

How to optimize a (Flash) website for search engines

Sonja Duijvesteijn
Communicatie en Multimedia Design



1.0 Introduction	3
2.0 Search engines	4
2.1 What is a search engine?	4
2.2 History of search engines	4
3.0 How does a search engine search?	5
3.1 Crawling	5
3.1.1 Selection policy	5
3.1.2 Revisit policy	6
3.1.3 Politeness policy	6
3.1.4 parallelization policy	6
3.2 Indexing	6
3.3 Query processing	7
4.0 Working of a seo specialist (slecht engels)	8
4.1 Illegal techniques	8
4.1.1 Cloaking	8
4.1.2 Keyword stuffing	10
4.1.3 Landing pages	10
4.2 Off page optimization	10
4.3 On page optimization	10
4.3.1 Titles and urls	10
4.3.2 File size	11
4.3.3 Semantic code	11
4.3.4 Content	11
4.3.5 Meta tags	12
5.0 Flash and seo	13
5.1 Dynamic content	13
5.2 Animations and file size	13
5.3 Flash and links	13
5.4 HTML mirror website	14
6.0 Seo tests	15
6.1 Results	15
6.1.1 Sandboxing	15
6.1.2 No content	15
7.0 Conclusion	16
8.0 literaturelist	17
8.1 Sources	17
Appendix A	18

1.0 Introduction

Whenever a new technology arises, there is a group of early adaptors. They then convince a big crowd to jump on the ship and make it sink. After that, new small ships come out of the blue, and in the end there will be a supertanker which uses all the facets of this new technology.

So, at first there was the internet. It took a long time, but from 1998 on there was a real .com hype. After the bursting of the bubble a lot of money was lost on stock exchanges all around the world and after the predicted demise of the web it is now back. More and more companies invest in their websites. This is an important part in the growing process of a technology.

Now there are literally billions of websites. At first it seemed like enough to just have a website. However, the money put in a company (or other) website should flow back in the company as profit. More and more companies however, see no revenues from their spendings.

According to research about 70% of the time people are online they're trying to find something. So, when you sell cars, you want to be found when someone is looking for cars. Unlike normal advertising, which is, even though for a specific target group, targeted at a very large crowd of not interested viewers. While, advertising for cars when someone is looking for cars does actually get an interested audience. This could mean more sales.

Since about a year there is a lot of emphasis on search engine optimization. This is the process of making a website more accessible to search engine spiders in the hope of getting a higher ranking on the search engine result pages (SERP). And, although a paid advertisement that is targeted correctly is nice, an unpaid genuine search result is even better.

My personal interest in this field comes from my general interest in web development and extensive use of search engines (specifically Google). Over the last year I've found more and more information about seo (search engine optimization). In my work I've taken a lead in the promotion and knowledge of seo.

The problem with seo however is that most knowledge is guesswork. There are a great number of different search engines and they work differently. Also, they are updated frequently, thus what is true today, might not be true tomorrow. In this paper I will look at a number of different techniques for seo and will try to prove or disprove them. Also, I'll give an overview of different search engines, how they (probably) work and why they are important.

2.0 Search engines

2.1 What is a search engine?

A search engine is a program or part of a program that is designed to find something based on the input it gets from the user. This can be either the search function in an operating system, on a mobile device, a corporate intranet, a large website or the entire Internet. For the purpose of this document only search engines are considered which index the Internet, and thus give cross-domain information.

2.2 History of search engines

The Internet started as ARPANET which was a link between the University of California and the Stanford Research Institute. Within weeks two more universities were added. Since that first ARPANET link universities have been guiding the way in the development of new features and technologies. So, it is no surprise that the first real search engine also came from a university.

"Archie", which stands for Archive was created in 1990 by Alan Emtage, a student at McGill University in Montreal. This program downloaded ftp listings and let users search on filenames. However, this was hardly a real search engine. Aliweb (Archie like indexing for the web) let webmasters submit their own websites since 1993, including a description, and searched through that. Though limited, the search was more substantial than just checking the filename.

