Proceedings of the IASTED International Conference
Internet and Multimedia Systems and Applications (EuroIMSA 2009)
July 13 - 15, 2009 Cambridge, UK

WEBSITE VISIBILITY – QUANTIFYING NEGATIVE SEARCH ENGINE RANKING ELEMENTS FOR OPTIMISATION

Melius Weideman
Faculty of Informatics and Design
Cape Peninsula University of Technology
Cape Town
South Africa
melius@gmail.com

ABSTRACT
The Internet has become a major marketing and sales channel, and there is strong competition between websites for high rankings in search engine results.

A variety of ways exist to ensure high rankings, some involving organic coding, others payment to search engines. The focus of this research paper is identifying and ranking the elements to be avoided during the design of a webpage. Failure to do so could earn a website permanent banning from search engine indices.

A number of empirical experiments and literature surveys produced models which identify, and some rank, website visibility elements. The top SEO experts also produced a ranking of elements from a practitioner angle, with both a positive and a negative effect on website visibility. However, no evidence could be found of a ranking system which combines the rigour of academic research with the reliability of industry expertise.

These models were scored, normalised, and elements which were the same but carried different names were identified and combined.

The result is a ranking system, triangulated between academic research, empirical experiments and industry expert opinion. Seventeen negative elements were ranked, with link spamdexing and keyword spamdexing being the top two elements to avoid.

Website designers should use this scale during mission critical website design to identify and avoid elements which could cause their website being banned from large search engine indices.

KEY WORDS
website, visibility, search, engine, ranking, optimisation

1. Introduction

There can be no doubt that the Internet has had a far reaching impact on business and our everyday lives. It is claimed that email and Internet searching are respectively the most and second most used Internet applications. Furthermore it is claimed that over 80% of Internet traffic is generated by searchers. Thus online communication and the use of search engines to find relevant information are the two dominant components of the way we use the Internet.

Research has proven that an average of 91% of searchers do not read results past the third Search Engine Result Page (SERP) [1]. Since searching and the use of search results play a major role in for example online purchase decisions, it has become imperative that mission critical websites occupy the top positions on SERPs. Website designers use either proper organic design of webpages or paid systems (or a combination) to achieve high rankings on SERPs. Guidelines exist on which elements to include in and exclude from organic design to achieve high rankings [2]. However, no evidence could be found of a ranked system based on academic research combined with the industry's expert opinion. The research problem can be formulated as: no clarity could be found as to the relative importance of negative website visibility elements, based on both academic research and practitioner expertise.

The purpose of this research was to establish a ranked, triangulated sequence of elements which could earn a website banning orders from search engines. These elements should therefore be avoided in design, although some of them (eg JavaScript and images) might have to be used under certain conditions. A ranking system will indicate whether or not certain negative elements can be included in website design.

The contribution of this research is to provide a ranked listing of website visibility elements to website designers. This list provides clear guidance on which elements to avoid at all costs (those high up on the list, including link and keyword spamdexing) and which ones could be used when a good reason exists to do so (those lower down on the list, including JavaScript and dynamic webpages). It furthermore has highlighted the existence of spamdexing in many more forms than the traditional ones. Finally, a warning goes out to designers against falling into the trap of designing and coding using any form of unnatural excess, which can be construed by search engine algorithms as spamdexing.

2. Previous Research

2.1 The Environment

Search Engine Optimisation (SEO) is the process of finetuning a website in such a way that crawler visits will result in high rankings and possibly large numbers of visitors [3]. Since large numbers of visitors could imply high sales and high profit, extreme attention has been focussed on website visibility and achieving the coveted top positions in search engine results. Unfortunately this has spawned the birth of search engine spam, better known as spamdexing (SPAMming the search engine inDEX). Spamdexing is any attempt to manipulate the relevancy of a webpage in an attempt to earn a higher ranking than what the content of the webpage deserves [1].

