler (2003: 281) state that this approach to research is more useful in situations where the researcher does not have clear idea of what to expect from the study as well as where the area is vague. Exploratory research, according to Struwig and Stead (2001: 7) can be carried out in the following ways:

- Information gathered from secondary sources.
- Selecting certain cases to analyse.
- Carrying out a survey with individuals that have opinions on the subject matter.

3.3.2 Descriptive research

Unlike exploratory research which tends to be more flexible, descriptive research aims at describing something in an attempt to give a more complete and accurate analysis of a situation (Struwig & Stead, 2001:8). According to Amaratunga, Baldry, Sarshar and Newton (2002:26), descriptive research is concerned with data which is gathered through the use of interviews or mailed questionnaires.

3.3.3 Case studies

According to Remenyi and Money, (2004: 72) a case study is “a sophisticated research tactic for establishing valid and reliable evidence for the research process as well as presenting findings that result from the research”. Case studies can be applied in both qualitative and quantitative studies (Näslund, 2002: 330). Remenyi and Money (2004: 72) indicate that case studies enable researchers to concentrate on specific instances in order to identify process that may not be visible in larger scale surveys.

3.3.4 Statistical methods

This research approach differs from the case study in that research is conducted on a larger sample while examining few variables in those samples (Struwig & Stead, 2001: 8). Cooper and Schindler (2003: 150) indicate that statistical methods are designed primarily for broad research and not in-depth research.

3.4 Survey environment

E-commerce has enabled the breaking down of physical barriers that were built by traditional business, and has evolved into a quicker, cheaper and more secure way of gaining better customer value (Wen *et al.,* 2001; Davidrajuh, 2003). The number of retail sites listed in South Africa was 215 at the end of 2001, but grew up to an estimated 719 at the end of 2003, with a value of R341 million sales achieved in 2003 (World Wide Worx, 2004).

However, an e-commerce website needs to be well optimized in order to achieve high ranking in search engines as a well optimized website is likely to attract more traffic, and subsequently boost sales (Van Steenderen, 2001). This process is often hindered by the fact that there is no set definition for what is considered spam by search engines.

3.4.1 Population

According to Cooper and Schindler (2003: 179), a population is an object of which a measurement or study is undertaken. It involves the entire collection of elements on which studies are to be carried out on. Saunders, Lewis and Thornhill (1997: 124) define a population as “the full set of cases from which the sample is taken”.

3.4.2 Sample

A good sample should be accurate and precise (Cooper & Schindler, 2003: 210). Saunders *et al.* (1997: 124-125) state that sampling techniques provide some methods that enable a researcher to reduce the amount of data collected. These authors further state the following as reasons to deploy sampling techniques:

- If it is impractical to survey the entire population.
- Budget constraints.
- Time constraints.
- All data has been collected but results are needed quickly.

Cooper and Schindler (2003: 181) have stated the following additional reasons for sampling:

- Greater accuracy of the results.
- Availability of population elements.

For this study, a total population of 4985 would not be practical to work with, as a lower population can produce valid results. Saunders *et al*. (1997: 125) indicate that a large population does not necessarily mean more valid results than a census survey. The organization of the data collection of a smaller population is more manageable. Furthermore, the costs implications of analysing a total population of 4985 are not as economical as a smaller sample size, as it is cheaper to work with a smaller sample (Saunders *et al*., 1997: 125).

According to Cooper and Schindler (2003: 183) there are two types of sampling, namely probability sampling and non probability sampling.

3.4.2.1 Probability sampling

This type of sampling is based on a random selection. The chances of cases being selected are equal. According to Saunders *et al*. (1997: 126), probability sampling is most commonly used in survey-based research. The following are the sample techniques associated with probability sampling (Cooper & Schindler, 2003: 199).

- **Simple random:**

  With this technique, each element has an equal chance of being selected.

- **Systematic:**
  Elements are selected from the population by assigning a number to the first element, and selecting the *k*th element after that.
- **Cluster:**
  The population is divided into subgroups with some of them being selected for the study.
- **Stratified:**
  This technique utilizes the dividing of the population in to sub populations based on certain attributes and using the simple random technique on each sub population (Saunders *et al.*, 1997: 137).
- **Double:**
  Within this technique, data is collected from a sample using a previously defined technique. A sub sample is then selected for further study, based on the information collected (Cooper & Schindler, 2003: 199).

3.4.2.2 Non probability sampling

According to Cooper and Schindler (2003: 200), elements do not have equal chances of being selected. The probability of each element being chosen is not known.

- **Convenience:**
  Saunders *et al*. (1997: 147) state that convenience sampling involves selecting cases which are easy to obtain from a sample. However the same authors also state that this type of sampling is prone to bias.
- **Purposive sampling:**
  Cooper and Schindler (2003: 201) state that there are two types of purposive sampling, being Judgement sampling and quota sampling.
  - **Judgement sampling:**
    This type is applied when a researcher selects samples that conform to a certain criteria.

- **Quota sampling**:

  Cooper and Schindler (2003: 201) state that this type of sampling is used to improve representativeness. This sampling is based on the logic that certain characteristics represent the dimensions of the population. According to Saunders *et al.* (1997: 143), quota sampling is normally used in instances where there is a large population.

- **Snowball:**

  A study done by Saunders *et al.* (1997: 147) has indicated that snowball sampling is more appropriate in cases where it is difficult to identify members of the required population. In this research, the author did not apply any snowball sampling as the population was not difficult to identify.

3.5 Methodology for this study

According to Cooper and Schindler (2003:152), document analysis involves the evaluation of records, reports, documents or opinions. The data in this study was collected by means of document analysis to which both qualitative and descriptive research methods were applied, but the results of the analysis were quantified to determine frequency of the investigated variables in the result list.

There has been some research done regarding what is considered SEO spam and what is good SEO practice. However, no evidence could be found of empirical work which has defined spam and the counter-measures. This may be complicated by the fact that practices that are acceptable with one search engine are not necessarily so with another. Exploratory research was used to determine what the five search engines that are the focus of this study consider as spam, and exactly how well they adhere to their own policies. More clarity on what other experts in the search engine industry consider as spam was also gained through the literature review.

Due to the number of search engines that had to be investigated, as well as e-commerce websites that were tested, the author decided to refrain from using a case study approach. A population of websites was obtained from the Cape Chamber of Commerce, which had an initial population of 4482, and onlineshopping.co.za which had an initial population of 141. Additional websites were added from Ananzi's shopping and auction home page, which had an initial population of 362. Looking at the census of the population at this stage, some sampling techniques had to be applied to the population to reduce it to a manageable size, as well as to eliminate sections of the population that did not meet the study criteria. For this study, the population required that the elements of study are fully functional e-commerce websites, and therefore the author applied judgemental sampling to extract functional e-commerce websites. From the initial population of websites, the author made the following selection:

- A selection of businesses with websites.
- This selection was filtered to exclude those businesses that are not e-commerce websites.
- From that list, more judgemental sampling was applied to obtain a final list of fully functional e-commerce websites (see Appendix A).

According to Saunders *et al*. (1997: 145) the choice of selecting judgemental sampling should be based on the research question. The research question of this study is based on the assumption that all websites investigated will be fully functional e-commerce websites. Another sampling technique was applied to the list to reduce it to a manageable number. The author applied a random number to all of the websites and sorted them according to the random number. The first 51 websites were then chosen for this study (see Figure 3.2). However, from the list of these randomly selected websites, four websites were no longer operational (see Appendix D), bringing the final website census to 47. The steps taken to design the research methodology for this thesis are summarized in Figure 3.1.

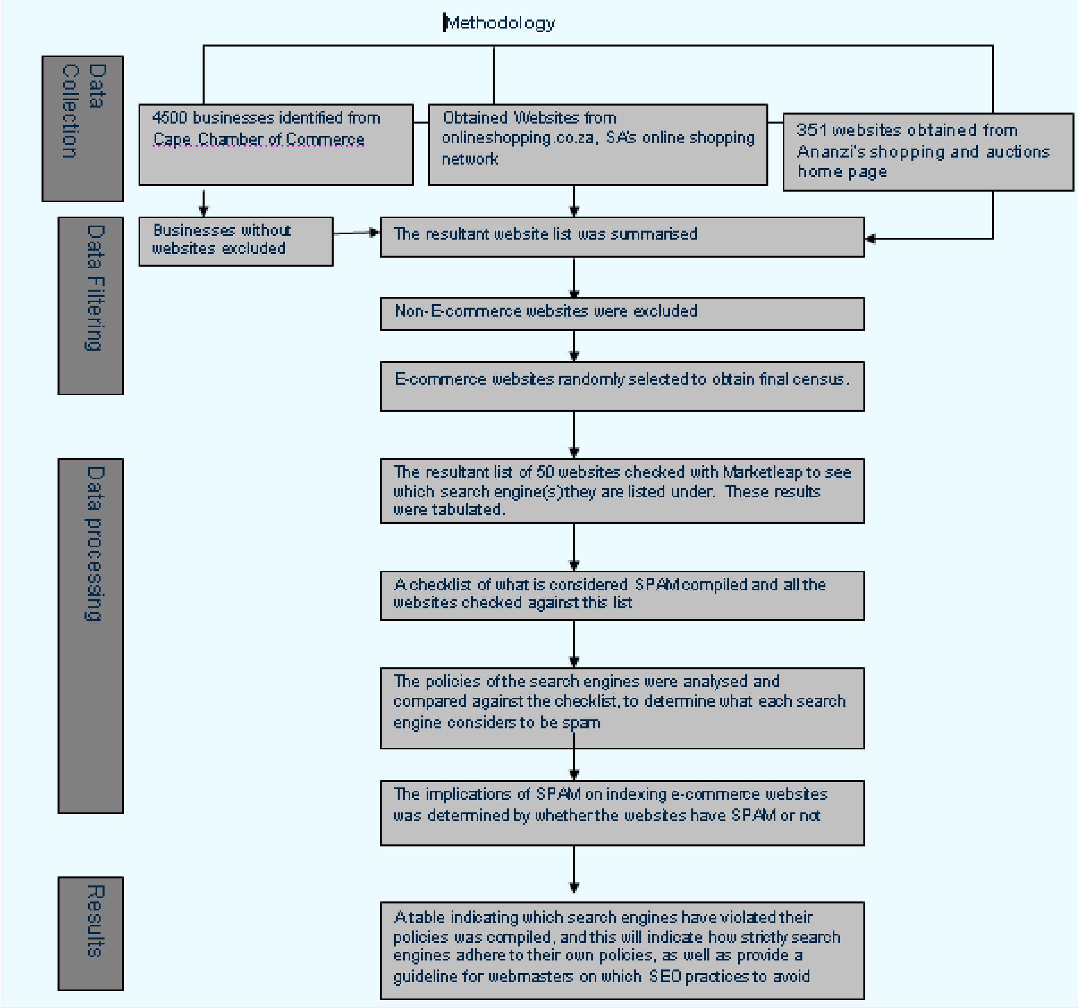
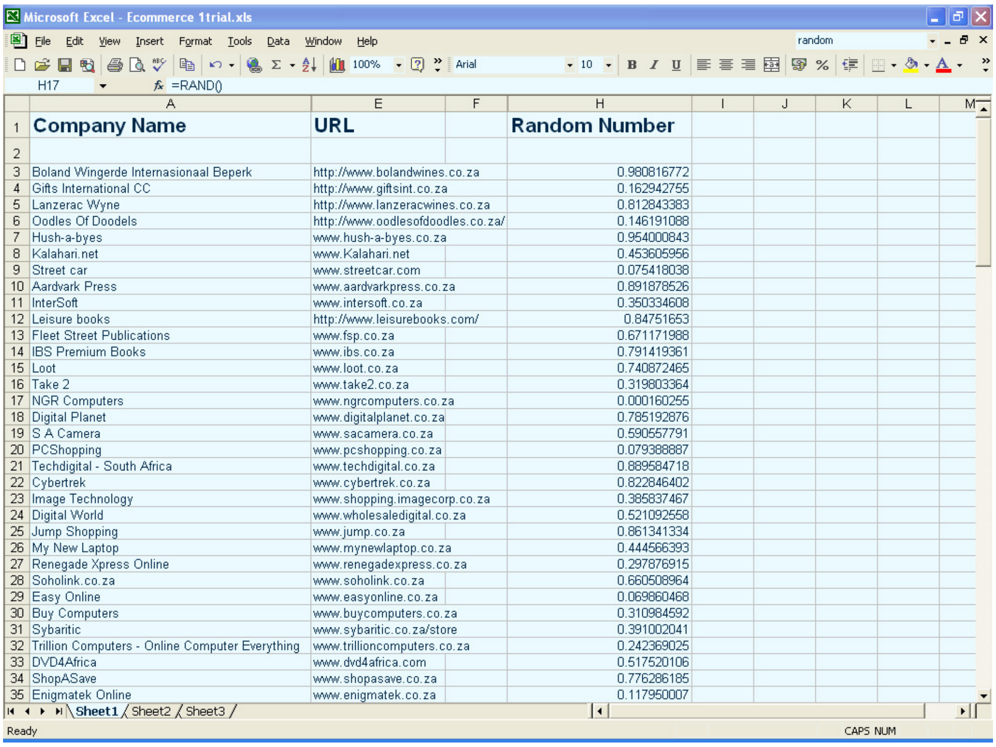

Methodology
Data Collection
Data Filtering
Data processing
Results
4500 businesses identified from Cape Chamber of Commerce
Obtained Websites from onlineshopping.co.za, SA's online shopping network
351 websites obtained from Ananzi's shopping and auctions home page
Businesses without websites excluded
The resultant website list was summarised
Non-E-commerce websites were excluded
E-commerce websites randomly selected to obtain final census.
The resultant list of 50 websites checked with Marketleap to see which search engine(s)they are listed under. These results were tabulated.
A checklist of what is considered SPAM compiled and all the websites checked against this list
The policies of the search engines were analysed and compared against the checklist, to determine what each search engine considers to be spam
The implications of SPAM on indexing e-commerce websites was determined by whether the websites have SPAM or not
A table indicating which search engines have violated their policies was compiled, and this will indicate how strictly search engines adhere to their own policies, as well as provide a guideline for webmasters on which SEO practices to avoid

FIGURE 3.1: Methodology process



Microsoft Excel - Ecommerce 1trial.xls
File Edit View Insert Format Tools Data Window Help
random
100%
Arial
10
Sheet1
Sheet2
Sheet3
Ready
CAPS NUM

H17 =RAND()