Then Webcrawler followed, first as a desktop application but from April 1994 as a website. Webcrawler send out hundreds of little spiders (affectionately called Spidey) which read through the entire page, indexing all words on the page. That full text search wasn't unheard of, but for the Internet it was new. With the thousands of websites it would take massive processor power, and storage to index everything and be able to search through it. And of course, Webcrawler also originated at a University.

In January of 1994 Yahoo! was founded (by Stanford graduates) but at first this was just a directory, without search functionality.

In 1998 Google was founded at Stanford University by Larry Page en Sergey Brin. They had an idea that would change the way search engines work (and would earn them millions). They both had parents who were a part of the academic world and as such were familiar with the way the importance of a scientific paper is rated. Depending on how many authors would cite from that paper, or reference it, a rating for the paper itself was made. Assuming that if many scientists cite and reference the paper the paper itself must be good.

They adopted this idea for the web. If a website gets referenced often it must mean it is an important website for that search term. This is the basis of "pagerank" the patented idea that made Google. At its peak in 2004 Google handled 80% of all searches on the Internet. Then, Yahoo choose to use its own search engine instead of continuing to buy Googles results and the market share of Google fell.

3.0 How does a search engine search?

3.1 Crawling

A search engine sends out spiders (occasionally called ants) that spider the web. A spider starts with a list of URL's that it will visit. In that visit the content of the webpage is downloaded, and checked for text that should be indexed. This means that HTML, or other scripting languages are normally not indexed, and the spiders needs to know what to index and what not. The HTML however is important for the next step in crawling. Identifying other links, which haven't been visited yet and visiting those. Because of this an seemingly endless list of URL's is generated, which are all visited by the different spiders.

The number of spiders a search engine has is an important factor in the amount of the web that can be visited in a relative short period of time. However, that's not all that makes a search engine successful at crawling.

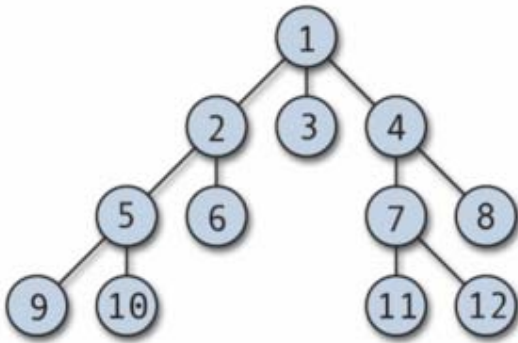
"Given that the bandwidth for conducting crawls is neither infinite nor free it is becoming essential to crawl the Web in a not only scalable, but efficient way if some reasonable measure of quality or freshness is to be maintained." (Edwards et al., 2001).

Because of this there are number of policies that decide which websites and pages within those website to visit.

3.1.1 Selection policy

According to a research by C. Lee Giles only 16% of the internet was indexed in 2003. Since the growth of webpages is enormous it is (at this time) not possible to index all webpages. Therefore a search engine decides which links to visit, and which to put on hold for later (if ever).

There are a number of different strategies that can be followed to determine which websites to visit and which not. Commonly the first rule of every strategy lists a number of resources that should not be spidered. For example, anything ending with .psd is skipped.



After this first step breadth-first might be used. A technique that puts every link it finds in a tree structure. It starts at page one, every link it finds there will be visited and checked for new links. Those new links are then followed, and all new links in that batch are put in batch three. This way no links will be forgotten, but it takes a long time and very irrelevant pages will be indexed as well.

Most modern search engines use a combination of different techniques.

Depending on the number of pages of a website that has been indexed already, the average rank of those pages. The link text, the position in the page, the age of the domain and other factors are weighed to determine the importance of the url. And if all else fails, there's still the breadth first technique.

3.1.2 Revisit policy

When should a page be revisited? Assuming that outdated content can cost customers (people can't find what they want, or don't get the expected results) it is vital for a search engine to have the current version of a website. The question is then, do you want to have a recent version of a page, or do you want to have as little outdated content as possible. The difference being, that in the first case 1 crawl per day might suffice, while in the latter 1 per hour might be better.