2.2 Models

A number of research projects have been done on the elements affecting website visibility. In an early model Binnedell compiles a simple list indicating positive and negative elements, without an attempt at ranking them [4]. Secondly, another model was found with the same attributes, except this time an attempt was made to rank the elements after empirical work [5]. Thirdly, a detailed account of a series of empirical experiments, combined with expert interviews and an in-depth literature survey produced a model with further ranking indication [6]. Finally, an industry model was analysed where 37 SEO experts' opinions were summarised [7].

The Binnedell model is presented in Figure 1. The elements preceded by a plus sign are claimed to have a positive effect on website visibility, and those with a minus sign a negative effect.

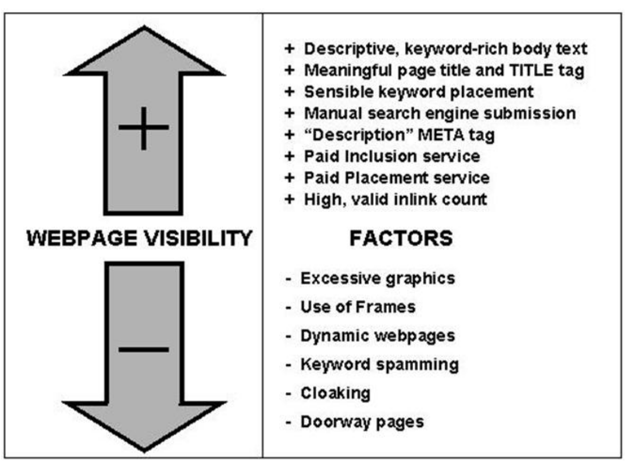

Figure 1. The Binnedell Model

No empirical work was done, and no attempt at ranking the elements was presented. However, the identification and grouping of the elements have all been confirmed in subsequent studies as being correct. A total of six negative elements are listed, in no particular order.

Secondly, Chamber's model was identified as being peer reviewed with tested academic research results - see Table 1. In this case ranking was done, using a points scheme based on an expert interview, literature survey and empirical experiments. These include 12 keywords being used on six search engines, over 444 websites. Manual searching and SERP interpretation was done. A lower "Rank" figure in the table indicates higher level of importance, and positive and negative elements are mixed.

Table 1
The Chambers Model

NO	VISIBILTY ELEMENTS	RANK
1	Inclusion of meta tags	1.5
2	Hypertext / anchor text	2
3	No Flash or fewer than 50%	3
4	No visible link spamming	4
5	Prominent link popularity	4.5
6	No frames	5
7	Prominent domain names	7
8	Prominent headings	7
9	No banner advertising	8
10	Prominent HTML naming	10
	conventions	

The negative elements from the model are, in order from "most negative" to "least negative":

- the use of Flash technology,
- link spamdexing,
- frames, and
- banner advertising.

Thirdly, Visser did empirical and other work to produce a model which subcategorises both positive and negative elements as being "more" or "less" crawler friendly or unfriendly [6]. See Figure 2.

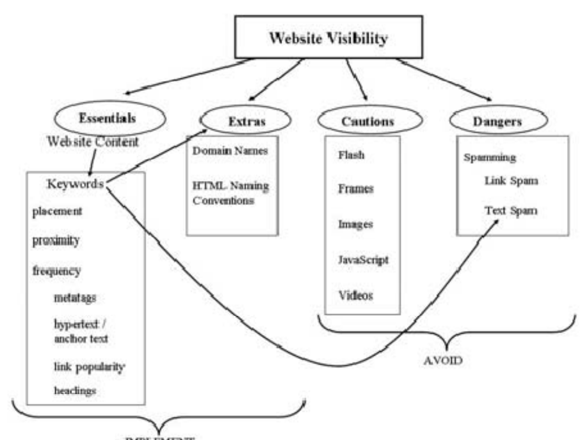

Figure 2. The Visser Model

Visser uses specific terms to indicate the level of "positiveness" or "negativeness" of elements:

- "Essential" is most positive,
- "Extra" is least positive,
- "Danger" is most negative and
- "Caution" is least negative.

