SEO Truth

A Bible for the Next Generation of Search Engine Optimisation

Phillip Midwinter, Search Engineer for the Advertising Industry

Preface

First off, I'd like to introduce myself. I'm a Search Engineer, a developer and programmer. I've worked with clients throughout the advertising industry at many different companies. My specialty is developing software that works with the search engines of companies like Google, Yahoo and MSN and attempts to influence the rankings of my client's websites, as well as report on those ranking changes. I've never been to a lecture on computer science, read a book on development methodology and yet I'm in demand. My skills lie in understanding the technology of a search engine and how to capitalise on their ranking algorithms, web crawlers and content filters and it's the ideas I generate in this area which have kept me in gainful employment.

SEO (Search Engine Optimisation) used to be a fairly simple task where you'd make sure every page on your client's site had Meta tags, descriptions and content unique to that page. You might then try to analyse the keyword density of your key terms to keep them somewhere between 4 and 7 percent. More often than not most SEO companies wouldn't even attempt that.

What most SEO companies would never tell you,

and this is the industry's most well kept secret, is that they're intrinsically lazy. If you had a good client, with good content and a product of interest then their SERs (Search Engine Rankings) would climb entirely naturally to the top spots, you'd have nothing to do but sit back and reap the benefits of your lack of work.

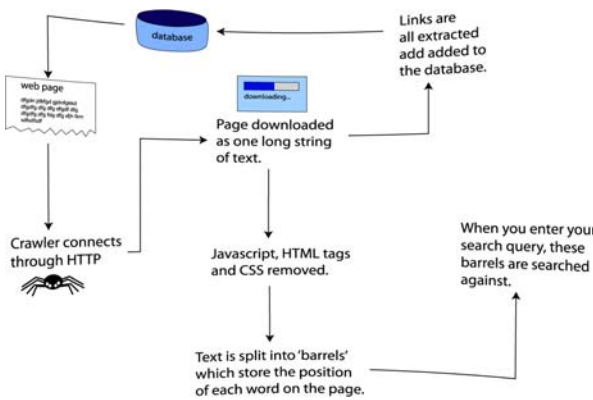
This is of course a sad state of affairs which no real SEO company would allow and part of this book will help you to spot the difference between a professional outfit and rank amateurs and define the widening gap between the two camps.

As the title suggests I'm writing about the next generation of SEO. It's becoming more difficult to increase the rankings of a particular website and it will only get more difficult to manipulate a website's ranking without any understanding of how new search engine technology works. Lucky for you, my field is semantics (how to correlate the relationship between one word and another essentially) and you're in for a whole chapter in manipulating a semantic index similar to those increasingly used by the major search engine players.

Chapter 1 - The Past

In order to proceed correctly in the future, the most important lesson is for us to understand what happened historically. There's no shortage of information on the internet and amongst SEOs and webmasters about how Google's original *PageRank* system worked. This is in large part thanks to a paper written by Google's founders, Larry Page and Sergey Brin, whilst they were still studying for their PhDs at Stanford University. Not long after that they received their first investment from a company called Sun Microsystems which enabled them to build upon the hardware they had in their university dorm room and create the international phenomenon we know today.

PageRank was essentially a very simple system. It counted each link from one site to another as a vote for the destination site. By voting for another site the original gave away some of its own *PageRank.* The idea came from Salton's Vector Space Model, which is a mathematical principal known to most Computer Science graduates today. This simple method of calculating which websites had the most votes, and therefore deserved higher rankings, is key to all search engine algorithms as it's extremely fast to calculate. The most important factor in any



The Anatomy of a Search Engine, based on the work of Larry Page and Sergey Brin whilst at Stanford.

search engine is its speed in returning and ranking results, especially when you're dealing with an index of billions of pages.

If you understand that all calculations undertaken by a search engine must be as fast as possible, it allows you to draw logical conclusions:

- Thinking about a page as a machine would (which struggles to actually understand rather than just read), rather than as a human, is key to analysing your websites content for SEO value.
- Is every single underlined heading, keyword color, font size, image location, keyword relationship and page title length analysed when a page is crawled? It's highly doubtful that anything too in depth is going to be indexed, when the crawler has another hundred thousand pages to visit and rank as quickly as possible, use some common sense here. Of course as processor speeds and bandwidth increase more in depth analysis will become possible in a shorter space of time.
- The search engine needs to maximise two things: the speed of its calculations and its measure of quality relevancy.

> Occasionally one is going to suffer at the importance of the other, if you were going to choose between indexing a page poorly - or not at all - which would you do?