There are two ways to handle this:

1. Visit every website with the same frequency. So start at A, and go on to Z, just to start over again.
2. Proportionally. A website that is expected or known to update frequently will be crawled more frequently as well.

Most modern search engines opt for the second choice.

3.1.3 Politeness policy

Not only the search engine itself influences which pages are crawled. Webmasters can also exclude certain documents and directories. With the use of a robots.txt spiders can be told to stay away from certain areas. Before a spider crawls a page in a domain it should always first look for this file to check whether it is allowed to do so. Unfortunately there are also spiders who do not follow this practice. All major search engines however support this protocol.

A good example of a robots.txt is that of the white house, which can be found at <http://www.whitehouse.gov/robots.txt>. It has a list of at this moment almost 650 directories which spiders aren't allowed to crawl.

3.1.4 Parallelization policy

A search engine has numerous spiders crawling the web at the same time. Some of these are bound to find the same URL. To save on downloading and processing time by not crawling the same page twice there is a parallelization policy according to which spiders get URL's to crawl. This is done either dynamically or statically. Statically means, that there is a fixed rule set from the beginning of the crawl that states how new URL's are divided over the different spiders.

Dynamic parallelization is done by a central server that gets all new URL's and determines on, for instance, the load of the spider, which spider crawls what.

3.2 Indexing

How exactly search engines index the web in 2006 is a trade secret. But in a paper¹ by Sergey Brin and Larry Page, the founders of Google from 1998 there is a good overview as to how Google worked at that time, and how they thought it should improve in the future.

When crawling a website new URL's are found, these are rewritten to absolute URL's (so, including the domain name) and given a docID, which makes it easier to keep track of URL's in the system. This URL was found as a link, and as such had an anchor text. This text is put into a forward index. After that, the new docID and the docID of the page where the link was found are put into a database; these are later used to calculate PageRank.

¹ The Anatomy of a Large-Scale Hypertextual Web Search Engine

The basis of indexing is the lexicon, a long list with all known words. The docID's are matched with the lexicon generating an inverted index. This inverted index contains for each word a list of references to all docID's that use that word, and where the word appears in the document. Besides that Google also maintains a 'hitlist' per document.

"A hit list corresponds to a list of occurrences of a particular word in a particular document including position, font, and capitalization information."²

3.3 Query processing

The last important task of a search engine is query processing, which actually exists of two parts. The query returns results with which the search term matches. And those results need to be ranked on the most relevant ones.

When searching for 'communication and multimedia design' the inverted index is consulted, this is already in place. The words 'communication', 'multimedia' and 'design' are consulted to see which document uses all three words. (The word 'and' is skipped as it is too common to be of any value.) The hit list for that document is then considered to calculate a rank for the document regarding that search term. Return to the inverted index to find more results while the total number of results is smaller than an x amount.

It's easy to say that the rank is calculated, but how that is done might be more interesting than the rest of the process. Again, this information is based on the paper by Larry Page and Sergey Brin. They difference between two types of hits, plain and fancy. A fancy hit includes everything occurring in the URL, the title, anchor text and meta tags. Plain hits are basically everything else. Of those plain hits the capitalization, font and position in the document is considered to determine how important it is.

To make sure these aren't tampered with, an algorithm is in place that checks for example how often a word occurs in a document. If a word occurs twice it will rank twice as high on that word, but if it occurs 20 times it might only rank 4 times as high. All variables that are considered for ranking are checked to try to filter out any anomalies.

The most important thing (for Google) are the links that point to a document, and to determine the worth of those links PageRank is used.

² page 9, The Anatomy of a Large-Scale Hypertextual Web Search Engine

4.0 SEO techniques

With the importance of findability rising and more companies expecting a solid business case for their return on investment it is only logical that a new profession like search engine optimization specialist or search engine optimization consultant would emerge. And that is exactly what has happened. These specialists have a range of techniques available to boost the ranking of a website.