According to Visser, the two most negative elements are link and text spamdexing.

All three models listed above were based on academic research and experiments, and the author deemed it necessary to add these results to those of a literature summary as well as industry expert opinion. Fishkin, himself a SEO expert, completed a research project by involving the top 37 SEO world experts in an extensive survey with over 200 questions [7]. The results were tabulated, indicating a score and standard deviation for each element.

An extract of the Fishkin model is presented in Figure 3. In this example, "Overuse of Targeted Keywords" has been rated at 3.3 out of 5, which indicates a medium level of importance, with a standard deviation of 1, which indicates an average level of consensus amongst the experts.

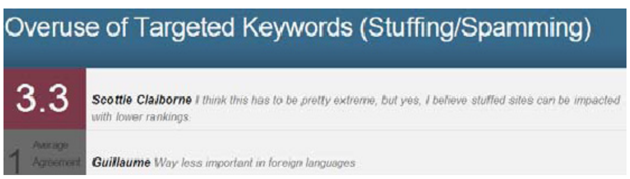

Figure 3. The Fishkin Model (extract)

Fishkin lists "Server is often inaccessible to Bots" and "Content Very Similar or Duplicate of Existing Content in the Index" as being the two most negative elements. However, the standard deviation figures for each score vary from "Average Agreement" to "Highly disputed", underlining the volatile nature of the SEO world.

2.3 Synthesis

The results of these models are summarised in Table 2.

It is clear that three negative elements appear three times across the four models: frames, keyword spamdexing and link spamdexing. However, at this stage there has been no weighting incorporated, and a new set of weighted points will have to be implemented before any ranking of elements can be done. It was considered prudent to triangulate the two sets of results so far (academic models and industry model) with a short literature survey. This would confirm or refute the validity of these elements being on the list.

Table 2

Combination of four Models

	BINNEDELL	CHAMBERS	VISSER	FISHKIN
Banner Advertisements		X		
Cloaking	X			
Doorway Pages	X			
Duplicate Title/Metatags				X
Dynamic Pages	X			
Ext. Links Low Quality				X
Flash		X	X	
Frames	X	X	X	
Graphics	X		X	
Inbound link spamdexing				X
JavaScript			X	
Keyword spamdexing	X		X	X
Link spamdexing		X	X	X
Low level of Visitors				X
Server Inaccessible				X
Slow Server				X

2.4 Literature Summary

Ample evidence exists to confirm that the three elements which have the highest density in Table 1 are well-known as being “suspect” in website design. These three; frames, keyword spamdexing and link spamdexing, are mentioned repeatedly in the literature.

Framing is often used on webpages to create a consistent navigation layout. However, a typical framed webpage contains very little for a search engine crawler to index, and often has no links to follow [8]. Webpages beyond the one where a crawler finds itself busy indexing, are likely to be ignored for this reason [9], [10]. Furthermore, some authors believe that a typical webpage should not exceed the vertical dimension of a screen since users have an aversion to scrolling down. This makes the use of frames unnecessary [11]. However, there are workarounds one could use to implement frames without the associated decrease in website visibility, at a cost in programming time and design [8].

Secondly, keyword spamdexing involves the repeated use of one or more keywords/key phrases on a webpage, beyond what is considered to be acceptable, readable text. The aim is to convince the visiting crawler that these keywords/phrases are really what the website is all about, and then gain (undeserved) high rankings with the search engine. A common methodology is to hide the keywords from human view, but allow them to be visible to crawlers. This can be done by making the text invisible, hide it behind layers, place it at the bottom of long pages, hide it inside metatags or make it so small that is appears as a thin line [12], [13], [2], [5], [14], [15].

Thirdly, the use of link analysis in search engine algorithms is common, which puts a premium on the quality and quantity of inlinks for a given webpage [16]. This has resulted in the growth of a type of abuse termed link farms or Free For Alls (FFA). A FFA is a webpage which simply lists thousands of links to other pages, with no content of value [17].