SEOs in the past were able to capitalise on this speed issue by choosing to concentrate on areas of a page such as the Meta tags, description and page title. The content itself gradually became more important as time went on but still was subject to the speed of indexing. SEOs quickly realised that keyword density (how many times a keyword appears on a page out of the total number of words) was a very quick way to determine some kind of relevancy, and that the search engines were using it too.

Once the search engines got wise they implemented filters that stopped SEOs from flooding a page with keywords. Arguments in the SEO community followed over exactly what was the ideal keyword density for a term, and this usually settled somewhere between 4 and 7 percent.

Of course the *PageRank* model meant that agencies were keen to build as many links to their client websites as possible. To make matters worse however they were after links that already

had high *PageRank* values to gain the maximum ranking as quickly as possible and this sprang up a cottage industry of people generating high *PageRank* links, purely to sell on. Google of course were unhappy about this and their anti-spam team began its work. Blacklisting of websites which 'farmed links' was becoming fairly common and this moved on to other aspects of 'black hat' SEO behavior - where an unfair advantage was being made by some nefarious companies and individuals.

Most SEO agencies at this stage relied heavily on staff who'd be subjected to some extremely tedious and repetitive labour. Going through page after page of a website and adjusting the number of keywords on a page, slightly changing each page title and Meta tag was a boring job and not well paid.

Directors and CEOs didn't have a whole stack of problems though, if they kept building up link relationships with ranking websites and making sure their Meta tags were in place, their job was done. Often enough they'd have clients who already had an interesting product which did most of the work itself, spreading links around the internet as people registered their interests.

This natural traffic increase was what Google was

looking for as they wanted sites which progressed on their own merits rather than trying to beat the system.

Chapter 2 - Changing the System

Search engines, such as Google, used to proudly display the colossal numbers of pages currently in their website indexes. In the long run the size of their index was to be a problem and one that would ultimately make it harder for them to deal with the problems created by the SEOs trying to game their system.

The numbers of indexed pages came down, it's not known how many are currently in the database but some theorise that there are even less pages today than there were a few years ago. In order to understand why we need to go back to thinking like a machine:

- Each database row uses a rigid structure.
- To change this structure, it must be changed across the entire system.
- If you want to implement wide ranging changes to how your engine works, at some point you will need to change this rigid structure.
- After you've indexed billions of pages, the cost and time needed to change the structure is monumental.

As a result of these restrictions, engineers needed to find ways to change the structure of their indexes as little as possible and yet still find new ways to show how relevant a page was as quickly as possible.

One of the results of this was using the change in ranking information for each page over time. In a nutshell, by studying the rate of ranking change against time engineers produced algorithms allowing machines to calculate whether a page was ranking higher as a result of natural spread across the internet or by outside intervention. If your website appeared to spreading its link roots too quickly, blacklisting wasn't far away.

Whilst these protections were being added in, it's likely Google had begun to work on other methods of search engine indexing. Realising that the original model could only be adapted so far before it lost value altogether they came up with models for new indexes which today appear in verticals (news, blogs, finance and others). They also hired specialists in artificial intelligence and machine learning who worked on putting together indexes where the meanings of words themselves could be used to calculate the relevance of a page to a particular search query.

Some of these verticals are now integrated into

Google's main search. Some SEOs believe that Google began completely re-indexing from scratch, afforded anonymity by the hidden numbers of indexed pages and using the time before the next generation of the web hit to amass suitably large indexes. Personally I think this unlikely but it is possible that one of the major algorithm shifts was in fact a total index change as well. The point is, nobody outside of their campus in Mountain View knows for sure what happens inside their walls.

Other major search engines were more open about the changes to their systems. In particular MSN were keen to be seen as coming up with entirely new search engine models, and their latest iteration; Windows Live Search, is not only visually different but technically as well. Google maintained their status, partly through shrouding their changes in secrecy, appearing to display the same engine throughout the years.

With the advent of Web 2.0 (a term used to define the shift in website usage over the recent past), new problems were created for the engineers at the major search engines. In the past it was next to impossible for a website to appear with hundreds or thousands of links to it out of nowhere without some serious SEO intervention. Social networking and news websites as well as

blogs changed all this. Within hours of an article on an almost unlinked backwater blog being submitted to Digg.com, it could gain huge traffic and build links at an unprecedented rate. Aiming to calculate whether a site had been manipulated by looking for a steady, natural growth pattern was no longer feasible.

It didn't take a long time for SEOs to realize this state of affairs and in short order they harnessed the power of social networking to submit stories on their client websites, specifically written to garner social interest.