Search engines aren't against optimization of pages, in fact they welcome it as long as it means that the true content that is on a page can be found more easily. In practice these legal techniques are divided into two different types. On page and off page optimization. But there is also a number of illegal techniques. These range from simple keyword stuffing to elaborate fake websites to generate more links to theirs.

4.1 Illegal techniques

Illegal techniques are not forbidden by law, but forbidden by the search engines. When a search engine finds that such a technique is used by a website it may be removed from the rankings completely or get a penalty on it's rank and thus drop significantly in the SERPs. In practice this means less hits on that website resulting in less sales and missed profit that might add up to tens of thousands of Euros.

To most search engines, anything that alters the rank for your page, that doesn't involve making more or better content, or is beyond the normal life of a website is illegal. So, getting new referrers to your website is only normal. But getting 500 new refers of domains that are all owned by the same person might very well be illegal.

The problem search engines have with illegal techniques is that it makes it easier for non relevant websites to rank high in the results. The last thing a search engine wants is too irritate it's visitors with wrong results as they might choose to use a different search engine instead.

4.1.1 Cloaking

When a spider visits a website it sends out a signature saying it is a spider. Browsers do this as well, as do mobile devices and other visiting applications. This way a website can send different information to view on a pda for example. But it can also send out specific information for a spider. This practice is called cloaking.

Search engines don't like this as the information they index is not the same as their customers will see. This means that someone looking for 'download e-book' might end up a website that sells regular books only because they told the search engine otherwise.

Frequently the information on the cloaked page will be similar to what can be found on that website, but not on that specific page. When searching however you expect to be brought to a relevant page instantly and are annoyed with the search engine when you get bad results.

A well known example of cloaking gone wrong in the Netherlands was in May of 2004. A number of insurance companies had been using cloaking. Figure 1 shows the website of Amev as a regular visitor would see the homepage. Figure 2 shows the same page, but the version that is send to a spider. Obvious is that there is a lot more text, and a massive amount of links on the left side.

However if you've searched for 'financieel adviseur' and you find the page above you're not finding the information you were looking for. Which is why search engines remove such websites from their rankings.



Figure 1

AMEV, financieel adviseur, verzekering

Mena

[aansprakelijkheidverzekering](#)
[aansprakelijkheidverzekering](#)
[AMEV](#)
[AMEV](#)
[arbeidongeschiktheid](#)
[arbeidongeschiktheid](#)
[autoverzekering](#)
[autoverzekering](#)
[belezen](#)
[belezen](#)
[brandverzekering](#)
[employee benefits](#)
[financieel adviseur](#)
[hypotheekadviseur](#)
[kapitaalverzekering](#)
[lijfschadeverzekering](#)
[lijfschadeverzekering](#)
[mlb-verzekering](#)
[motorverzekering](#)
[motorverzekering](#)
[onverzekering](#)
[onverzekering](#)
[opstalverzekering](#)
[opstalverzekering](#)

Context: home - index.html

AMEV biedt geen online verzekering. AMEV kiest voor het intermediair. Dat houdt in dat wij een AMEV verzekering aanbieden via een onafhankelijke verzekering financieel adviseur zorgt ervoor dat al jouw relevante financiële, juridische en fiscale omstandigheden worden meegenomen in een berekening. Een verzekeringsadvise ingewikkelde producten toelichten. Ben je als zelfstandig ondernemer op zoek naar een mlb-verzekering of een verzekering zelfstandige ondernemer of als particulier aansprakelijkheidsverzekering. AMEV is voor alle markten thuis. En de financieel adviseur helpt jou bij de keuze voor een goede verzekering. Bij AMEV is er mogelijk heel andere wensen. Je kunt bijvoorbeeld niet alleen terecht voor een verzekering, ook geprojectieerde en vernieuwende basisre- en beleggingsproducten worden tot ons je beleggen of juist sparen? De autoverzekering van AMEV is uniek door de wijze van premiehoogte berekenen. Hierbij tellen we ruim twintig persoonlijke factoren in is een eerlijke premie die een stuk stabielder is dan die van andere aanboders.