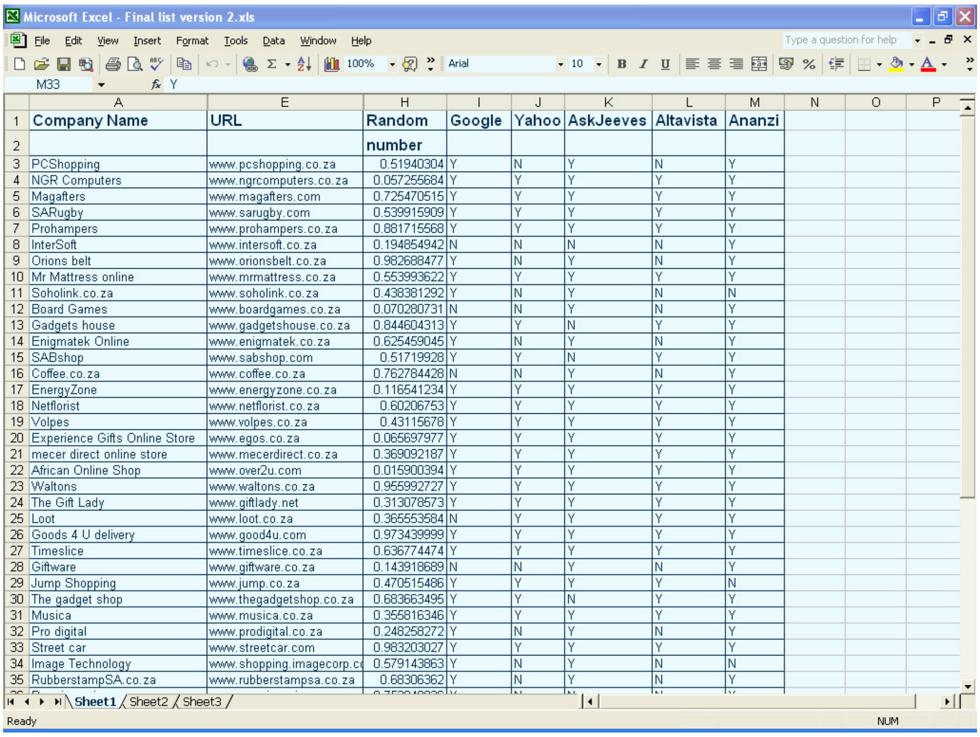
	A	E	F	H
1	Company Name	URL		Random Number
2				
3	Boland Wingerde Internasionaal Beperk	http://www.bolandwines.co.za		0.980816772
4	Gifts International CC	http://www.giftsint.co.za		0.162942755
5	Lanzerac Wyne	http://www.lanzeracwines.co.za		0.812843383
6	Oodles Of Doodels	http://www.oodlesofdoodles.co.za/		0.146191088
7	Hush-a-byes	www.hush-a-byes.co.za		0.954000843
8	Kalahari.net	www.Kalahari.net		0.453605956
9	Street car	www.streetcar.com		0.075418038
10	Aardvark Press	www.aardvarkpress.co.za		0.891878526
11	InterSoft	www.intersoft.co.za		0.350334608
12	Leisure books	http://www.leisurebooks.com/		0.84751653
13	Fleet Street Publications	www.fsp.co.za		0.671171988
14	IBS Premium Books	www.ibs.co.za		0.791419361
15	Loot	www.loot.co.za		0.740872465
16	Take 2	www.take2.co.za		0.319803364
17	NGR Computers	www.ngrcomputers.co.za		0.000160255
18	Digital Planet	www.digitalplanet.co.za		0.785192876
19	S A Camera	www.sacamera.co.za		0.590557791
20	PCShopping	www.pcshopping.co.za		0.079388887
21	Techdigital - South Africa	www.techdigital.co.za		0.889584718
22	Cybertrek	www.cybertrek.co.za		0.822846402
23	Image Technology	www.shopping.imagecorp.co.za		0.385837467
24	Digital World	www.wholesaledigital.co.za		0.521092558
25	Jump Shopping	www.jump.co.za		0.861341334
26	My New Laptop	www.mynewlaptop.co.za		0.444566393
27	Renegade Xpress Online	www.renegadexpress.co.za		0.297876915
28	Soholink.co.za	www.soholink.co.za		0.660508964
29	Easy Online	www.easyonline.co.za		0.069860468
30	Buy Computers	www.buycomputers.co.za		0.310984592
31	Sybaritic	www.sybaritic.co.za/store		0.391002041
32	Trillion Computers - Online Computer Everything	www.trillioncomputers.co.za		0.242369025
33	DVD4Africa	www.dvd4africa.com		0.517520106
34	ShopASave	www.shopasave.co.za		0.776286185
35	Enigmatek Online	www.enigmatek.co.za		0.117950007

FIGURE 3.2: Website list after random selection.

3.5.1 Document analysis

3.5.1.1 Websites

After the sampling of the initial population, this author utilized Marketleap to determine if the websites that are to be used in the study were registered with search engines. Another program known as WebPositionGold, which is also able to determine if websites are registered with search engines was available but the cost implications of obtaining it resulted in the author deciding to refrain from using it. The websites were then tabulated to indicate which search engine(s) they were registered with. The resultant list reflected which website was registered with which search engine as well as websites that are not indexed by any search engine (see Figure 3.3).



Microsoft Excel - Final list version 2.xls

M33 Y

	A	E	H	I	J	K	L	M
1	Company Name	URL	Random	Google	Yahoo	AskJeeves	Altavista	Ananzi
2			number					
3	PCShopping	www.pcshopping.co.za	0.51940304	Y	N	Y	N	Y
4	NGR Computers	www.ngrcomputers.co.za	0.057255684	Y	Y	Y	Y	Y
5	Magafters	www.magafters.com	0.725470515	Y	Y	Y	Y	Y
6	SARugby	www.sarugby.com	0.539915909	Y	Y	Y	Y	Y
7	Prohampers	www.prohampers.co.za	0.881715568	Y	Y	Y	Y	Y
8	InterSoft	www.intersoft.co.za	0.194854942	N	N	N	N	Y
9	Orions belt	www.orionsbelt.co.za	0.982688477	Y	N	Y	N	Y
10	Mr Mattress online	www.mrmattress.co.za	0.553993622	Y	Y	Y	Y	Y
11	Soholink.co.za	www.soholink.co.za	0.438381292	Y	N	Y	N	N
12	Board Games	www.boardgames.co.za	0.070280731	N	N	Y	N	Y
13	Gadgets house	www.gadgetshouse.co.za	0.844604313	Y	Y	N	Y	Y
14	Enigmatek Online	www.enigmatek.co.za	0.625459045	Y	N	Y	N	Y
15	SABshop	www.sabshop.com	0.51719928	Y	Y	N	Y	Y
16	Coffee.co.za	www.coffee.co.za	0.762784428	N	N	Y	N	Y
17	EnergyZone	www.energyzone.co.za	0.116541234	Y	Y	Y	Y	Y
18	Netflorist	www.netflorist.co.za	0.60206753	Y	Y	Y	Y	Y
19	Volpes	www.volpes.co.za	0.43115678	Y	Y	Y	Y	Y
20	Experience Gifts Online Store	www.egos.co.za	0.065697977	Y	Y	Y	Y	Y
21	mecer direct online store	www.mecerdirect.co.za	0.369092187	Y	Y	Y	Y	Y
22	African Online Shop	www.over2u.com	0.015900394	Y	Y	Y	Y	Y
23	Waltons	www.waltons.co.za	0.955992727	Y	Y	Y	Y	Y
24	The Gift Lady	www.giftlady.net	0.313078573	Y	Y	Y	Y	Y
25	Loot	www.loot.co.za	0.365553584	N	Y	Y	Y	Y
26	Goods 4 U delivery	www.good4u.com	0.973439999	Y	Y	Y	Y	Y
27	Timeslice	www.timeslice.co.za	0.636774474	Y	Y	Y	Y	Y
28	Giftware	www.giftware.co.za	0.143918689	N	N	Y	N	Y
29	Jump Shopping	www.jump.co.za	0.470515486	Y	Y	Y	Y	N
30	The gadget shop	www.thegadgetshop.co.za	0.683663495	Y	Y	N	Y	Y
31	Musica	www.musica.co.za	0.355816346	Y	Y	Y	Y	Y
32	Pro digital	www.prodigital.co.za	0.248258272	Y	N	Y	N	Y
33	Street car	www.streetcar.com	0.983203027	Y	Y	Y	Y	Y
34	Image Technology	www.shopping.imagecorp.co	0.579143863	Y	N	Y	N	N
35	RubberstampSA.co.za	www.rubberstampsa.co.za	0.68306362	Y	N	Y	N	Y

FIGURE 3.3: Website search engine listing

3.5.1.2 Search engine policies

The policies of the search engines used in this study were analysed to determine what each search engine considers to be spam. Discrepancies emerged when the search engine policies were compared, as some search engine optimization practices were considered as spam by certain search engines, while with other search engines it was considered as acceptable (See Table 3.1).

TABLE 3.1: Search engine spam summary

SPAM	GOOGLE	YAHOO!	ASKJEEVES	ALTAVISTA	ANANZI
Keyword Stuffing	0	0	0	0	0
Hidden Text	0	0	0	0	0
Tiny Text	0	0	0	0	0
Hidden Links	0	0	1	0	0
Artificial Link farms	1	0	1	0	1
Page Swapping	1	0	0	0	1
Sneaky page Redirects	0	0	0	0	1
Duplicate Pages	0	0	0	0	Duplicate entries allowed under special circumstances
Doorway pages	0	0	0	0	0
Cloaking	0	0	0	1	0
Automated Queries	0	0	0	0	No guarantee
Meta-tags stuffing	1	1	1	1	Sites without meta -tags will be rejected
Southern African based content	1	1	1	1	0
Meta-tags must accurately describe content	1	1	1	1	0
Full entry in capital letters	1	1	1	1	0
Use of affiliate or referral programs	1	1	0	1	1

KEY: Policy advises against spam – 0
Not stated on policy – 1

Some judgemental sampling was applied to Table 3.1. This author initially selected the spam tactics that are mentioned in most of the policies. From this list, spam tactics that were detectable by humans were selected. The final results of this sampling are indicated in Table 3.2.

Table 3.2: Detectable Spam practices

Spam
Keyword stuffing
Invisible text
Tiny text
Hidden links
URL spam
Page redirects
Doorway pages
Meta-tag stuffing

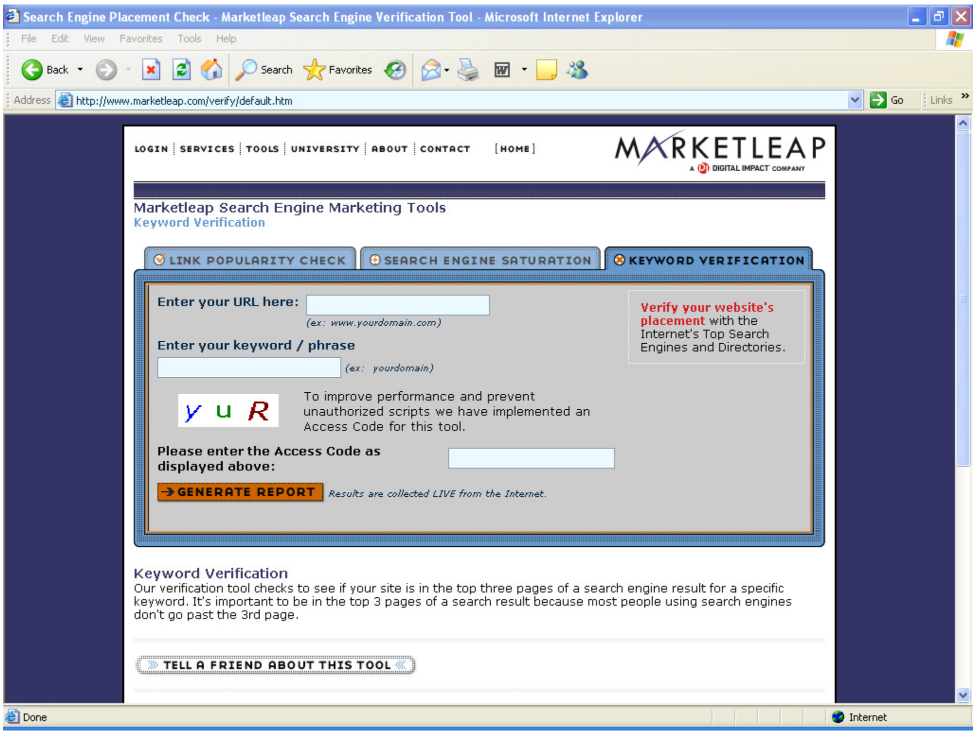
3.5.1.3 Types of data

According to Struwig and Stead (2001: 40), data can either be numeric or non-numeric. It can also be verbal or non verbal. The data that has been collected for this research is non-numeric. This data is in the form of search engine policies, common SEO practices and SEO tactics that are generally considered as spam.

3.5.1.4 Sources of data

Struwig and Stead (2001: 41) have stated that the sources of data can be primary data, secondary data or commercial data. Under primary data, the researcher personally collects data, while secondary data is data that has already been collected. Commercial data refers to the data that has been collected for market research purposes. For purposes of this study secondary data was utilized. The author collected search engine policies from the search engine websites. Websites that were used in this study were collected from the Cape Chamber of Commerce, onlineshopping.co.za and Ananzi’s shopping and auctions home page. Other data was attained during the through analysis of the literature review, contained within the ambit of Chapter 3.

All the websites that were identified were analysed to ensure that they were fully functional e-commerce websites. The SEO practices that are considered as spam were identified and listed in Table 3.1. The source code of the e-commerce websites was analysed to determine if it contained any of the practices classified as spam. Marketleap was used to identify websites that were listed in search engines. This software also indicated whether the websites listed are in the top 10, 20 or 30 search engine result lists, as well as websites that had page redirects.


Search Engine Placement Check - Marketleap Search Engine Verification Tool - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
LOGIN | SERVICES | TOOLS | UNIVERSITY | ABOUT | CONTACT [HOME]
MARKETLEAP
A DI DIGITAL IMPACT COMPANY
LINK POPULARITY CHECK
SEARCH ENGINE SATURATION
KEYWORD VERIFICATION
y u R
GENERATE REPORT
TELL A FRIEND ABOUT THIS TOOL
Done
Internet

http://www.marketleap.com/verify/default.htm

Marketleap Search Engine Marketing Tools
Keyword Verification

Enter your URL here: (ex: www.yourdomain.com)

Enter your keyword / phrase (ex: yourdomain)

To improve performance and prevent unauthorized scripts we have implemented an Access Code for this tool.

Please enter the Access Code as displayed above:

Results are collected LIVE from the Internet.

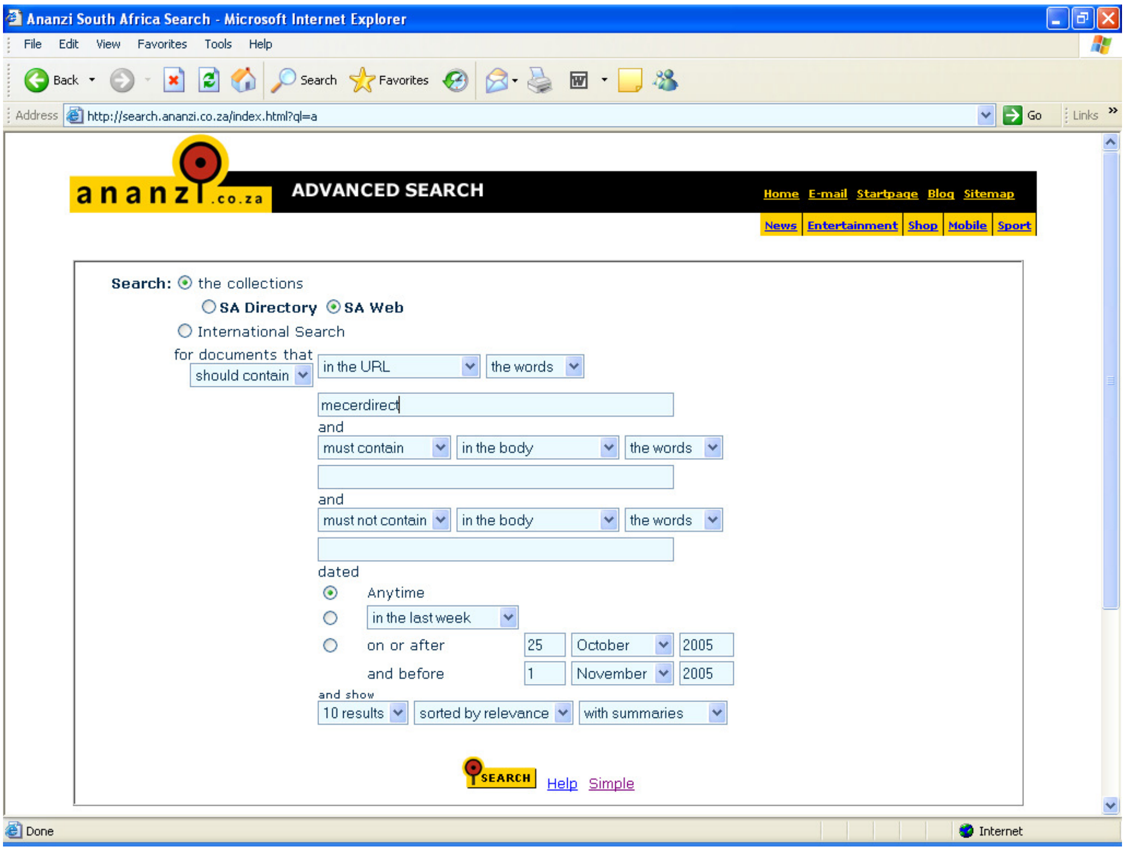
Verify your website's placement with the Internet's Top Search Engines and Directories.