Many other technologies, programming features and ways of presenting or manipulating elements of webpages have been used to present a webpage to a search engine in a way so as to undeservedly enhance its ranking. These include doorway pages and cloaking. A detailed discussion of these and other negative factors is outside the scope of this paper.

3. Methodology

3.1 Background

In order to achieve the objective, the next step was to allocate points to each element of each model, combine identical (but differently named) elements and produce a ranked list. This combination of elements was based on the author's expertise and extensive research of the past, since it required an understanding of the factors involved in the implementation of these elements.

A random value of 40 points was allocated to the oldest academic model, increasing to 50 and 60 respectively for the younger ones. The argument was that in the fast changing world of SEO, more recent research should carry more weight. To allow theory and practice an equal input into the system, the (only) practitioner model received 40 + 50 + 60 = 150 points to divide amongst its elements.

Where no ranking was done (eg the Binnedell model), equal points would be allocated to each element. With ranked models, points would be divided according to rank. A higher score indicates a "more negative" effect on website visibility.

3.2 Binnedell Model

According to Figure 1, this model lists six negative elements without ranking. Each one of these elements would thus receive 40/6 = 6.66 points – see Table 3.

Table 3
Binnedell Model Points Allocation

ELEMENT	POINTS EARNED
Excessive graphics	6.66
Use of Frames	6.66
Dynamic webpages	6.66
Keyword spamdexing	6.66
Cloaking	6.66
Doorway pages	6.66 **(TOT: 40)**

3.3 Chambers Model

This model lists four negative elements – see Table 1. Their respective weights were taken to be 4, 3, 2 and 1, for a total of 10. The "most negative" element was allocated 4/10 x 50 = 20 points, down to 1/10 x 50 = 5 points for the "least negative" one – see Table 4.

Table 4
Chambers Model Points Allocation

ELEMENT	EFFECT	POINTS EARNED
Flash content	Most Neg	20
Visible link spamdexing	Neg	15
Frames	Neg	10
Banner advertising	Least Neg	5 **(TOT: 50)**

3.4 Visser Model

This model proposes seven negative elements, ranked into two categories. Equal points had to be allocated inside each category, but with the "more negative" elements (labelled "Dangers") receiving a higher score than the "less negative" ones (labelled "Cautions") in Figure 2.

Using the same calculation system as with the other models, the five "Caution" elements received 6.4 points each, and the two "Danger" elements 14 points each – see Table 5.

Table 5
Visser Model Points Allocation

ELEMENT	EFFECT	POINTS EARNED
Link spamdexing	More Neg	14
Text spamdexing	More Neg	14
Flash	Less Neg	6.4
Frames	Less Neg	6.4
Images	Less Neg	6.4
JavaScript	Less Neg	6.4
Videos	Less Neg	6.4 **(TOT: 60)**

3.5 Fishkin Model

This model lists nine negative elements [7], giving a total weight of 9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1 = 45. The highest score was thus 9/45 x 150 = 30 points, down to 1/45 x 150 = 3.3 points - see Table 6. The "ranking" column in this table lists the score received out of 5, as allocated by Fishkin, where a higher score indicates a "more negative" element. Since some rankings were equal (for example, elements 4, 5 and 6 each achieved a ranking of 3.3), these groups of scores had to be averaged. The second figure in italics in Table 6 is the new, averaged weight, used for the final calculations.