SEOs also took advantage of Google's new verticals, and their connection to the main index. Company blogs started to appear, often in fact written by an SEO and full of content that was shaped to the key search terms their clients were looking for. These blogs were then indexed by Google in their dedicated vertical, giving a much broader content picture for their client and generating many more opportunities for links.

There was now a new emphasis upon SEO: quality content and lots of it. Copywriters began their jobs in earnest churning out pages of well written content on every facet of their client's business, which would then be blogged upon and submitted to social networking. A new breed of

SEO also sprang up, the independent writer, who endeavored to cover key niches of content and fill their websites with advertising to generate an income. All of these combining factors drove Google's now well known Anti-Spam team into overdrive and rules were established to cover these epidemics of social advantage.

There may be those of you thinking that this is the present in SEO terms, or perhaps the near future. We're almost there but first we need to look at the difference between past and the future.

Chapter 3 – The Gap

The gap has always been present. It's the distance between those SEO professionals who use their acumen, experience and expertise to do a good job, and the amateurs who take your money and allow the natural increase in ranking of your website to suggest that they're doing their job correctly.

It may make more sense to some of you were this chapter at the end of the book; I've chosen to put it here to illustrate the gap more effectively. Any of the practices you've read about previously are old, they're a relic and if you or your SEO agent is using them, then I'd suggest reading the chapters that follow.

You can use the chapter numbering to assign a score to your agent. If they use a practice in chapter one, assign one point and so on. If they refuse to tell you what practices they do use – then they're wasting your time.

There's an important difference in the structural makeup of the SEO agencies that are committed to new practices and ideas; they'll employ programmers and artificial intelligence specialists as well as the copywriters, HTML scripters and

account management team. The programmers, or search engineers as we are now known, are all important to these businesses.

Their job is to advise the SEO staff on what makes good content, to program reporting tools that allow their clients to see how effective their campaigns are and to create 'scale models' of the search engines they're attempting to manipulate. Often they'll be from a web development background and will be part of the development team; meaning that your site is created with SEO in mind.

However, in an agency that's still committed to old way of thinking which is becoming less and less effective on a monthly basis, you'll likely be continually told about:

- Meta tags, a throwback to the old search engines are so close to obsolete as to be irrelevant. Remember time spent adjusting this is money that could be better spent elsewhere.

- Duplicate content, when they have no idea what constitutes duplicate content. Ask them exactly how a machine understands one page is similar to another – that'll usually stump them.

- Keyword densities, if quoted to you, are a real sign that this is a backwards agency. Real SEOs will use semantic profiles to establish the key content of a page.

- Gaining number 1 rankings, especially across a small number of key terms. Your company should be optimizing for at least 100 key terms – if not 1000. This is known as long tail theory and comes from PPC (Pay Per Click advertising). It generates more traffic in the long run even though you may have some slightly lower rankings for individual terms, so don't worry.

- Fantastically short time periods are (which as any SEO who is worth their salt knows) a bad idea for your company's long term rankings. You run the risk of being blacklisted and are unlikely to attain them without black hat intervention (this unethical and poor practice).

- Manually collated keyword reports, where some poor sap sits at a computer and manually checks every keyword on your report. Firstly, if they have the time

to do this they're wasting it when they could be doing SEO. Secondly, if they have a search engineer who's half decent they can throw together an automated system inside a week.

These points are just a few of the many areas that illustrate the gap. In the coming months and years, the gap is going to grow wider as those who have invested time and money into quality programmers (who understand the systems they're developing websites for) go leaps and bounds ahead of the copywriters. Many will try to justify their position and state previous records. In search engine technology past results are precisely worthless, because like everything else on the internet, SEO is continually changing.

On the other hand professionals may seem slow to start on occasion whilst they work out the correct system for your business, but in the long run they'll be streaks ahead and will have the reporting software to prove it to you.

Chapter 4 - Cat and Dog

Semantics and ontologies are the basis of the next generation analysis of content being carried out by the major search engines. When people ask me to explain to them the concept of semantics as it applies to search I try and illustrate with the following points:

- The human mind is able to understand the meaning of a word.

- A machine, intrinsically, cannot understand the meaning of a word. It needs software programmed to do so.

- If I were to ask a human for a word related to, 'cat', they could answer, 'dog'. They understand both are animals, both are domestic and that often appear in popular culture together.

- The human knows that these words are related, because they understand the semantics of language.