Keyword Verification
Our verification tool checks to see if your site is in the top three pages of a search engine result for a specific keyword. It's important to be in the top 3 pages of a search result because most people using search engines don't go past the 3rd page.

FIGURE 3.4: Homepage of Marketleap (**Source:** Marketleap.com, 2005)

Furthermore, Marketleap was used to determine whether websites were listed in search engines or not. The websites that were not listed in any search engine were also used in the study. This instrument was sufficient in that it collected the results live from the Internet; however results from AskJeeves and

Ananzi were not part of the results list. This author then manually searched the above mentioned search engines for the same websites to determine if the websites are listed in those search engines. The resultant list was tabulated and reflected in Appendix B.



Ananzi South Africa Search - Microsoft Internet Explorer

http://search.ananzi.co.za/index.html?ql=a

Search: ● the collections

○ SA Directory ● SA Web

○ International Search

for documents that should contain | in the URL | the words

mecerdirect

and must contain | in the body | the words

and must not contain | in the body | the words

dated

● Anytime

○ in the last week

○ on or after 25 October 2005

and before 1 November 2005

and show 10 results | sorted by relevance | with summaries

FIGURE 3.5: Ananzi advanced searching page (**Source:** Ananzi, 2005d)

Figure 3.5 represents the interface of the advanced searching feature for Ananzi. This author made use of this feature to manually determine if the websites are listed under Ananzi as the software which was used (see Figure 3.4) did not check for websites listed under Ananzi. Four possible outcomes of webpages with and without spam, were identified and are listed in Table 3.3.

TABLE 3.3: Possible conclusions

CLASS	DESCRIPTION
Class 1	Websites containing what is considered as spam, but listed under one or more search engines
Class 2	Websites that contain spam and are not listed in any of the search engines
Class 3	Websites that do not contain any form of spam and are listed in one or more search engines
Class 4	Websites that do not contain spam, but are not listed in any of the search engines

3.6 Conclusion

In this chapter, the author identified several approaches to research. From analysing the literature, as well as the search engine policies, this author was able to compile a table of what is generally considered as spam (see Table 3.1). Judgemental sampling was applied to this table to reduce it to a manageable size (see Table 3.2), and this table in turn was used in the analysis of the websites.

CHAPTER 4
RESULTS AND ANALYSIS

4.1 Introduction

In this chapter, an analysis and interpretation of the data that was collected in the previous chapter will be reported upon. This author has followed the principle of maintaining objectivity in this chapter to ensure that the data that has been collected is not misrepresented and that conclusions reached are not distorted (Saunders *et al.,* 1997: 114). The results of determining if any of the websites contained spam will be reported on, and compared to the literature review, as well as search engine policies.

Websites that were not listed under any search engines were also evaluated in an attempt to reach a conclusion on possible reasons for their exclusion. Possible conclusions for failure of listings of such websites were drawn.

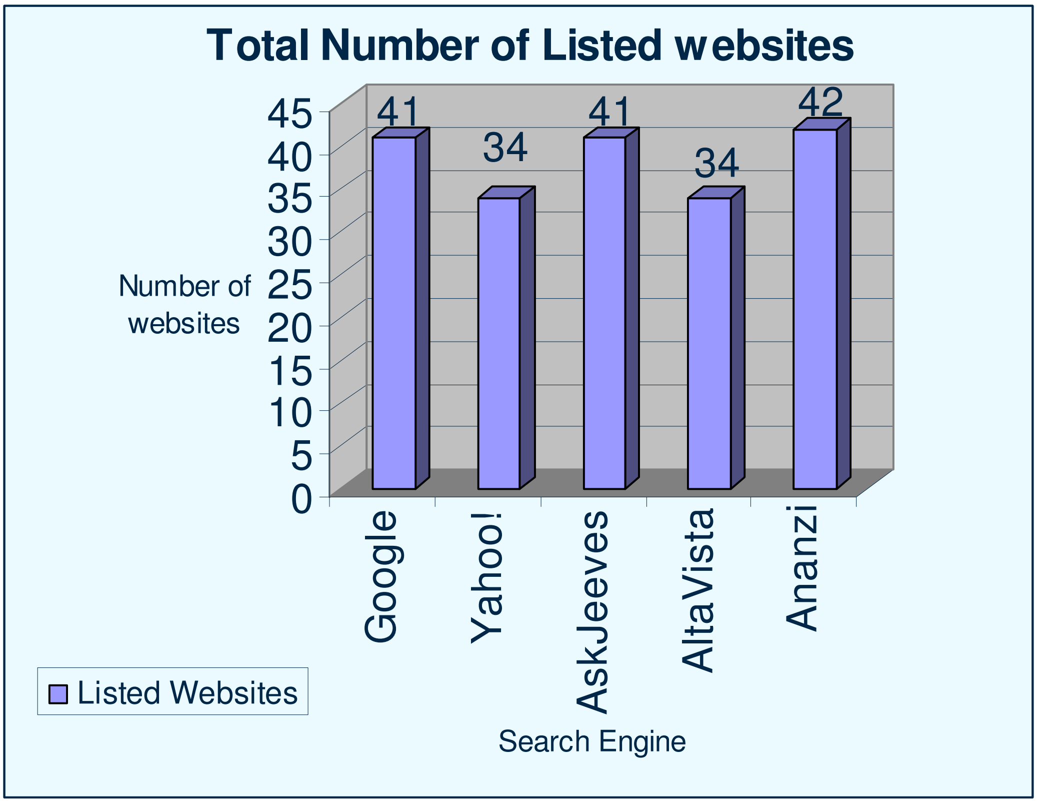
4.2 Website analysis

The aim of this thesis was to determine the impact of search engine exclusion policies on indexing e-commerce websites. One way of determining these implications was through analysing the source code of some websites to determine if they contain any of the practices that were identified as spam. As explained in the previous chapter, a table containing a number of practices that were identified as spam was compiled (see Table 3.1). Due to financial constraints, the websites could not be checked against the entire list of spam. The author made use of judgemental sampling to identify spam that could be detected by human editors (see Table 3.2). The websites were analysed on the basis of this table. The result of this analysis is elaborated more in the following paragraphs.

The following process was adhered to during the analysis of the websites:

- The websites were classified to indicate whether they are listed by the five search engines. This classification was presented in a graph (see Appendix B, Graph 4.1).
- According to Saunders *et al*. (1997:125), a large population does not necessarily mean more valid results. During the analysis of the data, the author found four websites from the list of 51 randomly selected websites were no longer available (see appendix D); reducing the sample list to 47.
- The source code as well as the home pages of the sample list of websites was evaluated to determine if they contain any of the identified spam practices.

Graph 4.1 is an indication of how many of the 47 websites were listed by individual search engines. From the graph, Google indicates that 41 of the 47 websites were registered.



GRAPH 4.1: Website listing

The following process was adopted to detect the occurrence of spam, which was conducted by professional web designers (see Appendix E).

- **Keyword stuffing:** the meta-tags of the specific pages were checked to see if they contained any keyword stuffing. According to Thurow (2003: 221) keyword stuffing/stacking can be placed in any HTML code (see Appendix C).
- **Invisible/Tiny text:** these were checked by highlighting the page, dragging the cursor from the top of the page to the bottom of the page. Any invisible text or tiny text would have been highlighted.
- **Meta-tag stuffing:** Meta-tags were checked to see if they contained repeated keywords or keywords that were not related to the website.
- **Page redirects:** This was identified by looking at the URL in the status bar before clicking a link and the source code. By looking at the status bar before clicking a link, Java redirect code in the link could be

detected. The URL was observed closely to determine if it changed to another URL before opening the webpage.

- **Doorway page:** the author searched for the websites and clicked on the resulting links. According to Wikipedia.com (2005c), doorway pages usually have a 'click here to enter'.
- **URL spam:** The URL was examined to detect if it is long and contains sequences of spam terms

The results of the analysis are summarised in Table 4.1.

TABLE 4.1: Overall analysis results

Search Engine	Number of websites listed	Number of websites with spam	Number of listed websites with spam	Number of unlisted websites with spam	Percentage of websites with spam	Number of occurrences of spam identified(can be one or more per website)	
						Page redirects	Keyword stuffing
Google	41	10	9	1	21.9%	7	3
Yahoo!	34	10	7	3	20.6%	5	3
AskJeeves	41	10	7	3	17.1%	5	3
AltaVista	34	10	7	3	20.6%	5	3
Ananzi	42	10	8	2	19.1%	6	4

As reflected in Table 4.1 the two most commonly identified spam elements in the websites that were analysed were keyword stuffing and page redirects. During the analysis of the websites, it was discovered that two of the websites both types of spam, resulting in the total number of times that spam was detected being 12, spread over 10 websites. Therefore, some of the websites contained more than one type of spam, and hence the discrepancy between the total number of websites with spam, and the occurrences of the types of spam identified (see Table 4.1). The number of occurrences of spam identified column indicates the websites that had spam but were listed by search engines. For example, Google has a total of nine listed websites with spam. However, number of occurrences of spam is 10. This is due to the fact that one of the websites had both keyword stuffing and page redirects (see Appendix C).

In the sections that follow the author has presented the analyses of these two spam practices. The implication of the results will be further discussed in the conclusion. Table 4.2 reflects the total percentage of websites that had spam.

TABLE 4.2: Percentage of websites with spam.

Total number of analysed websites	Total number of websites with spam	Percentage of websites with spam
47	10	21.3%

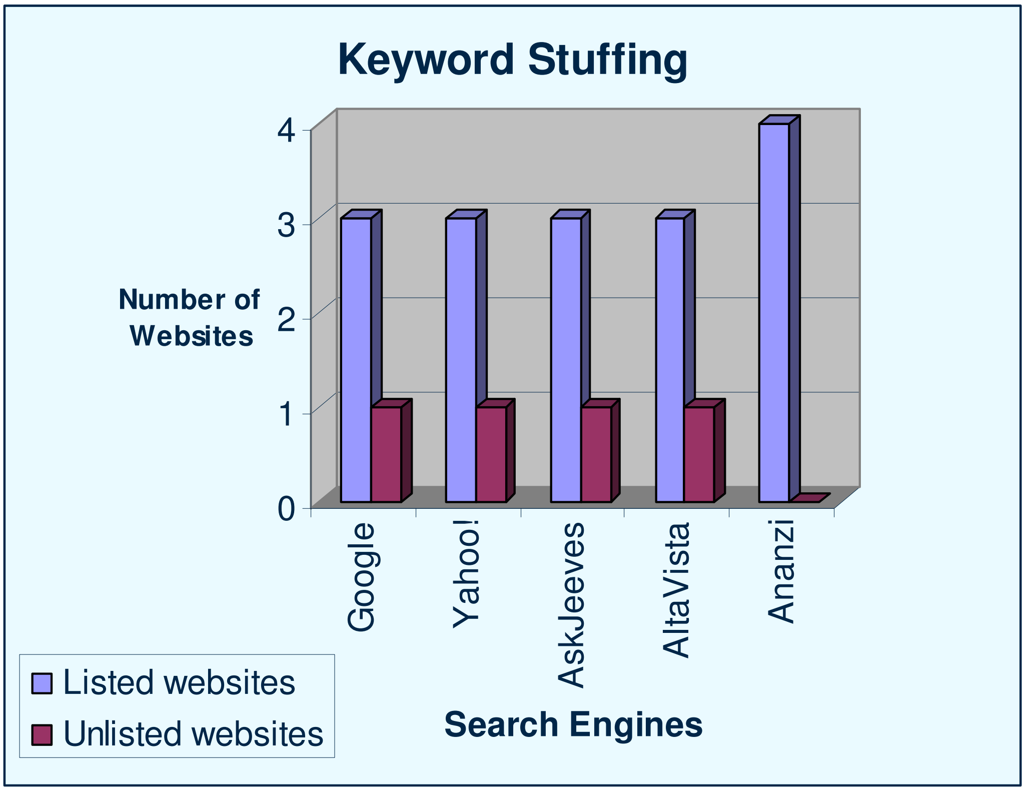
4.2.1 Keyword stuffing results

Although keyword stuffing was detected in some of the analysed websites (see Appendix C), it was not a major finding as compared to the other result. This is illustrated in Table 4.3 and Graph 4.2.

TABLE 4.3: Keyword stuffing results.

Search engine	Listed websites	Unlisted Websites
Google	3	1
Yahoo!	3	1
AskJeeves	3	1
Altavista	3	1
Ananzi	4	0

Table 4.3 is a reflection of how many of the websites contained keyword stuffing.



GRAPH 4.2: Keyword stuffing results

4.2.2 Keyword stuffing analysis

Keyword stuffing is considered to be an unethical SEO technique. However, if the keyword is repeated many times it will raise a red flag to the search engines and they will likely place a spam filter on the website or webpage (Thurow, 2003: 221; Wilkinson, 2004).

Graph 4.2 indicates that although keyword stuffing is considered as spam, all the search engines contravened their policies by listing websites that contained this type of spam. The graph also depicts websites that contained keyword stuffing which were not listed by the individual search engines. In evaluating the results, it is evident that although the five search engines state that webpages with keyword stuffing will not be indexed, some were indexed by the same search engines, indicating non compliance of the search engines to their policies.

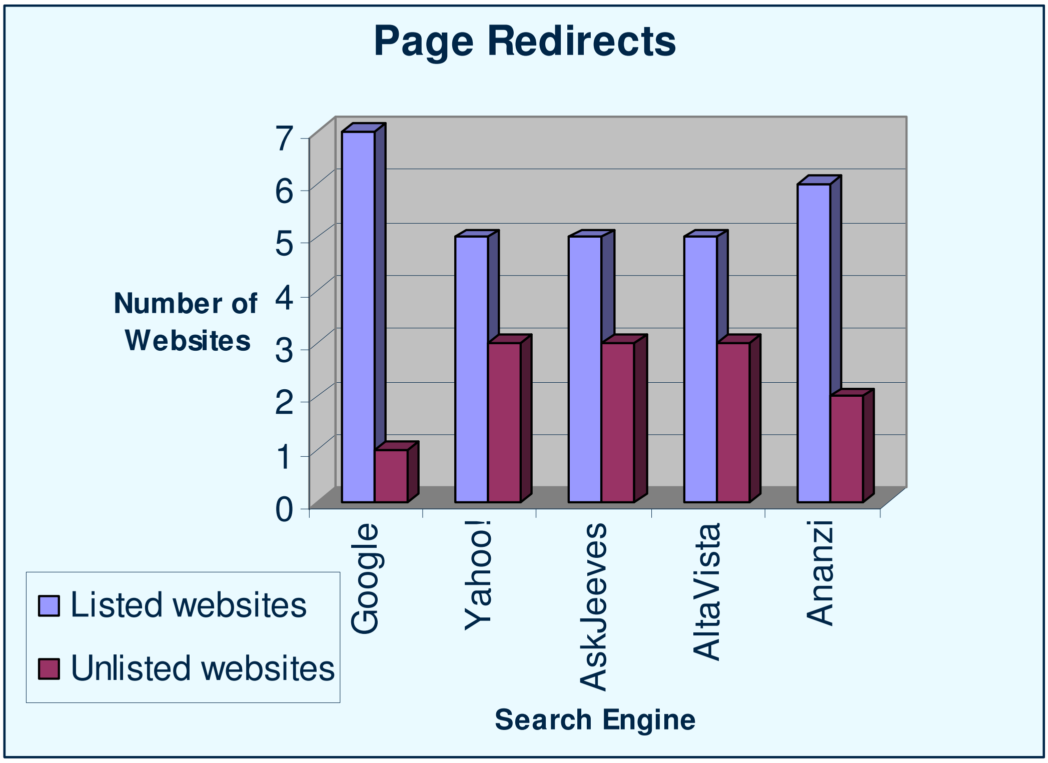
Although Ananzi had the highest number of indexed websites with keyword stuffing, some of the websites listed by Ananzi were not listed by the other search engines implying a possible exclusion of the websites by the search engines based on the presence of keyword stuffing.