Table 6
Fishkin Model Points Allocation

ELEMENT	RANKING	POINTS EARNED
1 Server is often inaccessible to bots	3.8	30
2 Content is very similar or duplicate	3.6	26.7 *25*
3 External links to low quality/spamdexing sites	3.6	23.3 *25*
4 Duplicate title/meta tags on many pages	3.3	20 *16.7*
5 Keyword stuffing	3.3	16.7 *16.7*
6 Participation in link schemes	3.3	13.3 *16.7*
7 Very slow server response times	2.8	10
8 Inbound links from spamdexing sites	2.1	6.7 *5*
9 Low levels of visitors	2.1	3.3 *5*

3.6 Weideman Model

Continuing in the tradition set by this paper of using author surnames as model names, the next step would be to combine the first four models into the final one, termed the Weideman model. This kind of "marriage" between academic research and practitioner expertise has never been done in the SEO world, according to the author's knowledge. The 26 negative elements identified in the first four models were listed together, and duplicates combined – this left nine unique negative elements. For example, the following three elements are identical:

- Visible link spamdexing (Chambers) 15
- Link spamdexing (Visser) 14
- Participation in link schemes (Fishkin) 16.7

Creating a new name for all three (Link spamdexing) and adding the points, produced one entry in Table 7 and Figure 4 with a value of 45.7 points.

Another eight elements did not have any duplicates, so they were calculated separately, and combined with the nine elements summarised already.

4. Result and Analysis

In summary, Table 7 and Figure 4 provides a presentation of the relative magnitudes of the 17 (9 + 8) validated, ranked negative elements considered to have a negative effect on website visibility. Elements without prefixes originate from more than one model, while those with a prefix are from one model only (b = Binnedell, c= Chambers, etc).

Table 7
Weideman Model Points Summary

NO	ELEMENTS	POINTS
1	Link spamdexing	45.7
2	Keyword spamdexing	37.36
3	f-Server is often inaccessible to bots	30
4	Flash	26.4
5	f-Content is very similar or duplicate	25
6	f-External links to low quality/spamdexing sites	25
7	Frames	23.04
8	Graphics	19.46
9	f-Duplicate title/meta tags on many pages	16.7
10	f-Very slow server response times	10
11	b-Cloaking	6.66
12	b-Doorway pages	6.66
13	b-Dynamic webpages	6.66
14	v-JavaScript	6.4
15	c-Banner advertising	5
16	f-Inbound links from spamdexing sites	5
17	f-Low levels of visitors	5

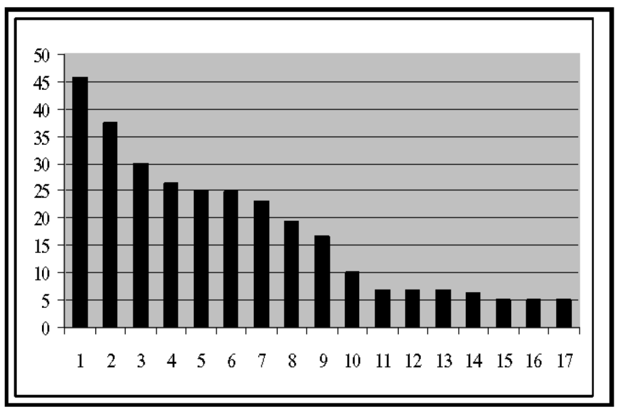

Figure 4. Weideman Model Relative Magnitudes

When analysing these results, it is clear that two kinds of obvious spamdexing occupy the first two positions, while a number of others have similar attributes to the spamdexing ones.

The following elements (over and above numbers one and two) are considered to contain a form of unnatural excess, which is typically associated with spamdexing:

5 Content is very similar or duplicate
6 External links to low quality/spamdexing sites
9 Duplicate title/meta tags on many pages
12 Doorway pages
16 Inbound links from spamdexing sites

It is therefore clear that website designers should avoid at all costs any form of unnatural excess, generally termed spamdexing. Search engine algorithms can and are often reprogrammed to detect this kind of excess, since it involves mindless repetition of keywords, text, hyperlinks or complete webpages. Webpages found guilty of spamdexing are normally either blacklisted or completely banned from search engine indices.

The first nine elements have a substantially higher weight than numbers 10 to 17, which implies that the avoidance of these nine is not negotiable. The last eight appear to be less critical.