When computers are trained to understand language, this is usually a manual process. It involves building an ontology, which is like a tree

structure of categories which words may fall into any number of branches of. Then, when they look up a particular word, they see which other words fall into the same branches allowing them to have a limited understanding of a finite number of words.

There is another way to solve this problem and here we go back to the aspect of speed, and the cost of upgrading search engine systems.

I'd like you to look at the page prefacing this one; the first page of the chapter. If we take a look at one particular word, say, 'meaning', in the second bullet point. In a standard search index, that word would be stored on its very own row. In the rows immediately before would be the previous words in the sentence and vice versa. A poor measure of which words are related would be to take those immediately next to the word in question. A better measure would be to take those words, as well as assigning a 'distance' score to those 3 or 4 words before and beyond.

Still, this is very crude and the chances are we won't find words which are related in any way. When chance is involved in any aspect of mathematics, to increase your probability you simply repeat your experiment as many times as possible. Computers are excellent at repeating

simple operations like this and it's no difficult task for them to make this calculation thousands of times in a second.

What results from this is an extreme probability that the words with the highest distance scores, when we repeat this over the thousands upon thousands of pages in our search index, are going to be meaningfully related to our original word.

You can take this even further though as if we were to narrow down the pages we carried this operation out on, perhaps by time, so a different picture emerges. What we're actually calculating is which words are meaningfully related, for that particular part of history. If we narrow by age group, or by industry, we're finding the semantic relevance for each of those key groups. There's a lot of slang on the internet, and there are a lot of acronyms and buzzwords that vary across social groupings.

This kind of 'on the fly ontology' is immensely powerful for judging the subject of a page and allows search engineers to create a kind of 'semantic profile' for a particular website. This is the future of the keyword density chart.

Nobody currently knows how Google's semantic engines work but from Larry Page's and Sergey

Brin's original paper this is a highly possible (though simplified) method for them to be using. The arrival of Google Trends and Google's (experimental beta) Timeline feature also support this theory.

Chapter 5 – Speechless

It's very easy for SEOs to get too involved in the concepts of semantic linguistics. There is another new arrival on the internet that's given rise to changes in other search engine verticals; rich media. The term rich media, as I'm using it, pertains to images, video, flash and other non textual content. Indexing an image is a very different exercise to indexing a page of content and most SEOs are happy to just not try.

Part of the reason for the rise in the use of image search and video search comes from social networking. A social networking site allows quick, interlinked propagation of links and people don't like to waste time. A fine example of the phenomenon at work is *icanhascheezburger.com* (ICHCB) which simply posts pictures of cats, with simple captions superimposed over the top. ICHCB nonetheless enjoys high search engine rankings, massive traffic and is hugely popular. The time they spend on each post? It's probably less than a minute.

How, exactly, do we use image and video vertical search in SEO? There are many ways to do so and ultimately it comes down to what you're targeting. I'm going to go through just a few different advantages that can be drawn from image or video search, if it's something you're serious about then I'd go to an SEO who specializes in this field because it's wide ranging enough to justify that for the right content.

- As you would with an html page, name your video or image appropriately. Use underscores between words and remember that not only will the image name and its associated alt tag be used to gauge its context, but so will surrounding content.

- Always allow in-content images to be clicked on and then opened in a separate page. Underneath the full size picture simply add related tag words and a brief description or title. This isn't duplicate content but a site feature.

- Like other verticals, targeting images specifically can help to increase your rankings on the standard Google search index. There are also less checks and balances in place on these systems

because they're newer.

- If you are attempting to have rich media content crawled, submit it to a special image or video themed blog with little written content. Tag all the media effectively and ping it through to news feed aggregators through services like *pingomatic.com*, you'll be indexed much more quickly. In Google's case, just submit to YouTube and Google Video.

As technology is improving, new methods of indexing video and images are coming to the fore and will each have to be optimized for in the same manner as written content:

- Image recognition, by recognizing shapes and patterns computers may be able to discern the elements of a picture and correctly label them. Of course you'll want to check the picture you have is being recognized correctly, and then indexed as such.

- Speech recognition is already around, but by applying that to video, companies hope to be able to discern content without the user creating their own labels. Adding speech to videos, even at a

level where a human can't hear it, but perhaps a computer might has already been suggested by some (black hat leaning) SEOs.

The key here is experimentation, SEO companies in the past have been used to running many different tests to try and predict how a search engine is indexing a page. The lazy ones will ignore this area altogether but a good company will be ahead of the game already and suggest strategies that work with your content and target audience to make sure all the effort that's gone into producing your site is fully utilized.