AMEV: de producten van AMEV koop je bij de financieel adviseur.

AMEV, financieel adviseur, verzekering Heb je een schadeverzekering van AMEV? De saakle schadeafhandeling van AMEV is alom bekend. Maar wist je dat AMEV ruim twintig factoren meetelt bij het vaststellen van de premiehoogte van je motorverzekering?

AMEV, financieel adviseur, verzekering De hypotheekadviseur helpt je een hypotheek voor jouw woning uit te zoeken. Je woning is je thuis. Je basis. Met de wooftha de brandverzekering van AMEV stel je je meest waardevolle bezittingen veilig. Meest er ooit. Bij het plaatsvinden of schade ontstaan, dan komt AMEV snel en volledig een pensioen of een pensioenkort, bent u optimaal de pensioentjeuring. Bijna de pensioentjeuring van AMEV die je voor je medewerkers hebt getroffen, kun je een extra bijpaarmodule openen. Gratis. Een mooie kans om je employee benefits op een eenvoudige manier uit te breiden!

AMEV financieel adviseur

De professionele verzekeringsadviseur neemt alle relevante financiële omstandigheden mee in een berekening. Een aansprakelijkheidsverzekering of een verzekering zelfstandig ondernemer, de financieel adviseur licht kosteloos alle verzekeringsp AMEV toe.

AMEV, financieel adviseur, verzekering *Verder is een pensioentjeuring bij AMEV niet, moet je eerst voor jezelf besliden wat je ermee wilt gaan doen. Wil je een pensioen

Figure 2

4.1.2 Keyword stuffing

Keyword stuffing is one of the oldest techniques in use to improve ranking. By adding a lot of relevant words to a page it's bound to rank higher than without those words. Previously these words would be hidden away in the keywords meta tag. Search engine developers realized this as well and those meta tags are hardly ever used anymore.

This hasn't put an end to this practice however. Keywords are now placed in hidden fields, or with a colour equal to or resembling the background color. This way normal visitors don't notice that they're there, but search engines do index them.

4.1.3 Landing pages

By making a massive amount of portals for a specific topic that all link to each other and most prominently to your own website you can raise the ranking of your own website. Because, when a large number of relevant websites refer to your website, yours must be the best.

This is partially stopped by decreasing the importance of websites/pages with a large number of links and little content, but not completely. The new trick is to display dynamically generated content that is stolen from other relevant websites and add links to that. It is very hard to stop such practices.

4.2 Off page optimization

As seen before an important part of the ranking of a website is based on the number of referrers to that website. So, to increase the findability you need to increase the number of inbound links. By adding your website to portals, to directories and link pages on different (relevant) websites you can increase your ranking.

A good directory to be part of is DMOZ, which uses high standards to determine which websites get added and which don't. A vast number of volunteers goes through the inclusion requests and manually filters out any websites that aren't a good source of original content. Also, Google has a special place in it's algorithm (heart) for DMOZ and websites that are included in DMOZ automatically rank higher.

4.3 On page optimization

Factors outside your own website are much harder to alter so a large part of the optimization process is done on the website itself. The purpose of all these techniques is to make it easier for the search engine to find the actual content of the website, and to determine what is most important on a page.

4.3.1 Titles and urls

The title of a page is (partially) shown in the taskbar in Windows and completely in the top bar of the browser window. Since this is always seen by users and can't be hidden it is assumed that whatever is in the title must be important, and should give a clue about the content of the page itself. So, all words in the title carry extra value in giving the page a ranking.

By making sure a title fits a specific page instead of giving all pages of a website the same title you can improve your ranking. And why would the title of your page be 'welcome to the website of Communication and Multimedia Design' when you can use 'Communication and multimedia design, minor multimedia engineering, admittance'.

Also, most search engines show the title of a page on the result pages, so a descriptive title will also tempt people to actually click on the link.