4.2.3 Page redirects results

Table 4.2 indicates that out of the 47 websites analysed, 21.3% contained spam. Page redirects were detected in most of the analysed websites. This is reflected in Table 4.4 and Graph 4.3.

TABLE 4.4: Page redirects results.

Search engine	Listed websites	Unlisted Websites
Google	7	1
Yahoo!	5	3
AskJeeves	5	3
Altavista	5	3
Ananzi	6	2



GRAPH 4.3 Page redirects results

4.2.4 Page redirects analysis

According to Dunn (2004), not all redirects are considered as spam, however Wu and Davison (2005) are of the opinion that redirection is used to refer users to another URL and therefore constitutes search engine spam.

During the analysis of page redirects, Google had the highest number of websites with page redirects that had been indexed followed by Ananzi. This implied a possible exclusion of the websites by the other three search engines based on the presence of page redirects. As in the instance of keyword stuffing, the use of page redirects points to the non-compliance of search engines to their own policies.

4.2.5 Results of other spam in the analysis

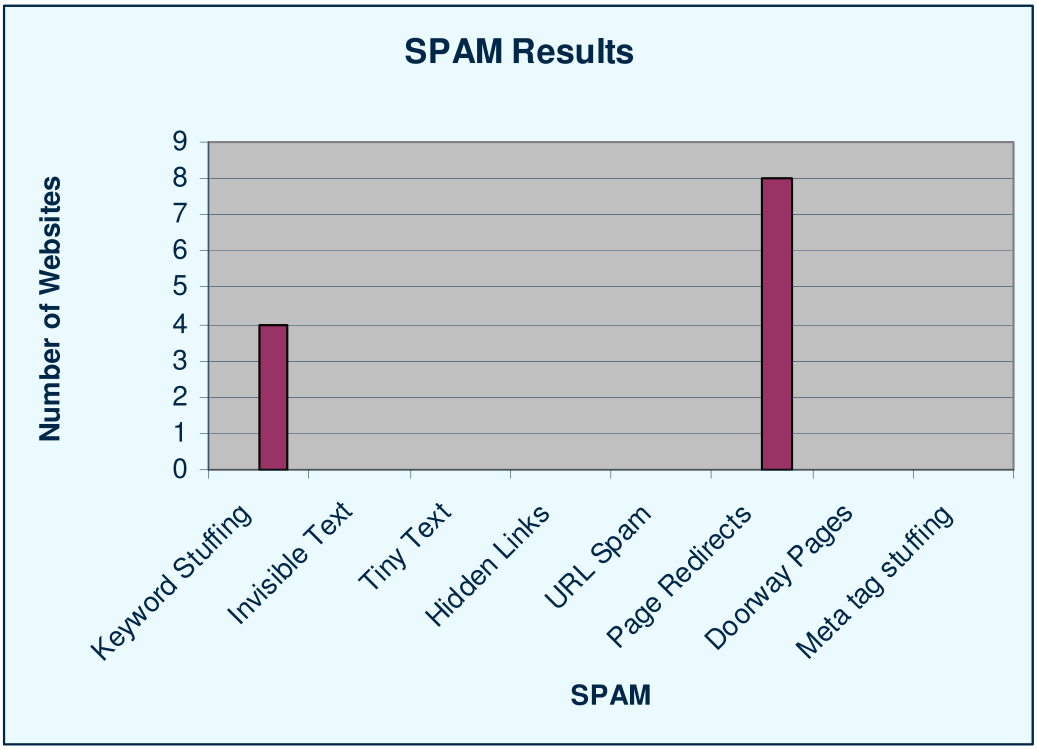
The following types of spam was not evident in the websites analysed:

- Invisible text.
- Tiny text.
- Hidden links.
- Artificial link farms.
- URL spam.
- Doorway pages.
- Meta-tag stuffing.

This could imply that search engines take the above mentioned spam techniques seriously and exclude all websites containing this type of spam. Web designers should refrain from applying these tactics when optimizing their websites. Overall spam results are reflected in Table 4.5.

TABLE 4.5: Overall spam results

Spam	Number of Websites
Keyword Stuffing	4
Invisible Text	0
Tiny Text	0
Hidden Links	0
URL Spam	0
Page Redirects	8
Doorway Pages	0
Meta-tag stuffing	0

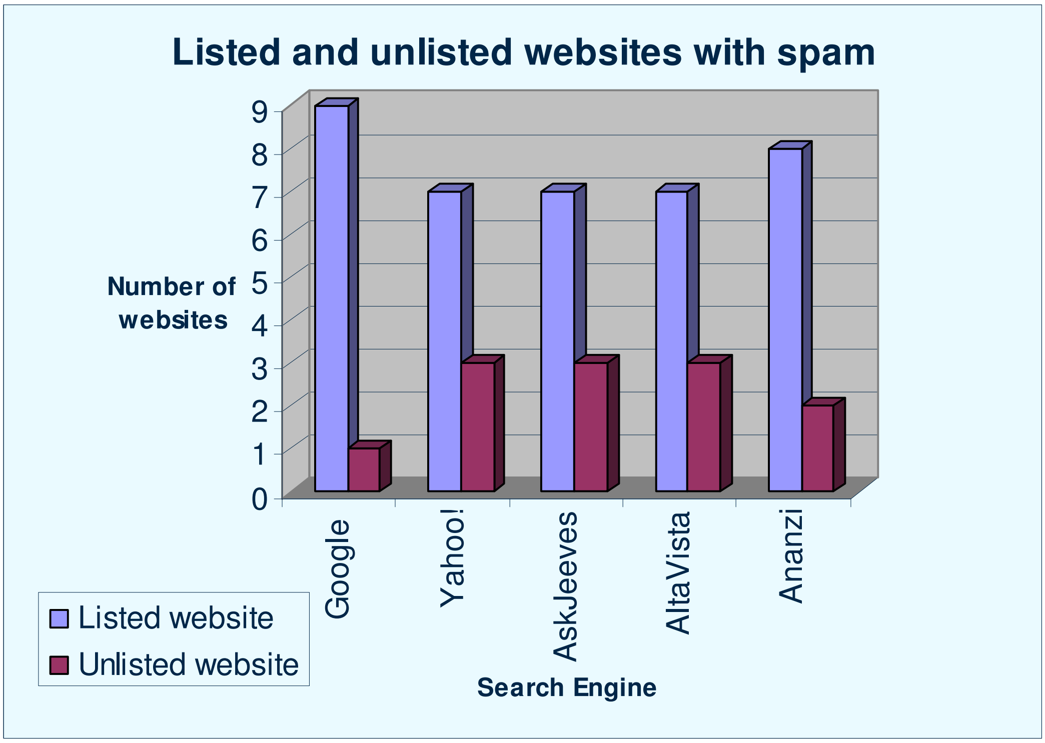


GRAPH 4.4: Overall spam results

Graph 4.4 indicates that of the eight different types of spam, only two were detected during the analysis of the websites. Table 4.6 reflects the overall number of websites that contained all types of identified spam but were listed by search engines, as well as those websites that contained spam and were not listed by the search engines.

TABLE 4.6: Spam listed and unlisted websites

Search engine	Listed websites	Unlisted websites
Google	9	1
Yahoo!	7	3
AskJeeves	7	3
AltaVista	7	3
Ananzi	8	2



GRAPH 4.5: Listed and unlisted websites with spam.

Graph 4.5 categorises the websites that contained spam and whether they were listed or not listed by search engines. It is evident that of the websites that were analysed, the search engines indexed more websites that contained spam than those that did not contain spam.

4.2.6 Result categories

The following categories have been identified as being representative of the potential conclusions:

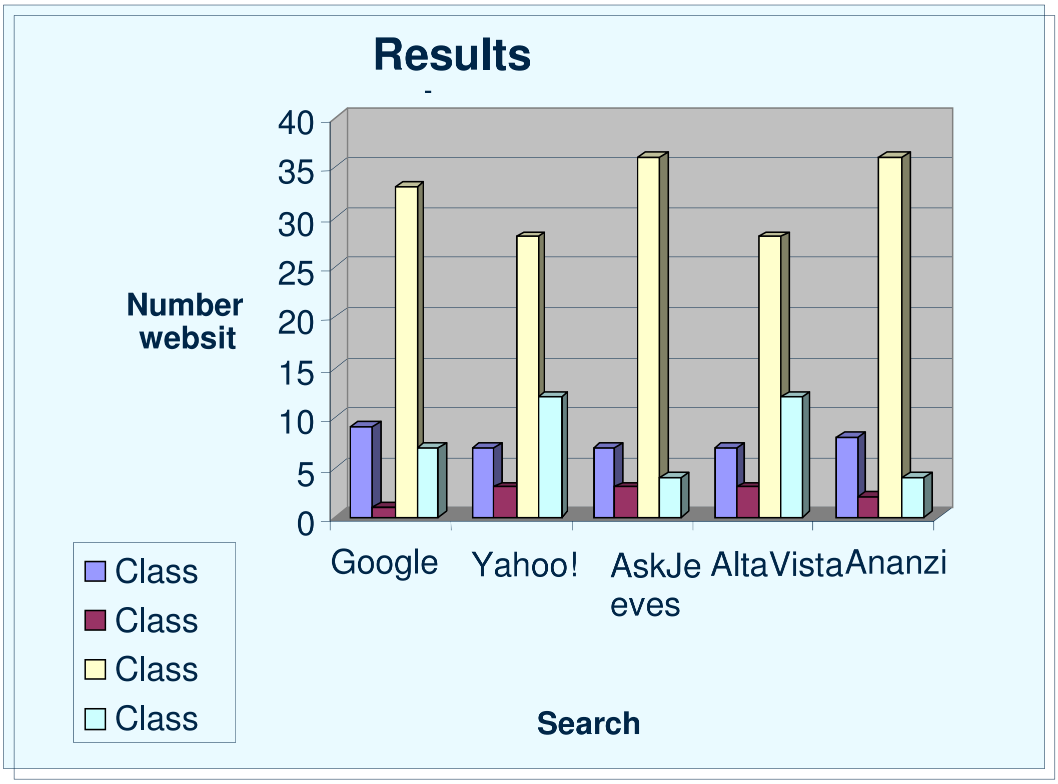
- **Class 1:** Websites containing what is considered as spam, but listed under one or more of the identified search engines.
- **Class 2:** Websites that contain spam and are not listed in any of the identified search engines.
- **Class 3:** Websites that do not contain any form of spam and are listed in one or more of the identified search engines.

- **Class 4:** Websites that do not contain spam, but are not listed in any of the identified search engines.

Table 4.7 and Graph 4.6 is an indication of how many websites were within which category.

Table 4.7: Results category table

Category	Google	Yahoo	AskJeeves	AltaVista	Ananzi
Class 1	9	7	7	7	8
Class 2	1	3	3	3	2
Class 3	33	28	36	28	36
Class 4	7	12	4	12	4



GRAPH 4.6: Results Category

4.3 Conclusion

From Graph 4.6, it is evident that class 2 websites had the least amount of spam occurrences. This class was a category of websites that contained spam and were not indexed by search engines. It was the author's expectation to find more websites falling within this category. However class 1, which was websites that contain spam and are listed under search engines, had more occurrences than class 2. The significant number of class 1 occurrences proves to an extent that search engines do not actually adhere to their own policies because they indexed websites that contained spam.

Class 3, which was websites that do not contain any form of spam, and are listed under one or more search engines, had the highest amount of occurrences. It proved the theory that using the correct SEO practices increases the chances of a website being indexed by search engines. However, another surprising outcome was class four, which was websites that did not contain spam, and were not listed. It was the author's expectation to find the least amount of websites falling within this category. However, reasons for this category could be on the part of the designer not submitting their website for indexing, not linking to other websites even though they do not contain any spam, or search engines not indexing some websites even though the correct SEO practices have been applied.

Graph 4.6 has indicated that in general, web designers do not apply spam tactics. The literature review had initially indicated that search engines take spam seriously, and even impose penalties on websites that contravene the policies. However, Graph 4.6 also indicates that search engines lack tight controls in instances where websites with spam were indexed. Ananzi and AskJeeves had the highest number of websites that did not contain spam and were listed.

CHAPTER 5
CONCLUSION

5.1 Introduction

This study aimed to investigate the search engine exclusion policies and the implications these policies have on indexing e-commerce websites. Only e-commerce websites were analyzed and as a result, the findings cannot be generalized to the listing of all other websites outside e-commerce. This chapter aims to reach a conclusion of the study, based on the literature review, the research conducted, and the results of the analysis.

5.2 Research Findings

As presented in Chapter 4, the results of this study were obtained from analysing 47 e-commerce websites, to determine if they contained any spam. Table 5.1 reflects search engine optimization practices that are perceived as spam. This list was compiled from the literature review as well as from analysing the policies of the five search engines that were used in this study.

TABLE 5.1: List of identified spam (**Sources:** AltaVista, 2005; Ananzi, 2005a; Anon, 2002; AskJeeves, 2005; Google, 2005c; Gikandi, 1999; Konia, 2002:312-316; Sullivan, 2003a & Thurow, 2003).

Spam list derived from literature and search engine policies	
1	Artificial link farms
2	Automated queries
3	Automatic submission
4	Cloaking
5	Domain spam
6	Doorway pages
7	Duplicate pages
8	Hidden links
9	Invisible text
10	Keyword stacking/stuffing
11	Metadata not describing the contents of webpage
12	Meta-tag stuffing
13	Page redirects
14	Page swapping
15	URL spam

As indicated in Chapter four, the analysis of the websites showed two findings. Firstly, there was a small number of websites that contained spam, but were listed in search engine results. Secondly, there were only two types of spam identified in the results. These types of spam were keyword stuffing and page redirects (see Appendix C; section 2.4.2.5; section 2.4.2.7). None of the other spam tactics that were used for this study were detected in the websites (see Appendix C).

5.3 Analysis of search engine policies

By comparing the literature review, the search engine policies as well as the data gleaned from the analysis phase, the following list is suggested to be acceptable and correct search engine optimization practices.

- Choose terms for the title that match the content of the document.
- Use description meta-tags and write descriptions accurately.
- Keep relevant text and links in HTML and not in graphics or images to ensure all the pages are crawled.

- Use ALT text for graphics.
- Build rich linkages between related pages.
- Use short and relevant page title to name pages.
- Have a site index to ensure that the entire site is indexed.

5.4 Limitations of the study

Although the author believes that the analysis was conducted in a professional manner there are some weaknesses of the study that need to be pointed out. Firstly, the author used purposive sampling. This could have introduced selection bias (only selecting websites that are not a representative sample).

Secondly, the results from some search engines could have been influenced by such engines that drives or powers them. This is particularly true in the case of AskJeeves that is powered by Yahoo!, meaning that the search results under AskJeeves might contain duplicates of results found under Yahoo!. Thirdly, there was no careful distinction between page redirects that are spam and those that are not. This could have resulted in overestimation of websites containing spam. Since this study was done at one point in time (cross sectional) and was not following website listing over time, it was not possible to determine whether websites containing spam would end up being banned sometime in the future.

5.5 Final Conclusion

From the research in this thesis, it is evident that spam can be applied to distort search engine results. As previously highlighted, web designers apply “unethical” SEO practices with the sole purpose of achieving high rankings in search engines. However, from the analysis, it was evident that a considerable number of e-commerce website designers are not making using of such spamming techniques. Of the e-commerce websites that were analysed, only 21.3% contained spam (see Table 4.2).