Chapter 6 – Something Clever

All of the above is excellent advice, I should know, I wrote it.

However there is a topic we haven't discussed yet and without it any SEO campaign is completely worthless; you need to have a metric of how well your campaign is performing.

This metric should **not** be determined by doing one of the following things:

- Paying small monkeys to manually enter terms into Google and hit next page until they find your client's website at position 1337, page 133.

- Continually hammering Google for results with an outdated automated system until your entire office is banned from said search engine. Your colleagues will not be impressed.

- Making up numbers that sound better than the ones you made up last month and sticking them in what you perceive to be a terribly whizzy PowerPoint presentation.

As usual, 33% of you will have totally ignored the bold 'not' above and continue to carry those actions out as gospel.

To the 67% remaining I congratulate you and will now offer you the list of things you will need for your SEO reporting system. Just try to ignore the 'laser' sound effects the person on your left is adding on their next slide:

- The system is going to be polling a large list of search engines to retrieve results. Search engines don't agree with this and often state as much in their Terms of Service agreement. It's possible (but not necessarily acceptable) to use multiple IPs with proxies to retrieve as many listings as you like.

- You need to report on as many search engines as possible. I don't mean Google, Yahoo and MSN. Reports fetched from around 10-30 search engines is a good measure, you should also try to keep your results updated with new engines as they come out.

- Your search reporting software should either be updated extremely often (daily if possible) or adaptive. If the latter then

it should be able to calculate for itself which are the results on a search engine page and never need updating.

- Not only do you want to report on multiple engines, but a long list of keywords as well. Around 200 for each client is a good start, we're going for the PPC style long-tail theory remember so volume is key.

- Your software should ideally output graphs (or charts, depending on your side of the Atlantic). It should do this because it's the single most effective way to see relationships, increases and blips in a campaign. It also saves some poor sod from manually compiling them in Excel.

You can and should use your reporting system to identify how changes you have made to the client site have impacted upon your campaign. If you change a header tag, make a note of it, write down every single thing you do and always put a date and a time. It may be that changes won't appear in your reports system until later on but you should be able to correlate major changes with your graphs to establish your timeframe correctly. Once you have your timeframe you can run minor experiments to see how you've

affected the search engine rankings across the board.

Over time you will learn how different engines respond and can accurately deduce when a campaign will kick into gear. This kind of knowledge is invaluable in online strategy.

Chapter 7 – Conversion

Many websites have a major failing when it comes to SEO; conversion is often ignored with the focus being on increasing the natural traffic to your pages. The conversion rate is the percentage of traffic you actually manage to make a sale from, or a lead depending on your strategy.

Many websites are optimized and then feature no way for the client to tell if they actually have any tangible benefit from all the extra traffic – or if people just go to the site and then leave without ever thinking of it again.

If you're not selling a product, then there are other ways for you to calculate conversions. Track the number of guests that leave comments on articles, how many people vote on a poll or fill in your contact form. Use Google's Analytics, it's a brilliant product and free. Observe how people move throughout the site, if they even do, and try to think of your website like a maze that you want to channel people through to certain, key, targeted pages.

Optimising your website for search engines is one thing – but optimising it for visitors first is far more important. Simple features like fully formed, structured urls allow your users to see at

a glance where they are in relation to the rest of the website. Organising your text with headings and subheadings isn't just a benefit to search engine rankings – it improves the readability of your website.

Drive customers to key areas by utilising clean, text based banners in side columns with a strong call to action. Suggest something that might not otherwise feature in their current thought pattern. Be consistent; if you add a banner on the top right of one page – do it for all – because they'll expect a leading message to be there each time.

If you keep an eye on your Analytics reports you can see which pages are most trafficked – perhaps there's a feature on a certain page which people load up again and again from a bookmark. Don't then be content to let visitors sit on this one page, encourage them to move from that page and check others, perhaps a news section relevant to your industry with continual fresh content which will also help your search engine ranking.

In Future

The most important area of SEO is to stay current. Keep abreast of the latest industry news, follow new technologies and don't think that because somebody says that a method or technology will increase your ranking that it's true.

Perform research, or hire a company that is prepared to explain to you their own.

About

Phill is a director of Grant Midwinter Limited and writes his general search musings down at http://www.surrch.com. He provides the search expertise whilst business partner Derrick Grant is the creative vision.

They both enjoy making pretty websites, which do incredible things.

You can contact us at info@grantmidwinter.com or send us a regular letter (legal threats only please) at:

Grant Midwinter Limited
Calls Landing
36-38 The Calls
Leeds
LS2 7EW