The same goes partially for the url of a page. It is a direct link to the information on that page and does generally show the directory structure of the website. So, a logical build up for an url could include domain name, category and subcategory. These all give important clues about the information to be found on that pages. For example 'http://www.multimedia-engineer.nl/2005/css/css3_selectors' implies that the information on that page is about multimedia engineering, was written in 2005, is about CSS, and more specifically, about CSS 3 selectors.

4.3.2 File size

Content at the bottom of a page is read less often than content at the top of the page, basically anything that needs scrolling is hidden away for most users. Page length also implies file size. And this is the easiest way for search engines to measure length. As such, content that is further down in the source code, is considered less important.

Unfortunately there are some web development techniques that increase the number of bytes used. One of the first things a seo specialist will do is rebuild a website in such a way that the total page length (in bytes) is decreased.

Tables used for layout are infamous for bloating code. But also inline CSS and Javascript can increase the file size significantly. By putting all CSS and Javascript in external files the total file size is reduced. And this reduced file size puts the content of the page higher in the file, resulting in a (slightly) better ranking.

4.3.3 Semantic code

Code itself can give clues as to what content is important and which is not. The tags h1 to h6 for example are used for headings. So the information between those tags must be more important than other information on the page. But also (emphasis) gives clues as does (bold) and <i> (italics).

By using the correct building blocks in HTML you can easily point the search engine in the right direction. The only thing you still need to do is make sure your content is good.

4.3.4 Content

Good content is the most important and most fail proof way to get good ranking. No matter what other tricks you use to get a higher ranking, or how many links you have to your website. If the content on your website isn't good visitors will not stay on your website. And your ranking will drop slowly.

Also, with good content a lot of your ranking will develop naturally. People will link to your website because they found something they want to share. Also, search engines index your content, and by having good content you can be sure of a good ranking in the future when the way search engine index changes.

So what is good content? First of all, it is spelled correctly. A search query consists of small number of words, and you should rank high for those words. If you've misspelled a word they won't find your website anymore. Obviously, if you use 'auto' 5 times on your website, and once as 'uato' you can still be found, but will rank slightly lower as you're less relevant.

Good content also uses keywords; an example is the homepage of a large number of companies.

‘Welcome to the website of <companyName>. We are a company that thinks that service is one of the most important parts in a professional relationship and as such we offer that. And besides this professional attitude we also excel in creativity and we value the customer highly”.

This must be a great company, but they forgot to mention what it is they actually do. Since people search for a specific branch or service this company will not be found.

4.3.5 Meta tags

There are a number of different attributes you can use in a meta tag that you can use to give information about your website. The most (mis)used of these are the description and keywords attribute. Most search engines do not use these meta tags any more, but some, like Ilse still, do.

In the keywords meta tag you can put up to 20 words to describe the content of your page. The description meta tag is used for a short description of your website. Some search engines show this text in their result pages. Although meta tags don't help your ranking much, every little bit helps.

There is also a meta tag, which is meant for search engines 'revisit-after'. With this you can tell search engines when to revisit your website and index it again. However, no major search engine actually supports it. The only known search engine that did support it was searchBC, a search engine specifically for Vancouver.

5.0 Flash and seo

Normal websites exist of a number of plain text pages. This makes it easy to index them as all the needed information can be extracted out of the document easily. That is not true for Flash content, or even complete Flash websites. The problem here is that all recognizable content has been compiled into an .swf and it's hard work to filter it out.

Adobe, formerly known as Macromedia recognized this problem as well, and they came with a software developers kit to decompile a Flash movie in such a way that only searchable information would be left. Search engines like Yahoo and Google have now adopted this technology making Flash movies better searchable.

However it is still rare to find a flash movie in the search results. Which must mean that it is still hard to rank well with one. But why is that? Keep in mind that the following information is mostly based on guesswork of the seo community, how exactly a search engine works is a trade secret and such guesswork is the best there is.

5.1 Dynamic content

Generally when an entire website is made in Flash not all information is actually put into the .swf, this would be too big to download. This means that using actionscript external content will be loaded into the Flash movie to show the information.