The author concludes that the research question which reads - *How do the exclusion policies of the five search engines impact the chances of indexing a website which contravenes the accepted SEO practices of the search engines?* – has been answered by this research. The implications of spam on indexing e-commerce websites have been determined. **Even though some websites that had spam were indexed, the overall number of websites that had spam was minimal. However, one other finding was that not all search engines registered all the websites found to contain spam (see Table 4.1).**

There seems to be some relaxation of rules by search engines when applying their exclusion policies, as experienced by websites that were indexed while they contained some form of spam (refer Table 4.1). Apart from influencing the change of policy and practice, it is not possible to explain why in a few cases some search engines, contrary to their policies, listed some websites with spam.

From the analysis of the websites, it is evident that most e-commerce web designers apply the correct practices when optimizing websites, and that search engines strictly adhere to the policies.

5.6 Future research

Future research could include investigating a wider selection of websites and not just e-commerce websites, as is the case in this research. More “unethical” SEO practices could be investigated with a larger sample. The results could be used to produce a list of SEO tactics that should be avoided and which should be applied to any website. It will be important to look at how search engines monitor their listed websites so as to continuously exclude those websites that contain spam, but were somehow listed by search engines, as is the case with some websites that were used in this study. As indicated by Wu and Davidson (2005), detecting search engine spam is a challenging research area.

BIBLIOGRAPHY

Adam, R. 2002. Is e-mail addictive? *Aslib Proceedings*, 54(2):85-94.

Alimohammadi, D. 2003. Meta-tag: a means to control the process of Web indexing. *Online Information Review*, 27(4):238-242.

Alimohammadi, D. 2004. Measurement of the presence of keywords and description meta-tags on a selected number of Iranian web sites.
Online Information Review, 28(3):220-223.

AltaVista. 2005. *Submit a site.*
http://www.altavista.com/addurl/default
[09 March 2005].

Amaratunga, D., Baldry, D., Sarshar, M. & Newton, R. 2002. Quantitative and qualitative research in the built environment: application of 'mixed' research approach. *Work Study,* 51(1): 17-31.

Ananzi. 2005a. *Add your site to Ananzi.*
http://search2.ananzi.co.za/Add_site/
[09 March 2005].

Ananzi. 2005b. *Ananzi home page.*
http://www.anazi.co.za
[12 September 2005].

Ananzi. 2005c. *Ananzi advertising and services.*
http://www.ananzi.co.za/comments/ratecard/rate_card.html
[09 March 2005].

Ananzi, 2005d. *Advanced search.*
http://search.ananzi.co.za/index.html?ql=a
[5 October 2005].

Anon, 2000. *Top site promote.*
http://www.topsitepromote.com
[4 September 2005].

Anon, 2001. *Get listed in the directories.*
http://www.123-search-engine-optimization.com/directories.html
[20 July 2005].

Anon, 2002. *SEO code of ethics.*
http://www.searchengineethics.com
[3 July 2005].

Anon, 2004a. *Black-hat search engine positioning tactics.*
http://www.beanstalk-inc.com/positioning-tactics/black-hat.htm
[23 August 2005].

Anon, 2004b. *How to suggest a site to the Open Directory.*
http://dmoz.org/add.html
[19 September 2005].

Anon, 2004c. *Report Spam Now.*
http://www.engine-spam/report-spam.html
20 September 2004].

Anon, 2005a. *Hidden Links and text.*
http://www.aim-pro.com/helpfiles/hiddenlinks.html
[07 September 2005].

Anon, 2005b. *Internet usage statistics - the big picture.*
World internet users and population stats.
http://www.internetworldstats.com/stats.htm
[19 August 2005].

AskJeeves. 2005. *Site submit service terms and conditions.*
http://www. ask.ineedhits.com/programterms.asp?n=u
[28 May 2005].

Barker, J. 2005. *Meta-search engines.*
http://www.lib.berkeley.edu/Teachinglib/Guides/Internet/metasearch.html
[03 September 2005].

Barnes, S.J. & Vidgen, R.T. 2002. An interactive approach to the assessment of e-commerce quality. *Journal of Electronic Commerce Research,* 3(3):114-127.

Boyes, J.A & Irani, Z. 2004. An analysis of the barriers and problems to web infrastructure development experienced by small businesses. *Information Technology Management,* 11(2):189-207.

Brightplanet 2005.
http://www.brightplanet.com/
[7 September 2005].

Brinkley, M. & Burke, M. 1995. Information retrieval from the internet: an evaluation of the tools. *Internet Research: Electronic Networking Applications and Policy,* 5(3):3-10.

Chambers, R. & Weideman, M. 2005. Search engine visibility: a pilot study towards the design of a model for e-commerce websites. *Proceedings of the 7th Annual Conference on World Wide Web Applications,* Cape Town, 29 -31 August 2005,http://www.uj.ac.za/www2005.

Collins, G. 2004. *Latest search engine spam techniques.*
http://www.sitepoint.com/print/search-enginespam-techniques
[12 September 2005].

Cooper, B. 2000. *Searching the Internet.* New York: Dorling Kindersley.

Cooper, D.R. & Schindler, P.S. 2003. *Business research methods.*8th ed. New York:McGraw-Hill Education Publishing.

Courtois, P.M. & Berry, M.W. 1999. Results ranking in web search engines. *Online Information Review*, 23(3)
http://www.onlineinc.com/onlinemag
[17 March 2005].

Cox, J. & Dale, B.G. 2002. Key quality factors in web site design and use: an examination. *International Journal of Quality & Reliability Management*, 19(7):862-888.

Createtraffic.net. 2001. *Doorway pages generator.*
http://createtraffic.net/tour.p?page=features/doorway
[6 September 2005].

Dahm, T. 2000. *Getting (and keeping) a top search engine ranking.*
http://www.webdevelopersjournal.com/articles/get_keep_top_ranking.html
[20 May 2005].

Darch, H. & Lucas, T. 2002. Training as an e-commerce enabler. *Journal of Workplace Learning*, 14(4):148-155.

Davidrajuh, R. 2003. Realizing a new e-commerce tool for formation of virtual enterprise. *Industrial Management & Data Systems,* 103(6):434-445.

Drott, M.C. 2002. Indexing aids at corporate websites: the use of robots.txt and Meta tags. *Information Processing and Management*, 38(2002):209-219.

Duffy, D.L. 2005. Affiliate marketing and its impact on e-commerce.
Journal of Consumer Marketing, 22(3):161-163.

Dunn, R. 2004. *The top 10 worst SEO tactics.*
www.stepforth.com
[18 September 2005].

Epstein, M. J. 2005. Implementing successful e-commerce initiatives. *Strategic Finance,* March:23-29.

Fetterly, D., Manasse, M. & Najork, M. 2004. Spam, damn spam, and statistics. Using statistical analysis to locate spam web pages. *Proceedings of the Seventh International Workshop on the Web and Databases,* Paris, 17-18 June.

Garofalakis, J., Kappos, P.& Makris, C. 2002. Improving the performance of web access by bridging global ranking with local page popularity. *Internet Research: Electronic Networking Applications Policy,* 12(1):43-54.

Gikandi, D. 1999. *Doorway pages go mainstream.*
http://www.webdevelopersjournal.com/articles/search_engines.html
[27 July 2005].

Goh, D.H. & Ang, R.P. 2003. Relevancy rankings: pay for performance search engines in the hot seat. *Online Information Review,* 27(2):87-93.

Google. 2004. *Technology overview.*
http://www.google.com/intl/en/corporate/tech.html
[13 July 2005].

Google. 2005a. *Google home page.*
http://www.google.co.za/
[24 August 2005].

Google. 2005b *Google information for webmasters: how do I get my site listed on Google?*
http://www.google.com/intl/en/webmasters/1.html
[24 August 2005].

Google. 2005c. *Google information for webmasters: my pages are currently not listed.*
http://www.google.com/intl/en/webmasters/2.html
[24 August 2005].

Google. 2005d. *Google information for webmasters: webmaster guidelines.*
http://www.google.co.za/webmasters/index.html
[01 March 2005].

Green, D. 2000. The evolution of web searching. *Online Information Review,* 24(2): 124-137.

Gyöngyi, Z. & Garcia-Molina, H. 2005. Web spam taxonomy.
Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web 2005, Chiba, Makuhari Messe, May 10, http://airweb.cse.lehigh.edu/2005/#proceedings.

Hart, T. & Rolletschek, G. 2003. The challenges of regulating the web. *Info,* 5(5):6-24.

Henning, E., Van Rensburg, W. & Smit B. 2004. *Finding your way in qualitative research.* Pretoria: Van Schaik Publishers.

Henzinger, M.K., Motwani, R. & Silverstein, C. 2002. *Challenges in web search engines.*
http://www.acm.org/sigir/forum/f2002/henzinger.pdf
[12 August 2005].

Hines, T. 2000. An evaluation of two qualitative methods (focus group interview and cognitive maps) for conducting research into entrepreneurial decision making. *Qualitative Market Research: An International Journal,* 3(1):7-16.

Hsieh, C. & Lin, B. 1998. Internet commerce for small businesses. *Industrial Management & Data Systems*, 98(3):113-119.

Jasco, P. 2005. Google scholar: the pros and the cons. *Online Information Review,* 29(2):208-214.

Joint, N. 2005. Aspects of Google: bigger is better – or more is less. *Library Review,* 54(3):145-148.

Kim, S., Shaw, T. & Schneider, H. 2003. Web site design benchmarking within industry groups. *Internet Research: Electronic Networking Applications Policy*, 13(1):17-26.

Kirkpatrick, C.H. 2002. Increase your website's search engine ranking. *Marketing Library Services,* 16(8).
http://infotoday.mondosearch.com
[13 September 2005].

Kline, V. 2002.Missing links: the quest for better search tools. *Online information review*, 26(4):252-255.

Konia, B.S. 2002. *Search engine optimization with WebPosition GOLD^{TM} 2.* Texas: Wordware Publishing.

Machill, M., Neuberger, C. & Schindler, F. 2003. Transparency on the net: functions and deficiencies of internet search engines. *Info - The Journal of Policy, Regulation and Strategy for Telecommunications,* 5(1):52-74.
http:/www.emeraldinsight.com/1463-6697.htm
[02 March 2005].

Marckini, F. 2000. *How to avoid trouble with the engines*.
http://www.inc.com/articles/2000/02/17232.html
[3 August 2005].

Marketleap.com. 2005.
http://www.marketleap.com/verify/defaault.htm
[12 September 2005].

McGuigan, G. 2003. Invisible business information: the selection of invisible websites in constructing subject pages for business. *Collection Building,* 22(2):68-74.

Moxley, D., Blake, J. & Maze, S. 2004. Web search engine advertising practices and their effect on library service. *The Bottom Line: Managing Library Finances,* 17(2): 61-65.

Näslund, D. 2002. Logistics needs qualitative research – especially action research. *International Journal of Physical Distribution & Logistics Management,* 32(5):321-338.

Nielsen, J. 2004a. *Statistics for traffic referred by search engines and navigation directories to Useit.*
http://www.useit.com/about/searchreferrals.html
[14 August 2005].

Nielsen, J. 2004b. *When search engines become answer engines.*
http://www.useit.com/alertbox/20040816.html
[14 August 2005].

Nielsen/Netratings. 2005. *Nielsen/Netratings releases top 10 search engine share rankings for July 2005.*
http://www.netratings.com/pr/pr_050824.pdf
[29 October 2005].

Nobles, R. & O'Neil, S. 2000. *Maximize web site traffic: build web site traffic fast and free by optimizing search engine placement*. Massachusetts: Adams Media Corporation.

Oppenheim, C., Morris, A., Mcknight, C. & Lowley, S. 2000. The evolution of WWW search engines. *Journal of Documentation*, 56(2):190-211.

Palumbo, F. & Herbig, P. 1998. International marketing tool: the Internet. *Industrial Management & Data systems,* 98(6): 253-261.

Peng, Y., Trappey, C.A. & Liu, N. 2005. Internet and e-commerce adoption by the Taiwan semiconductor industry. *Industrial Management & Data Systems,* 105(4):476-490.

Perkins, A. 2001. *The classification of search engine spam.* http://www.silverdisc.co.uk/articles/spam-classification/
[12 September 2005].

Podesta, G. 2000. E-commerce: helping customers gain the competitive edge. *Plastics Engineering,* 56(7):73-74.

Post , G.V. & Anderson, D.L. 2003. *Management information systems*: s*olving business problems with information technology*. New York: McGraw-Hill.

Poulter, A. 1997. The design of world wide web search engines: a critical review. *Program,* 31(2):131-145.

Remenyi, D. & Money, A. 2004. *Research supervision for supervisors and their students.* London : Academic Conferences.

Rowlett, D. 2003. *Stop search engine spam!*
http://www.internetmarketingwebsites.com/spam-review.htm
[12 June 2005].

Ru, Y. & Horowitz, E. 2005. Indexing the invisible web: a survey. *On~~n~~line Information Review,* 29(3): 249-265.

Saunders, Lewis & Thornhill. 1997. *Research methods for business students.* London: Pitman Publishing.

Sekhar, C. 2002. *Internet marketing and search engine positioning: a "do it yourself guide".* Tennessee: Southern Star Publishing.

Shenton, J. 2001. *Search engines explained: an overview of search engines and their use in promoting web sites.*
http://www.globalmillenniamarketing.com
[20 July 2005].

Sherman, C. & Price, G. 2002. *The invisible web: uncovering information sources search engines can't see.* New Jersey: Information Today, Inc.

Sherman, C. 2001. *Google unveils more of the invisible web.*
http://searchenginewatch.com/searchday/article.php/2158091
[12 July 2005].

Sherman, C. 2002. *Yahoo! birth of a new machine.*
http://searchenginewatch.com/searchday/article.php/3314171
[12 June 2005].

Simeon, R. 1999. Evaluating domestic and international website strategies. *Internet Research: Electronic Networking Applications and Policy,* 9(4):297-308.

Singh, A.M. 2002. The Internet - strategies for optimal utilization in South Africa. *South African Journal of Information Management*, 4(1).
www.sajim.co.za
[07 August 2005].

Struwig, F.W. & Stead, G.B. 2001. *Planning, designing and reporting research.* Pearson Education Publishing. South Africa.

Sullivan, D. 2001a. *Consumer group asks FTC to investigate search ads.* www.searchenginewatch.com//sereport/07/07-ftc.html [25 July 2005].

Sullivan, D. 2001b. *Desperately seeking search engine marketing standards.*
http://searchenginewatch.com/sereport/article.php/2164371
[5 September 2005].

Sullivan, D. 2002a. *Google bombs aren't so scary.*
www.searchenginewatch.com/sereport/print.php/34721_2164611
[3 September 2005].

Sullivan, D. 2002b. *How search engines work*.
www.searchenginewatch.com/webmasters.php/34751_2168031
[25 August 2005].

Sullivan, D. 2002c. *Intro to search engine optimization*.
www.searchenginewatch.com/webmasters/print.php/34751_2167921
[08 March 2005].

Sullivan, D. 2002d. *Search engine features for webmasters.*
www.searchenginewatch.com/webmasters/34751_2167891
[25 August 2005].

Sullivan, D. 2002e. *Search engine link popularity.*
www.searchenginewatch.com/searchday/print.php/34711_2159711
[3 September 2005].

Sullivan, D. 2003a. *Ending the debate over cloaking*.
www.searchenginewatch.com/sereport/print.php/34721_2165321
[08 March 2005].

Sullivan, D. 2003b. *How search engines rank web pages*.
www.searchenginewatch.com/webmasters/print.php/34751_2167961
[08 March 2005].

Sullivan, D. 2003c. *Searches per day.*
http://searchenginewatch.com/reports/article.php/2156461
[26 July 2005].

Sullivan, D. 2004a. *Buying your way in: search engine advertising chart.*
http://www.searchenginewatch.com/webmasters/print.php/34751_2167941
[9 August 2005].

Sullivan, D. 2004b. *Google tops, but Yahoo switch success so far.*
http://searchenginewatch.com/searchday/article.php/3334881
[7 September 2005].

Sullivan, D. 2004c. *Major search engine and directories.*
http://searchenginewatch.com/links/article.php/2156221
[28 May 2005].

Sullivan, D. 2004d. *Search engine results chart.*
http://searchenginewatch.com/webmasters/article.php/34751_2167981
[15 August 2005].