5. Implementation

The results obtained through this research have been applied fully in the design and coding of a static website, www.book-visibility.com. None of the 17 negative elements were included in the design, as far as the website owner is capable of controlling these factors. Some are outside the ambit of the website owner – numbers 3, 10, 16, and 17, for example.

6. Conclusion

The Weideman model confirms through experimentation and expert knowledge that spamdexing in general is the biggest contributor to website banning, and that it manifests in some ways previously not identified as such.

The advantages of the results include the fact that a website designer can use the ranked list to identify which elements should be avoided at all times (numbers 1 to 9). Some of the others are outside the control of the designer and the website owner – numbers 16 and 17, for example.

Limitations of the study include that no indication was given of positive elements, and how they can be used to enhance website visibility. This will be done in a future study.

This research has produced a valuable contribution to the body of knowledge in the SEO world. Most academics and practitioners are aware of spamdexing and its effects on websites. However, this research sheds new light on how "innocent" practices are construed as spamdexing by crawlers, and ranks them accordingly.

References

[1] M. Weideman, *Website Visibility: The Theory and Practice of Improving Rankings* (Oxford: Chandos Publishing, 2009).

[2] S. Thurow, *Search engine visibility* (Indianapolis: New Riders, 2003).

[3] R.A. Malaga, Worst practices in search engine optimization, *Communications of the ACM, 51*(12), 2008, 147-150.

[4] M. Binnedell, *The WBPN model: a proposed design approach to maximize website visibility to search engines*. (Bachelor of Technology mini-thesis, Cape Technikon, 2003).

[5] R. Chambers, *Search engine strategies: a model to improve website visibility for SMME websites* (Masters Thesis, Cape Peninsula University of Technology, 2005).

[6] E.B. Visser, *Search engine optimisation elements' effect on website visibility: the Western Cape real estate SMME sector* (Masters Thesis, Cape Peninsula University of Technology, 2007).

[7] R. Fishkin, & J. Pollard, Search Engine Ranking Factors V2, *http://www.seomoz.org/article/searchranking-factors*, 2007, [22 May 2009].

[8] J. Whalen, How to Optimize a Framed Site for High Rankings, *http://www.webproguide.com/articles/How-to- Optimize-a-Framed-Site-for-High-Rankings/*, 2000, [23 May 2009].

[9] T. Dahm, Search Engine Optimization Tips, *http://www.webdevelopersjournal.com/articles/search_strategies_tips.html*, 2000, [20 May 2009].

[10] M. Weideman, & R. Chambers, The application of best practice towards improving website visibility to search engines: a pilot study, *South African Journal of Information Management*, *7*(4), 2006, *http://www.sajim.co.za* [23 May 2009].

[11] E. Zeeger, Why Frames are bad for search engine optimization, *http://www.seomatrix.com/newsletter/archive/2006/SEONewsletter-July-2006/frames-bad-for-optimization.php* 2008, [22 May 2009].

[12] F. Mbikiwa, & M. Weideman, Implications of search engine spam on the visibility of South African ecommerce Web sites, *South African Journal of Information Management, 8*(4), 2006, *http://www.sajim.co.za* [28 May 2009].

[13] R. Goldsborough, Get Your Site noticed without Spamming, *Black Issues in Higher Education, 21*(26), 2005, 40.

[14] M. Weideman, Ethical issues on content distribution to digital consumers via paid placements as opposed to website visibility in search engine results. *Proceedings of ETHICOMP 2004,* Syros, Greece, 2004, 904-915.

[15] L. Introna, & H. Nissenbaum, Shaping the web: Why the Politics of Search Engines Matters, *Information Society*, *16*(3), 2000, 169-186.

[16] A.H. Keyhanipour, M. Piroozmand & K. Badie, A GP-adaptive web ranking discovery framework based on combinative content and context features, *Journal of Informetrics, 3*(1), 2009, 78-89.
[17] M. Thelwall, Commercial web site links, *Internet Research: Electronic Networking Applications and Policy*, *11*(2), 2001, 114-124.