This is a problem since the search engine tries to filter out any information that is needed for indexing. But it will only find new url's to spider, not actual content. These new url's might be spidered depending on the selection policy, but that will hardly help the .swf's ranking.

5.2 Animations and file size

Another problem with indexing Flash movies is that not all content is readable. Part of the information might be stored in animations or in images. A motion tween (movement from point a to point b) should be interpreted as normal text, but a shape tween (text changes into a different shape/text) poses a bigger challenge.

Also images are a problem in Flash movies. In HTML you can add an alt(ernative) attribute to an image to convey it's meaning in words. No such thing in Flash, so if you draw an icon of a house to represent 'home' the meaning will be lost to search engines completely.

Also, search engines only index a Flash movie until a certain file size. The theory behind that being that most visitors won't wait to see a 20 second intro, so neither should they. However a visitor could decide to skip to a different part of the movie early on. A search engine indexes everything sequentially. Meaning it might miss important information your other visitors do get to see.

5.3 Flash and links

A different problem with Flash movies is found in the way most modern search engines rank websites. They check to see how often a website is referred to, and whether the topic on that page matches that of the referring page. So, a page about cars that links to a page about vegetables will probably not help each other to rank well.

Flash content is normally put into a HTML page from where the movie is loaded. But, the HTML itself frequently does not contain any information about the subject. Someone referring to that Flash movie however gives a link to the .HTML page instead of the Flash movie.

The result of that is that the content and the relevant link (which would boost the findability) aren't on the same target. This makes the swf loose a lot of potential ranking making it harder to find. And with over 100.000 websites on most subjects the loss of only a small bit of ranking is essential and might push your content back pages on the SERP's.

5.4 HTML mirror website

The only effective way to make a Flash website findable at this moment is by building the website in HTML as well as in Flash. This website should show the same content as the flash movie, this make sure your website isn't seen as cloaking. But it also offers the content to visitors that do not have a Flash plug-in.

By adding a small bit of Javascript it is possible to check whether a visitor has the Flash plug-in installed. If that is the case you replace the content with the Flash movie. You get the good indexability of the HTML website and the rich experience of Flash, but it does take a lot of extra work.

When the movie I robot came out in 2004 a website for this movie was made entirely in Flash. However, the first 3 months the official website didn't show up in the result pages. Only after they added a HTML version of the website the website would show up in the search engines.

6.0 Seo tests

Most information on search engine optimization is based on practice. By looking at websites that rank well, and copying what they are doing. But there is hardly any research that actually examines the effect of specific techniques. The only research available of that type focuses on optimization of HTML websites, while with the number of rich internet applications growing, the number of Flash websites is also growing.

So to find out how to make Flash websites rank better a number of different pages³ were put up that use different techniques. A number of 14 websites have linked to the index page of this research continually. These pages had been online for 3 months at the time a conclusions has been drawn from them. As it takes some time before a page is indexed, and then some more time before it is given it's final ranking.

6.1 Results

To find out how the different pages rank on search engines I've compared a number of different search engines, To be precise: Google, Altavista, Yahoo, Ask (previously Ask Jeeves), MSN, Overture, Excite, Lycos and Ilse.

The main result of this research is that Flash is still indexed very poorly by search engines. Some search engines don't index any Flash content at all, these include: Ilse, Lycos, and Overture.

Of the search engines that do index flash movies none show the movies included in the test after three months. Access logs of the particular pages do show they have been visited by a number of search engines. That means, although the search engines know of the existence of the pages they see them as of so little importance they do not show up in the results.

There are a number of possible reasons for this, the pages could've been placed in the sandbox or the content of the movies might not have been found

6.1.1 Sandboxing

Most search engines use a technique called sandboxing. This means that a new website or page isn't put in the main index right away but is first placed in a temporary environment. This is done to limit the use of quick fix black hat seo practices. So, this way it would take months longer before landing pages could be effectively used. And something that takes months to accomplish is less likely to be abused.

Unfortunately it also means that it can take months, and even up to a year for new websites to show up in the search result pages.