Sullivan, D. 2004e. *Search engine size wars erupts.*
http://blog.searchenginewatch.com/blog/041111-084221
[28 May 2005].

Sullivan, D. 2004f. *Spam rules require effective spam police.*
http://www.clickz.com/experts/search/opt/article.php/3348681
[10 November 2005].

Sullivan, D. 2004g. *Submitting to crawlers: Google, Yahoo, Ask/Teoma & Microsoft's MSN.*
http://searchenginewatch.com/webmasters/print.php/34751_2167871
[28 June 2005].

Sullivan, D. 2004h. *Submitting to directories: Yahoo & the open directory.*
http://searchenginewatch.com/webmasters/print.php/34751_2167881
[6 August 2005].

Sullivan, D. 2005. *Hitwise search engine ratings.*
http://searchenginewatch.com/reports/34701_3099931
[23 August 2005].

The Endless Links Page Company. 2005. *Welcome to my FFA links page.*
http://www.free-for-all-links-page.com
[12 September 2005].

Thelwall, M. 2000a. Commercial web sites: lost in cyber space? *Internet Research: Electronic Networking Applications and Policy*, 10(2):150-159.

Thelwall, M. 2000b. Effective websites for small and medium-sized enterprises. *Journal of Small Business and Enterprise Development,* 7(2): 149-159.

Thelwall, M. 2001. Commercial Web site links. *Internet Research: Electronic Networking Applications and Policy*, 11(2):114-124.

Thelwall, M. 2002a. Methodologies for crawler based web surveys. *Internet Research: Electronic Networking Applications and Policy*, 12(2):124-138.

Thelwall, M. 2002b. Subject gateway sites and search engine ranking. *Online information Review*, 26(2):101-107.

Thelwall, M. & Vaughan, L. 2004. New versions of PageRank employing alternative web document models. *ASLIB Proceedings,* 56(1):24-33.

Thurow, S. 2003. *Search engine visibility.* Indianapolis: New Riders Publishing.

Thurow, S. 2004a. *Doorway pages are bad*.
http://www.searchenginesbook.com/presskit.html
[13 March 2005].

Thurow, S. 2004b. *How to spot Search engine Spam: Doorway pages.*
http://www.clickz.com/experts/search/results/print.php/3325301
[03 August 2005].

Thurow, S. 2004c. *Keyword Repetition for search engine optimization*.
http://www.webpronews.com
[10 November 2005].

Van der Westhuizen, M. 2001. The invisible web. *South African Journal of Information Management,* 3(3/4). http:www.sajim.co.za
[24 August 2005].

Van der Walt, P.W. 1998. *Task analysis of the webmaster*. Unpublished, Rand Afrikaans University, Johannesburg (MI thesis).

Van Steenderen, M. 2001. Website management: making a web site more visible. *South African Journal of Information Management,* 2(4). http://www.sajim.co.za
[21 September 2005].

Vaughan, J. 1999. Considerations in the choice of an internet search tool. *Library Hi Tech,* 17(1): 89-106.

Wallace, D. 2003. *Spamming techniques that you will want to avoid.*
http://www.searchrank.com/resources/art003.htm
[10 November 2005].

Weideman, M. 2004. Ethical issues on content distribution to digital consumers via paid placement as opposed to website visibility in search engine results. *Proceedings of the Seventh International Conference ETHICOMP 2004,* Syros, University of the Aegean:904-915, 14-16 April 2004.

Wen, J.H., Chen, H. & Hwang, H. 2001. E-commerce website design : strategies and models. *Information Management & Computer Security*, 9(1):5-12.

Wikipedia.com. 2005a. *Search engine*.
http://en.wikipedia.org/wiki/search_engine [29 September 2005].

Wikipedia.com. 2005b. *Search engine optimization*.
http://en.wikepedia.org/wiki/serach_engine_optimization
[29 September 2005].

Wikipedia.com. 2005c. *Spamdexing.*
http://en.wikipedia.org/wiki/Spamdexing
[29 September 2005].

Wilkinson, T.A. 2004. *Just say no to SEO spam*.
www.w-edge.com
[23 August 2005].

Wilson, M. 2000. The development of the internet in South Africa. *Telematics and Informatics*, 16:99-111.

Wilson, K.C. 2002. *Automatic indexing : problems and solutions.* http://www.humbul.ac.uk/ltsn-humbul/survey/survey_appendix8.doc
[23 September 2005].

World Wide Worx. 2002. *The Goldstuck report: online retail in South Africa.*
http://www.theworx.biz/retail02.htm
[23 September 2005].

Wu, B. & Davidson, B.D. 2005. Cloaking and Redirection: A Preliminary Study. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web 2005,* Chiba, Makuhari Messe, May 10
http://airweb.cse.lehigh.edu/2005/#proceedings.

Yahoo! 2005a. *Avoiding search engine spam*.
http://smallbusiness.yahoo.com/resources/article.php?mcid=6&scid=35&aid=2731
[3 August 2005].

Yahoo! 2005b. *Search ranking help.*
http://help.yahoo.com/help/us/ysearch/ranking/index.html
[3 August 2005].

Yates, R. 2005. Web site accessibility and usability: towards more functional sites for all. *Campus Wide Information Systems*, 22(4):180-188.

Zhang, J. & Cheung, C. 2003. Meta-search-engine feature analysis. *Online Information Review,* 27(6):433-441.

Zhang, I. & Dimitroff, A. 2004. The impact of webpage content characteristics on webpage visibility in search engines (Part I). *Information Processing & Management,* 41(2005): 665-690
http://jis.sagepub.com/cgi/content/refs/30/4/310
[12 February 2005].

APPENDIX A
E-COMMERCE WEBSITES

Company Name	URL
Boland Wingerde Internasionaal Beperk	http://www.bolandwines.co.za
Gifts International CC	http://www.giftsint.co.za
Lanzerac Wyne	http://www.lanzeracwines.co.za
Oodles Of Doodels	http://www.oodlesofdoodles.co.za/
Hush-a-byes	www.hush-a-byes.co.za
Kalahari.net	www.Kalahari.net
Street car	www.streetcar.com
Aardvark Press	www.aardvarkpress.co.za
InterSoft	www.intersoft.co.za
Leisure books	http://www.leisurebooks.com/
Fleet Street Publications	www.fsp.co.za
IBS Premium Books	www.ibs.co.za
Loot	www.loot.co.za
Take 2	www.take2.co.za
NGR Computers	www.ngrcomputers.co.za
Digital Planet	www.digitalplanet.co.za
S A Camera	www.sacamera.co.za
PCShopping	www.pcshopping.co.za
Techdigital - South Africa	www.techdigital.co.za
Cybertrek	www.cybertrek.co.za
Image Technology	www.shopping.imagecorp.co.za
Digital World	www.wholesaledigital.co.za
Jump Shopping	www.jump.co.za
My New Laptop	www.mynewlaptop.co.za
Renegade Xpress Online	www.renegadexpress.co.za
Soholink.co.za	www.soholink.co.za
Easy Online	www.easyonline.co.za
Buy Computers	www.buycomputers.co.za
Sybaritic	www.sybaritic.co.za/store
Trillion Computers - Online Computer Everything	www.trillioncomputers.co.za
DVD4Africa	www.dvd4africa.com
ShopASave	www.shopasave.co.za
Enigmatek Online	www.enigmatek.co.za
e-Wise Knowledge Library	www.ewklibrary.com
WildlifeCampus	www.wildlifecampus.com
Handsfree.co.za	www.handsfree.co.za
Bss Online	www.bssonline.co.za
Zakspeed Electronics online	www.zakspeed.co.za
Netflorist	www.netflorist.co.za
Flowers.co.za	www.flowers.co.za
Giftwrap	www.giftwrap.co.za
iFlora Web Florist	www.iflora.co.za
Egames	www.egames.co.za

SoSimple	www.sosimple.co.za
Board Games	www.boardgames.co.za
Great Deal Shopping	www.greatdeal.co.za
SARugby	www.sarugby.com
Experience Gifts Online Store	www.egos.co.za
Fragrance	www.fragrance.co.za
HealthSpas	www.healthspa.co.za
Perkal Gifts	www.perkalgifts.co.za
Bossem	www.bossem.co.za
Prohampers	www.prohampers.co.za
Giftware	www.giftware.co.za
The Gift Lady	www.giftlady.net
GOURMET GODDESS	www.gourmetgoddess.co.za
Gift Basket South Africa - What's in the basket?	www.giftbasketsa.co.za
Skin Care Shop	www.skincareshop.co.za
Herbalife South Africa	www.healthier.co.za
Suncore Health Products	www.soladey.co.za
Ascot Direct	www.ascotdirect.co.za
Perfume Direct	www.perfumedirect.co.za
Jose Jewels	www.josejewels.co.za
Something Sexy	www.somethingsexy.co.za
Media Exchange	www.mediaexchange.co.za
One World	www.oneworld.co.za
RubberstampSA.co.za	www.rubberstampsa.co.za
Pharmacy4u	www.pharmacy4u.co.za
Woolworth	www.woolworths.co.za
Pick n pay	www.picknpay.co.za
Ewine	www.ewine.co.za
Cyber Cellar	www.cybercellar.co.za
ActionWeb	www.actionweb.co.za
African Online Shop	www.over2u.com
Beds-on-line	www.beds-on-line.co.za
Biltong2u	www.biltong2u.co.uk
Books24	www.books24.co.za
CaCell	www.cacell.co.za
CCTV toolbox	www.cctvtoolbox.co.za
CD companion	www.cd-companion.co.za
Cellphones Direct	www.cellphones.co.za
Cheaper	www.cheaper.co.za
Cigars	www.cigars.co.za
Coffee.co.za	www.coffee.co.za
Jamo	www.jamo.co.za
Digital World	www.wholesaledigital.co.za
Soundz	www.soundz.co.za
Drugsure	www.drugsure.com
Eagle applications	www.eagleapplications.co.za
EnergyZone	www.energyzone.co.za
Exclusive books	www.exclusivebooks.com
Foto Digital	www.fotodigital.co.za

Gadgets house	www.gadgetshouse.co.za
Glomail	www.glomail.co.za
Goods 4 U delivery	www.good4u.com
Hammer on Guitars	www.hammeronguitars.co.za
Have2have.co.za	www.have2have.co.za
Homemark	www.homemark.co.za
Impact video online	www.impactvideo.co.za
Linen drawer	www.linendrawer.co.za
Luggage warehouse	www.luggagewarehouse.co.za
Luipini guys den	www.luipiniguysden.co.za
Magafters	www.magafters.com
mecer direct online store	www.mecerdirect.co.za
Mr Mattress online	www.mrmattress.co.za
Nuts are us	www.thenutshop.co.za
Orions belt	www.orionsbelt.co.za
Parties 4 africa	www.parties4africa.co.za
Nuweb	www.nuweb.co.za
Bath biltong	www.bathbiltong.co.uk
Premium wines	www.premiumwines.co.za
Pro digital	www.prodigital.co.za
ProGPS	www.progps.co.za
Redlorry	www.redlorry.co.za
Rxnam.com	www.rxnam.com
SABshop	www.sabshop.com
Saleys travel goods	www.stgbags.co.za
Satooz	www.satooz.com
SOHO online shopping	www.soho.co.za
South quay import-export	www.impex.co.za
Sun-e-shop	www.sun-e-shop.co.za
Team101	www.team101.co.za
Techno toys	www.technotoys.co.za
The gadget shop	www.thegadgetshop.co.za
Paintball craze store	www.paintballcraze.co.za
The south african food shop	www.southafricanfoodshop.com
Timeslice	www.timeslice.co.za
Toys for boys	www.toysforboys.co.za
Toyworld	www.toyworld.co.za
Under african sky	www.underanafricansky.co.za
Vedic books	www.vedicbooks.net
Vet products online	www.vetproductsonline.co.za
Volpes	www.volpes.co.za
Waltons	www.waltons.co.za
Yebo electronics	www.fort777.co.za
Musica	www.musica.co.za

APPENDIX B
WEBSITES AFTER RANDOM SELECTION

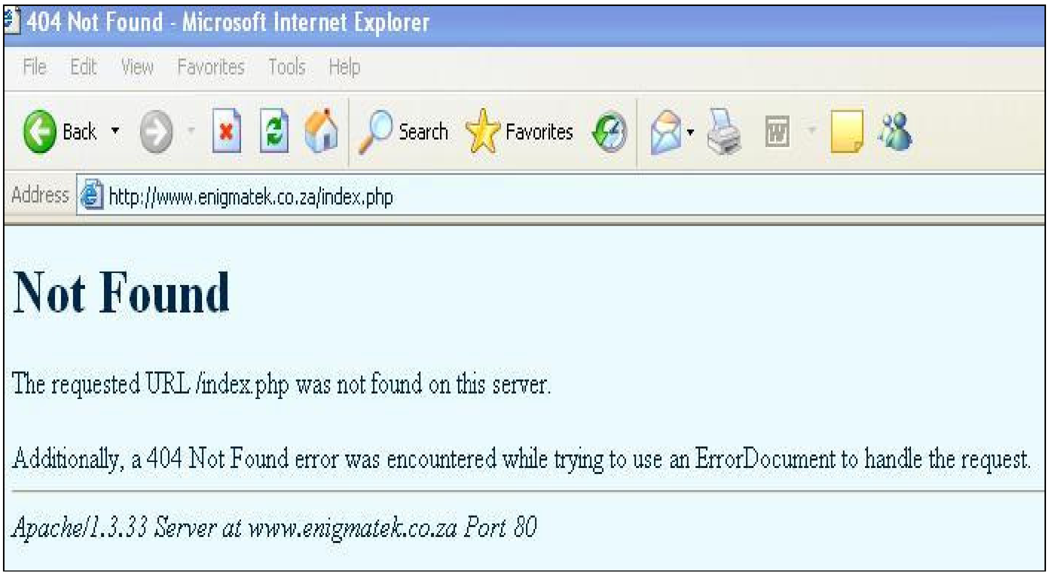
	Random Number	Google	Yahoo	AskJeeves	Altavista	Ananzi
www.pcshopping.co.za	0.529810049	Y	N	Y	N	Y
www.ngrcomputers.co.za	0.590298186	Y	Y	Y	Y	Y
www.magafters.com	0.99392923	Y	Y	Y	Y	Y
www.sarugby.com	0.764054842	Y	Y	Y	Y	Y
www.prohampers.co.za	0.7634205	Y	Y	Y	Y	Y
www.intersoft.co.za	0.114912137	N	N	N	N	Y
www.orionsbelt.co.za	0.139283534	Y	N	Y	N	Y
www.mrmattress.co.za	0.934011603	Y	Y	Y	Y	Y
www.soholink.co.za	0.781562232	Y	N	Y	N	N
www.boardgames.co.za	0.062169864	N	N	Y	N	Y
www.gadgetshouse.co.za	0.544735856	Y	Y	N	Y	Y
www.enigmatek.co.za	0.433696508	Y	N	Y	N	Y
www.sabshop.com	0.238287585	Y	Y	N	Y	Y
www.coffee.co.za	0.60258904	N	N	Y	N	Y
www.energyzone.co.za	0.53613596	Y	Y	Y	Y	Y
www.netflorist.co.za	0.894239366	Y	Y	Y	Y	Y
www.volpes.co.za	0.260735456	Y	Y	Y	Y	Y
www.egos.co.za	0.28816306	Y	Y	Y	Y	Y
www.mecerdirect.co.za	0.097452076	Y	Y	Y	Y	Y
www.over2u.com	0.92881439	Y	Y	Y	Y	Y
www.waltons.co.za	0.173680632	Y	Y	Y	Y	Y
www.giftlady.net	0.439258971	Y	Y	Y	Y	Y
www.loot.co.za	0.591205369	N	Y	Y	Y	Y
www.good4u.com	0.496423977	Y	Y	Y	Y	Y
www.timeslice.co.za	0.684989828	Y	Y	Y	Y	Y
www.giftware.co.za	0.696107694	N	N	Y	N	Y
www.jump.co.za	0.660878198	Y	Y	Y	Y	N
www.thegadgetshop.co.za	0.655579834	Y	Y	N	Y	Y
www.musica.co.za	0.614544777	Y	Y	Y	Y	Y
www.prodigital.co.za	0.803625213	Y	N	Y	N	Y
www.streetcar.com	0.603359362	Y	Y	Y	Y	Y
www.shopping.imagecorp.co.za	0.583992672	Y	N	Y	N	N
www.rubberstampsa.co.za	0.400811798	Y	N	Y	N	Y
www.premiumwines.co.za	0.647203898	Y	N	N	N	Y
www.healthspa.co.za	0.478814365	Y	N	Y	N	Y
www.team101.co.za	0.306309692	Y	Y	Y	Y	Y
www.progps.co.za	0.942887179	Y	Y	Y	Y	Y
www.wholesaledigital.co.za	0.855486362	Y	Y	Y	Y	Y
www.hammeronguitars.co.za	0.542882004	Y	Y	N	Y	Y
www.underanafricansky.co.za	0.879475772	N	Y	N	Y	N
www.cacell.co.za	0.483114566	Y	Y	Y	Y	Y
www.fotodigital.co.za	0.543047733	Y	Y	Y	Y	Y
www.southafricanfoodshop.com	0.017139978	Y	Y	Y	Y	Y
www.mynewlaptop.co.za	0.028897853	Y	Y	Y	Y	Y
www.sun-e-shop.co.za	0.387111045	Y	Y	Y	Y	Y
www.ascotdirect.co.za	0.567882678	Y	Y	Y	Y	N
www.healthier.co.za	0.564540938	N	N	Y	N	Y
www.cheaper.co.za	0.772055108	N	N	Y	N	Y
www.shopasave.co.za	0.43198867	Y	Y	Y	Y	Y
www.homemark.co.za	0.104288836	Y	Y	Y	Y	Y
www.josejewels.co.za	0.619885242	Y	Y	Y	Y	N