6.1.2 No content

Part of this research was in order to see which techniques used in flash allow the content to be indexed. Therefore it is likely that at least some of the flash movies have, according to the search engines no content. And, something without content can't be found.

But, at least some of the flash movies have content that can be extracted using the appropriate software.

³ The complete setup of this research can be found in appendix A

7.0 Summary and conclusion

Search engines have been in use since 1990, and ever since the number of people using them have grown significantly. With the growth of the customer base of search engines, the importance of showing up high in the result pages has also grown. Some new business only get their revenues from being found in search engines so for them ranking high is incredibly important.

Thus it is needed to understand how search engines work and rank websites, and how to influence that ranking. How exactly the different search engines work is a trade secret but some of the working is clear. Most modern search engines are based on the ranking system for scientific papers. The more frequent a paper is referred to in other scientific papers the higher it will rank itself.

This same method can be applied to websites/pages. By calculating how often a webpage is referred to, by websites about the same topic, a ranking for that topic can be found. So, generally speaking the more (quality) links to a page, the higher it will rank.

But there's more to seo than external links, on the page itself some changes can be made to improve the ranking of a page. These all come down to making the content better, and fixing the code around the content to give extra clues about it's importance.

At least that's how it works for HTML pages. But Flash is a major factor on the web and complete websites are build in it. So optimization for Flash websites is starting to be more relevant as well. Unfortunately, due to Flash movies being compiled into a .swf it is hard to get the content out again. And search engines rely on that content to index the file correctly.

Fortunately there is a way to get most of the content out of the Flash movie and some search engines do this, but not all. And even on those that do, Flash movies don't rank well. For any commercial website it would be unwise to only have a Flash version without an accompanying HTML website as this would significantly decreases revenues.

Generally speaking you should not build a commercial website in flash but in HTML. And in the HTML, content is most important, but after making good content it is best to spend your time on generating inbound links.

8.0 literaturelist

NPD Search and Portal website Study

Danny Sullivan, <http://searchenginewatch.com/sereport/article.php/2162791>

Evolving strategies for focused crawling

C. Lee Giles, 2003, <http://clgiles.ist.psu.edu/papers/ICML-2003-focused-crawling.pdf>

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin, Lawrence Page, 1998,

<http://www.public.asu.edu/~ychen127/cse591f05/anatomy.pdf>

Google Het verhaal achter het media succes

David A. Vise, Mark Malseed, 2005, ISBN 90 225 4365 X

8.1 Sources

http://en.wikipedia.org/wiki/Search_engines

<http://www.webcrawler.com/info.wbcrawl/search/help/about.htm>

http://en.wikipedia.org/wiki/Web_crawling

http://www.google.com/newsletter/librarian/librarian_2005_12/article1.HTML

<http://searchenginewatch.com/>

Appendix A

Goal of the test	NR	Description	link name	Text
Is the content of an swf counted to the content of the html page? And does the html with swf score higher, or the loose swf.	1a	Link naar html met swf	test 1	Litera termani
	1b	Link naar swf	test 1	Litera termani
Is dynamic content indexed in search engines and how does this effect the ranking?	2a	swf with text on stage	test 2	Crocidus lavus
	2b	swf with text from variable	test 2	Crocidus lavus
	2c	Swf with dynamicly loaded text from a static link.	test 2	Crocidus lavus
	2d	Swf with dynamicly loaded text from a dynamic link.	test 2	Crocidus lavus
Can you tween text and still keep it findable in search engines?	3a	Swf with regular text	test 3	Achtus remisii
	3b	swf with motion tweened text	test 3	Achtus remisii
	3c	swf with shape tweened text	test 3	Achtus remisii
How much text is index in swf's?	4a	swf with 5mb of different text	test 4	No specific search frase
Does a meta tag increase the ranking of a page?	5a	keywords meta tag	test 5	Nalatera kadius
	5b	description meta tag	test 5	Nalatera kadius
	5c	description meta tag	test 5	Nalatera kadius
	5d	no meta tags	test 5	Nalatera kadius