APPENDIX C
SPAM INDICATION

URL	Keyword stuffing	Invisible text	Tiny text	Hidden links	URL SPAM	Page redirects	Doorway pages	Meta tag stuffing
www.pcshopping.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.ngrcomputers.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.magafters.com	NO	NO	NO	NO	NO	NO	NO	NO
www.sarugby.com	NO	NO	NO	NO	NO	Yes	NO	NO
www.prohampers.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.intersoft.co.za	Yes – Image 11	NO	NO	NO	NO	Yes	NO	NO
www.orionsbelt.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.mrmattress.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.soholink.co.za	NO	NO	NO	NO	NO	Yes	NO	NO
www.boardgames.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.gadgetshouse.co.za	NO	NO	NO	NO	NO	Yes	NO	NO
www.enigmatek.co.za		NO	NO	NO				
www.sabshop.com	NO	NO	NO	NO	NO	NO	NO	NO
www.coffee.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.energyzone.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.netflorist.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.volpes.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.egos.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.mecerdirect.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.over2u.com	NO	NO	NO	NO	NO	NO	NO	NO
www.waltons.co.za	NO	NO	NO	NO	NO	Yes	NO	NO
www.giftlady.net	Yes – image 33	NO	NO	NO	NO	NO	NO	NO
www.loot.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.good4u.com	NO	NO	NO	NO	NO	NO	NO	NO
www.timeslice.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.giftware.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.jump.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.thegadgetshop.co.za	No	NO	NO	NO	NO	NO	NO	NO
www.musica.co.za	NO	NO	NO	NO	NO	Yes	NO	NO
www.prodigital.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.streetcar.com	NO	NO	NO	NO	NO	NO	NO	NO
www.shopping.imagecorp.co.za	NO	NO	NO	NO	NO	Yes	NO	NO
www.rubberstampsa.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.premiumwines.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.healthspa.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.team101.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.progps.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.wholesaledigital.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.hammeronguitars.co.za	NO	NO	NO	NO	NO	Yes	NO	NO

www.underanafricansky.co.za								
www.cacell.co.za	Yes- image 50	NO	NO	NO	NO	NO	NO	NO
www.fotodigital.co.za	Yes- image 51	NO	NO	NO	NO	Yes	NO	NO
www.southafricanfoodshop.com	NO	NO	NO	NO	NO	NO	NO	NO
www.mynewlaptop.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.sun-e-shop.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.ascotdirect.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.healthier.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.cheaper.co.za								
www.shopasave.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.homemark.co.za	NO	NO	NO	NO	NO	NO	NO	NO
www.josejewels.co.za								

APPENDIX D
SUSPENDED WEBSITES

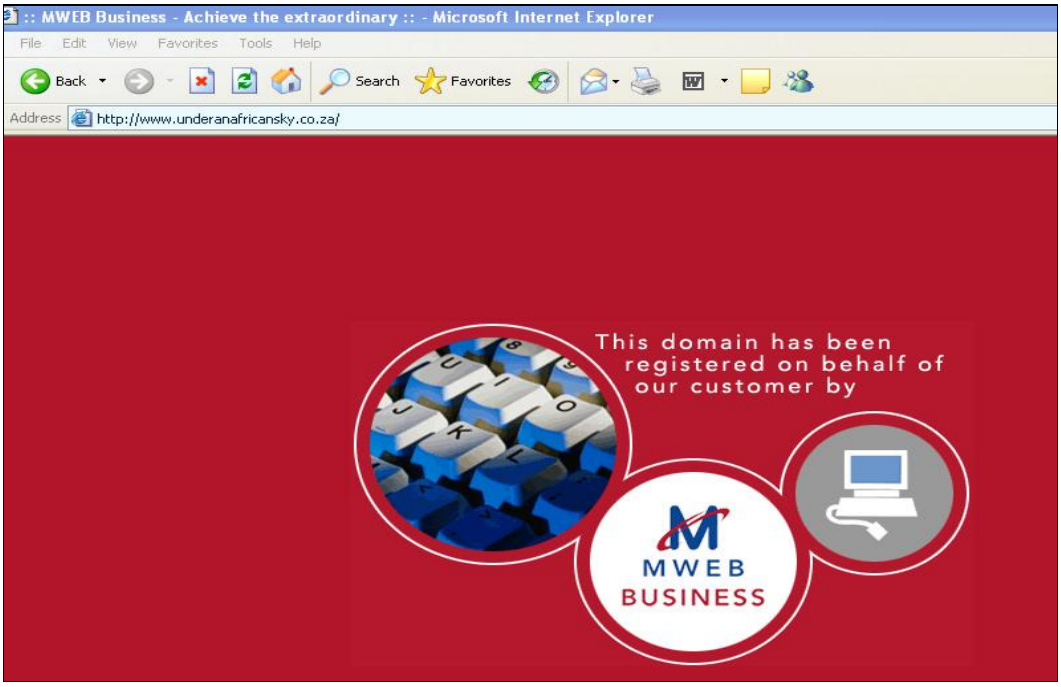

404 Not Found - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.enigmatek.co.za/index.php

Not Found

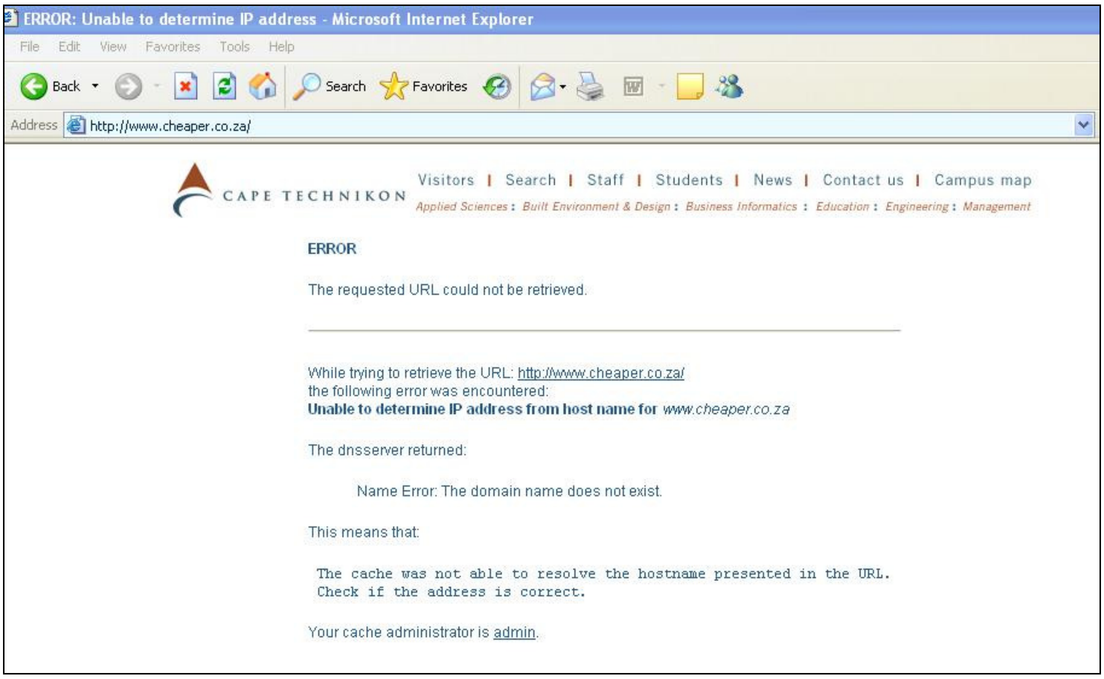
The requested URL /index.php was not found on this server.

Additionally, a 404 Not Found error was encountered while trying to use an ErrorDocument to handle the request.

Apache/1.3.33 Server at www.enigmatek.co.za Port 80


:: MWEB Business - Achieve the extraordinary :: - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.underanafricansky.co.za/
MWEB
BUSINESS

This domain has been registered on behalf of our customer by


ERROR: Unable to determine IP address - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.cheaper.co.za/
CAPE TECHNIKON
Visitors | Search | Staff | Students | News | Contact us | Campus map
Applied Sciences : Built Environment & Design : Business Informatics : Education : Engineering : Management

ERROR

The requested URL could not be retrieved.

While trying to retrieve the URL: http://www.cheaper.co.za/
the following error was encountered:
Unable to determine IP address from host name for *www.cheaper.co.za*

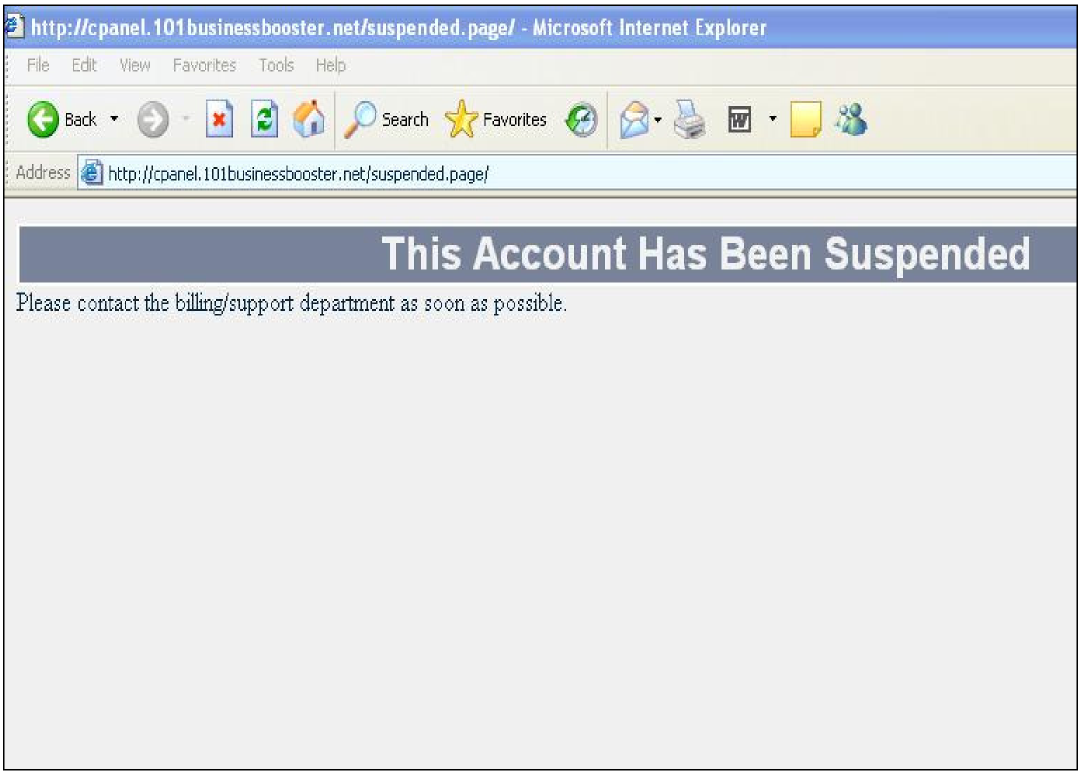
The dnsserver returned:

Name Error: The domain name does not exist.

This means that:

```
The cache was not able to resolve the hostname presented in the URL.
Check if the address is correct.
```

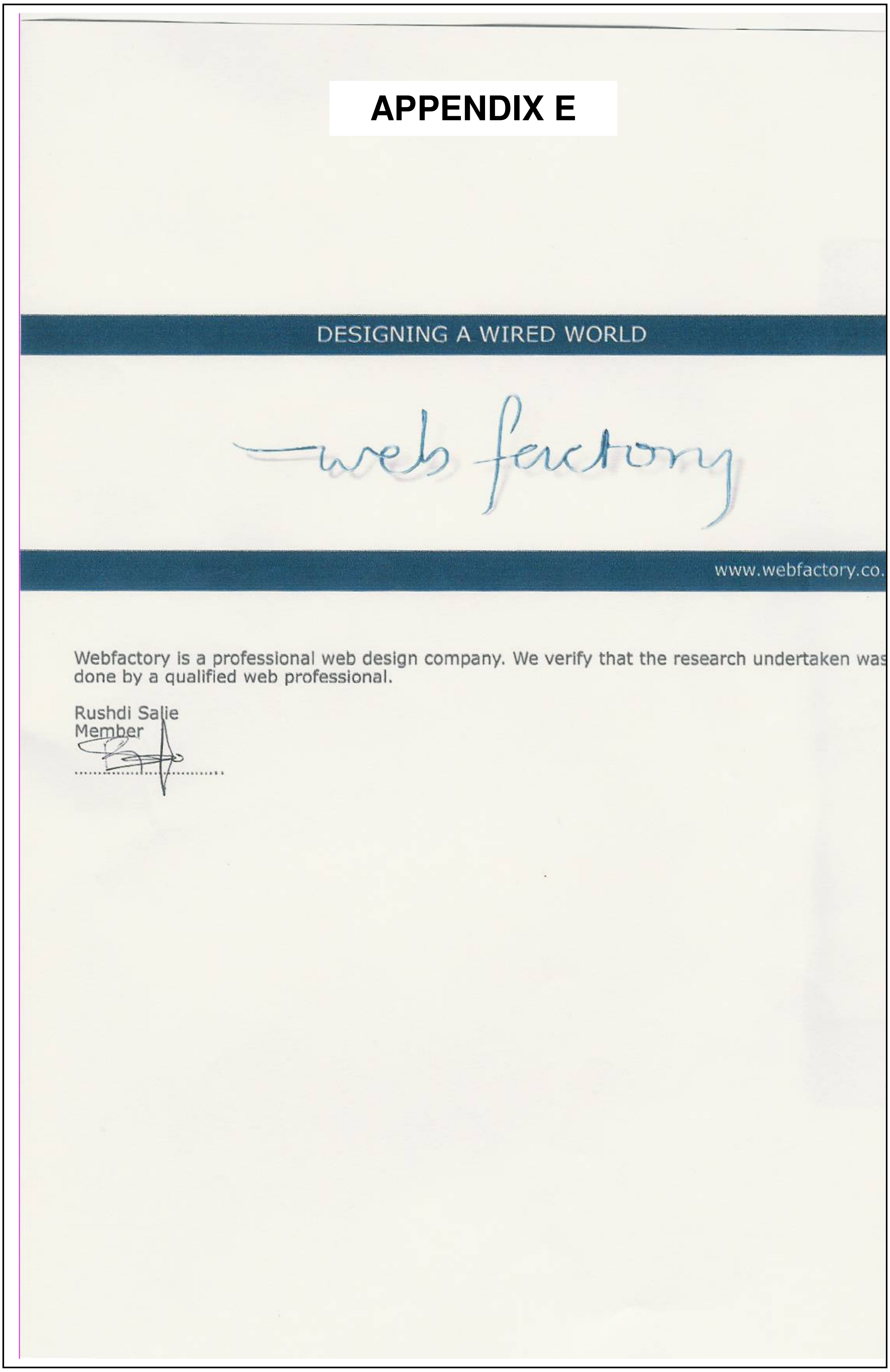
Your cache administrator is admin.


http://cpanel.101businessbooster.net/suspended.page/ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://cpanel.101businessbooster.net/suspended.page/

This Account Has Been Suspended

Please contact the billing/support department as soon as possible.

APPENDIX E



DESIGNING A WIRED WORLD

web factory

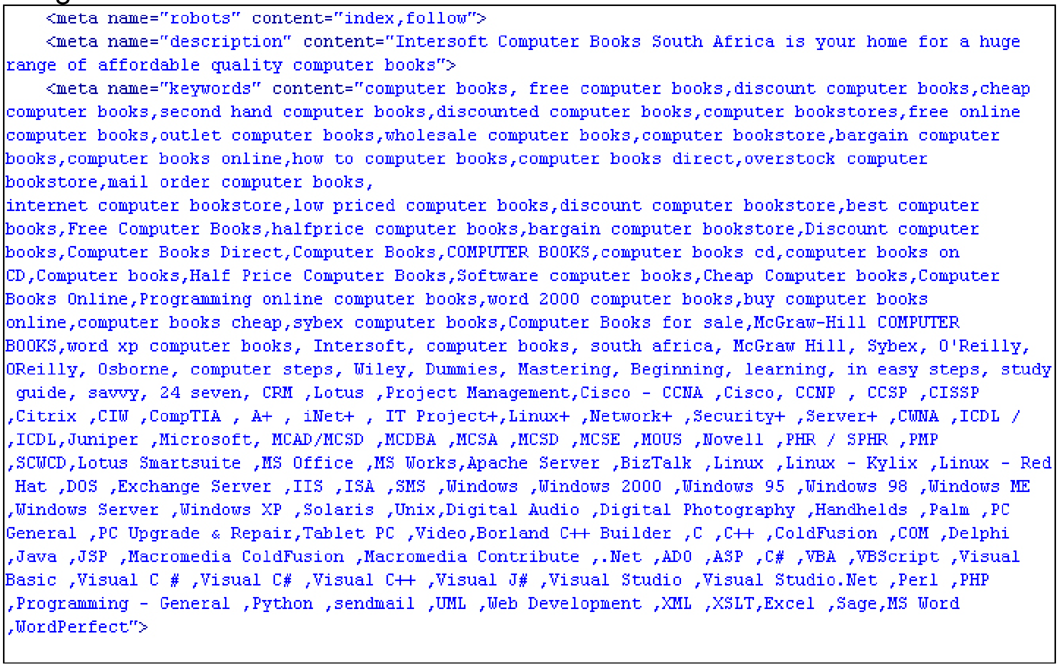
www.webfactory.co.

Webfactory is a professional web design company. We verify that the research undertaken was done by a qualified web professional.

Rushdi Salie
Member
....

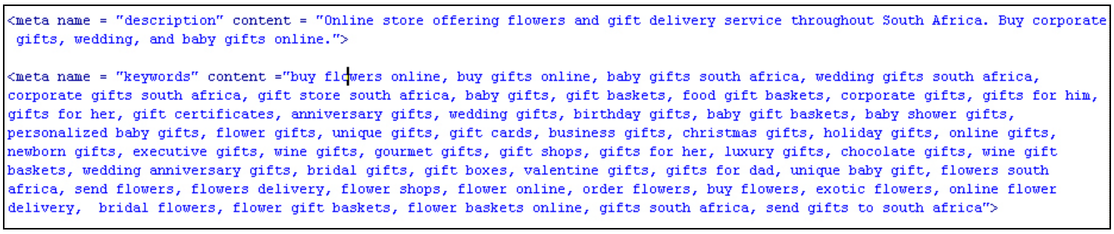
APPENDIX F
KEYWORD STUFFING IMAGES

Image 11



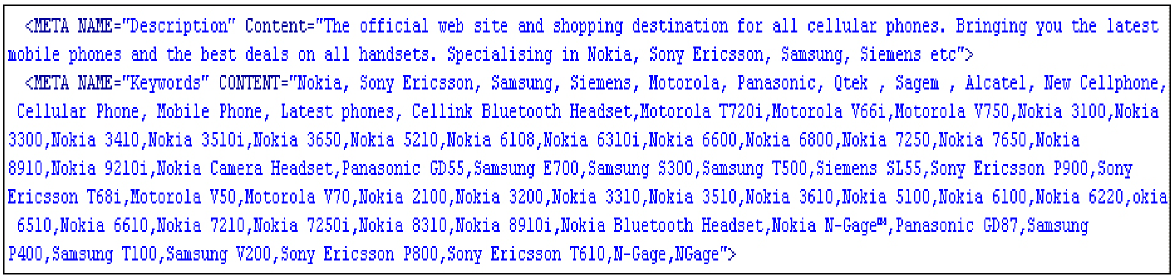
```
    <meta name="robots" content="index,follow">
    <meta name="description" content="Intersoft Computer Books South Africa is your home for a huge
range of affordable quality computer books">
    <meta name="keywords" content="computer books, free computer books,discount computer books,cheap
computer books,second hand computer books,discounted computer books,computer bookstores,free online
computer books,outlet computer books,wholesale computer books,computer bookstore,bargain computer
books,computer books online,how to computer books,computer books direct,overstock computer
bookstore,mail order computer books,
internet computer bookstore,low priced computer books,discount computer bookstore,best computer
books,Free Computer Books,halfprice computer books,bargain computer bookstore,Discount computer
books,Computer Books Direct,Computer Books,COMPUTER BOOKS,computer books cd,computer books on
CD,Computer books,Half Price Computer Books,Software computer books,Cheap Computer books,Computer
Books Online,Programming online computer books,word 2000 computer books,buy computer books
online,computer books cheap,sybex computer books,Computer Books for sale,McGraw-Hill COMPUTER
BOOKS,word xp computer books, Intersoft, computer books, south africa, McGraw Hill, Sybex, O'Reilly,
OReilly, Osborne, computer steps, Wiley, Dummies, Mastering, Beginning, learning, in easy steps, study
 guide, savvy, 24 seven, CRM ,Lotus ,Project Management,Cisco - CCNA ,Cisco, CCNP , CCSP ,CISSP
,Citrix ,CIW ,CompTIA , A+ , iNet+ , IT Project+,Linux+ ,Network+ ,Security+ ,Server+ ,CWNA ,ICDL /
,ICDL,Juniper ,Microsoft, MCAD/MCSD ,MCDBA ,MCSA ,MCSD ,MCSE ,MOUS ,Novell ,PHR / SPHR ,PMP
,SCWCD,Lotus Smartsuite ,MS Office ,MS Works,Apache Server ,BizTalk ,Linux ,Linux - Kylix ,Linux - Red
 Hat ,DOS ,Exchange Server ,IIS ,ISA ,SMS ,Windows ,Windows 2000 ,Windows 95 ,Windows 98 ,Windows ME
,Windows Server ,Windows XP ,Solaris ,Unix,Digital Audio ,Digital Photography ,Handhelds ,Palm ,PC
General ,PC Upgrade & Repair,Tablet PC ,Video,Borland C++ Builder ,C ,C++ ,ColdFusion ,COM ,Delphi
,Java ,JSP ,Macromedia ColdFusion ,Macromedia Contribute ,.Net ,ADO ,ASP ,C# ,VBA ,VBScript ,Visual
Basic ,Visual C # ,Visual C# ,Visual C++ ,Visual J# ,Visual Studio ,Visual Studio.Net ,Perl ,PHP
,Programming - General ,Python ,sendmail ,UML ,Web Development ,XML ,XSLT,Excel ,Sage,MS Word
,WordPerfect">
```

Image 33



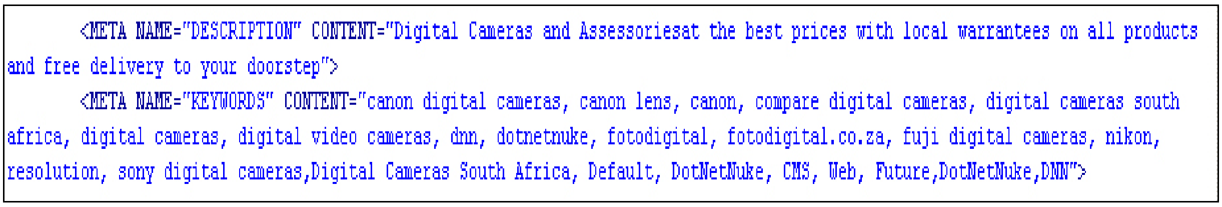
```
<meta name = "description" content = "Online store offering flowers and gift delivery service throughout South Africa. Buy corporate
 gifts, wedding, and baby gifts online.">

<meta name = "keywords" content ="buy flowers online, buy gifts online, baby gifts south africa, wedding gifts south africa,
corporate gifts south africa, gift store south africa, baby gifts, gift baskets, food gift baskets, corporate gifts, gifts for him,
gifts for her, gift certificates, anniversary gifts, wedding gifts, birthday gifts, baby gift baskets, baby shower gifts,
personalized baby gifts, flower gifts, unique gifts, gift cards, business gifts, christmas gifts, holiday gifts, online gifts,
newborn gifts, executive gifts, wine gifts, gourmet gifts, gift shops, gifts for her, luxury gifts, chocolate gifts, wine gift
baskets, wedding anniversary gifts, bridal gifts, gift boxes, valentine gifts, gifts for dad, unique baby gift, flowers south
africa, send flowers, flowers delivery, flower shops, flower online, order flowers, buy flowers, exotic flowers, online flower
delivery,  bridal flowers, flower gift baskets, flower baskets online, gifts south africa, send gifts to south africa">
```

Image 50



```
  <META NAME="Description" Content="The official web site and shopping destination for all cellular phones. Bringing you the latest
mobile phones and the best deals on all handsets. Specialising in Nokia, Sony Ericsson, Samsung, Siemens etc">
  <META NAME="Keywords" CONTENT="Nokia, Sony Ericsson, Samsung, Siemens, Motorola, Panasonic, Qtek , Sagem , Alcatel, New Cellphone,
 Cellular Phone, Mobile Phone, Latest phones, Cellink Bluetooth Headset,Motorola T720i,Motorola V66i,Motorola V750,Nokia 3100,Nokia
3300,Nokia 3410,Nokia 3510i,Nokia 3650,Nokia 5210,Nokia 6108,Nokia 6310i,Nokia 6600,Nokia 6800,Nokia 7250,Nokia 7650,Nokia
8910,Nokia 9210i,Nokia Camera Headset,Panasonic GD55,Samsung E700,Samsung S300,Samsung T500,Siemens SL55,Sony Ericsson P900,Sony
Ericsson T68i,Motorola V50,Motorola V70,Nokia 2100,Nokia 3200,Nokia 3310,Nokia 3510,Nokia 3610,Nokia 5100,Nokia 6100,Nokia 6220,okia
 6510,Nokia 6610,Nokia 7210,Nokia 7250i,Nokia 8310,Nokia 8910i,Nokia Bluetooth Headset,Nokia N-Gage™,Panasonic GD87,Samsung
P400,Samsung T100,Samsung V200,Sony Ericsson P800,Sony Ericsson T610,N-Gage,NGage">
```

Image 51



```
        <META NAME="DESCRIPTION" CONTENT="Digital Cameras and Assessoriesat the best prices with local warrantees on all products
and free delivery to your doorstep">
        <META NAME="KEYWORDS" CONTENT="canon digital cameras, canon lens, canon, compare digital cameras, digital cameras south
africa, digital cameras, digital video cameras, dnn, dotnetnuke, fotodigital, fotodigital.co.za, fuji digital cameras, nikon,
resolution, sony digital cameras,Digital Cameras South Africa, Default, DotNetNuke, CMS, Web, Future,DotNetNuke,DNN">
```

GLOSSARY

Boolean searching

The Use of operators AND, OR, NOT to combine search terms.

Cloaking

A technique used to display different or obscure pages to the search engine.

Crawler

The software that is used by search engines to find webpages to index. It is also known as a spider or robot.

Directory

A search tool that depends on human editors to review and index websites. Unlike search engines, a directory categorises websites according to topics.

Doorway pages

Two sets of webpages are sometimes designed. The first page, also called a doorway page, often contains keyword-rich text, which crawlers' index well. Human visitors are redirected to the second set which contains human friendly text and graphics. The first set therefore achieves a high ranking while the second set pleases the human visitor.

E-commerce

Denotes the selling of products over the Internet. Sales can be from business to consumer or from one business to another. E-commerce may also refer to Electronic Data Interchange (EDI), in which one company's computer queries and transmits purchase orders to another company's computer.

Exclusion Policy

Search engine policies often state the criteria for submission of websites. Some go further by listing web design practices to be avoided, which if found in submitted websites, can lead to that site being excluded from the search engine index.

Gopher

A protocol that was used to publish information on the Internet before the existence of the World Wide Web.

Indexing

Search engine indexing describes the process of reading a web page and extracting the contents into a search engine database.

Invisible web

This refers to information that is available on the web, but cannot be located through the use of search engines. This information is invisible to search engine crawlers for various reasons, including search engine policy decisions, (opting not to index certain formats) and information that is located behind a firewall.

Keyword Stuffing

The repetition of keywords in a website, in an attempt to increase its ranking in search engine results.

Link farm

A link farm is a collection of web pages that contain hyperlinks to one another or other pages. The main aim is to attempt to deceive search engines that place emphasis on the number of links in a website, when determining relevancy.

Meta search engine

A meta search engine is a search tool that retrieves results from a database of a number of search engines. When a searcher performs a search query on a meta search engine, the query is transmitted across several search engines and directories.

Page redirect

The process whereby a user is redirected to another site when attempting to access a website. This is viewed as spam because users visit a different webpage than the one that is viewed by search engine spiders.

Paid inclusion

This is a search mechanism advertising practice where webpages are included in search engine indices in exchange for payment. This practice does not guarantee top rankings.

Paid placement

Under paid placement, websites are guaranteed top rankings for certain keywords at a fee. Participating websites often bid for these rankings, and the highest paying website gets the top rank.

Paid submission

Paid submission refers to the process whereby web designers pay certain fees to have their websites reviewed by directory editors quicker than it would normally take. It does not affect ranking or chances of indexing websites.

Robot

See crawler.

Search engine

A search tool that uses an automated crawler to index words on web pages, thereby enabling full text searching of the Internet. It is designed to assist users to find relevant information on the Internet.

Search engine optimisation

The process of identifying factors in a website which could impact search engine accessibility to the website. It involves fine-tuning the many elements of a website so that it can achieve the highest possible visibility when a search engine responds to a relevant query.

Search Engine Spam/ Spam

Spam refers to the process whereby web designers attempt to influence the results of search engines by applying SEO practices that are perceived as unethical. Spam can also be defined as using words, HTML code or scripting on a webpage which is not intended to enhance user experience. This type of spam differs form e-mail spam. The latter refers to email that is unwanted, unsolicited and which often contains junk.

Search engine user

This is a general term describing all clients using search engines either for general search or website design.

Spider

See crawler.

Veronica

A search engine that was primarily designed for the Gopher protocol. When the Web superseded Gopher, it also led to the defunct of the Veronica search engine.

Visible web

Consists of all the webpages that are retrievable via search engine crawlers.

Website

A website is a collection of webpages designed to present information over the World Wide Web.