Università degli Studi di Genova

Facoltà di Ingegneria Informatica

Search Engine Optimization: a theoretical and practical approach

Corso: Basi di dati 2

Docente: Prof. Antonio Boccalatte

Supervisore: Dott. Christian Vecchiola

l.i.d.o. - DIST

http://www.lido.dist.unige.it

Allievo: Luca Vassalli

Indice

Introduction

My project for the course of Basi di Dati 2 consisted in writing an essay about the Search Engine Optimization and applying the aspects I deepened in the research to a real world website.

In the first part of this relation I will present what is implied in the optimization of a website for a Search Engine. I will explain why every webmaster should care about the ranking of its website in the major search engines results list; which ones are the means, and sometimes the tricks, to get a search engine friendly website and which ones are the drawbacks and the risks of a bad optimization.

In the second part I will present how I optimized a real world website applying the suggestions presented in the first part. I will explain why I used some tricks and why I did not use others. The website is called AgentService; it is developed with dynamic ASP pages and is about an agent oriented programming framework.

Search Engine Optimization

Introduction

Today the amount of information, we can get from the Web, is increasing more and more. Although it is very useful that everyone can retrieve all the information he needs in his work using just an Internet connection, on the other hand this requires a quick and efficient way to find what actually he is looking for.

Search engines are the world famous solution to the problem; they are an interface to the information: transcending user categories, geographic regions, and information seeking goals, they allow us to find in few minutes the main websites related to the argument we are interested in. Unfortunately the task to present us only the pages which are relevant to the keywords that you type is not trivial, it requires to crawl all the Web, index all the pages according to their content and then to rank these page to present them to the user in the correspondent order. In particular the critical part is when a search engine has to assign how much a page is relevantly related to a particular keyword; understanding the rules and algorithms used in this process is fundamental to increase the chances of gaining position in the ranking order of your website. The higher is this position and the greater is the probability that someone will arrive to your website through a search engines.

Search Engine Optimization (SEO) is the activity of optimizing a website in order to make it more search engine-friendly. Already in the 2003 these and directories average over 300 million searches per day and it was showed that the users usually trust, and consequently buy, more from sites found using a search engines rather than a banner advertisement[1]. Therefore it is useless to create a very functional and beautiful website which has a poor rating like it does not make sense to open a new shop without registering it with a phone book, since to contact or find it will be difficult.

Unfortunately, although you will follow all the basic rules of the SEO, that does not automatically guarantee top ranking. The main reason of this situation is that there are lots of pages which try to get the highest ranks with over optimization, so that the search engines are very careful in the ranking process; they try to detect suspicious over optimized sites which use the so-called black hat SEO techniques like hidden text, cloaking, doorway pages, keywords stuffing and the like.

Ethical SEO

If you look for information about SEO in the Web you will have no problem to find an endless list of suggestions, tools and examples about the argument. But looking with more attention you will be soon able to divide into two main groups the techniques

[1]"http://www.informit.com/articles/article.asp?p=31193&seqNum=2" extracted from "Search Engine Visibility" by Shari Thurow, ed. 2003

indicated: a first one which consists in a group of more or less basic SEO techniques based in the way the genuine content, that your website is already made of, should be enhanced from the SEO point of view; and a second one which consists in using some advanced means to present in a different way the website, and the pages themselves, to the human users or to the search engines.

In the second case for instance, pages are added containing links to an important page, or hidden text is added to a page; this content is not shown to the human users because it is meaningless but it can give a further boost to the rank position of the website in particular for very competitive search terms. The peculiarity of the this group of techniques is that, if a search engine spotted them, your website would lose position in the result list and in the worst case your website might be banned. It is true that also exaggerating in the application of the first group can be dangerous, but it is quite clear that the techniques of the second group are much more dangerous.

The techniques of the first group are usually called "ethical SEO" and the techniques of the second group SEO spam or black hat SEO techniques. Phil Craven[2], an influential SEO expert, in a series of articles points out his position. He claims that for the search engines SEO includes only: writing copy, giving advice on site architecture and helping to find relevant directories to which a site can be submitted. It means that for search engines both the ethical and the unethical techniques are deprecated, that is the reason you have to carefully optimize also using ethical SEO. Anyway it also true that a search engine cannot show all the relevant information at the top of the ranking, SEO spam are powerful techniques to achieve that your website, which maybe is relevant as much as others, will be considered relevant enough by the search engine to appear at the top of the ranking.

[2]Search Engine Optimization (SEO) and Google Phil Craven http://www.webworkshop.net/seo_and_google.html "Ethical" Search Engine Optimization Exposed! Phil Craven http://www.webworkshop.net/ethical-search-engine-optimization.html

Beside the ethical issues that everyone can deepen as much as he wants, it appears clear that as far as you have to optimize a website to a two or three word search sentences where there is not too much competition you can just use the ethical SEO, described in the chapter two of this paper, which is less dangerous, easier and faster. But whenever you will need to optimize a website for high competitive two word searches or, even worse, common one word searches you will need the SEO spam techniques, described in the chapter four.

I will also speak, in the chapter three, about the way a search engine sees your website, which affects the way you should develop it. Finally I will analyse other general topics about optimization under particular circumstances.

If it is true that you have to carefully optimize to not be banned, it is also sure that not following the main SEO rules will lead to very poor ranking positions, but, to understand how you can optimize a website, it is before necessary to analyse how a search engine actually works.

Optimization techniques

How a Search Engine actually works

Search engines use automated software programs, known as spiders or crawlers, to explore the Web and build their databases; a spider analyses page after page, following the links, which connect them one another. Then every page is added to a giant database, sometimes called the catalog, indexing it with its major keywords. A keyword is a word, which is relevant to the content of the page itself, or it is an expression that describes it. Usually the spider starts its research from the most relevant pages already in the database, so that pages already highly ranked are examined more frequently. When a new site is created the webmaster should submit a form to the major search engines to speed up the time to be included in the result list for the relevant searches. Anyway since the Web include a pretty fair amount of pages, if you changed something in your site to increase its ranking, it might take a month or even more the effects to be seen in its ranking.

After a page is in the database it is ready to be presented in a search, as a result every time a query is made, there is a program inside the search engine, which sifts through the millions of pages, recorded in the index, to find matches to the search.

The final step is retrieving the results: after a user makes a query the indexed pages are present to him in the ranking order, it appears already clear that the title of the page, which will be shown in the list of results, is very important in the ranking of the page, so a page is penalized if title and content mismatch. There are several different ranking algorithms used by different search engines, in fact the same query on different engines will usually have a different result. These algorithms are kept secret to avoid that studying them will help to find an easy way to get better position in the result list, so, if the general rules we will consider further on, can be considered always valid, you should keep in mind that achieving the top ranks requires optimizing in different ways for different search engines. For instance roughly Google considers more important the links whilst Yahoo the keywords; links and keywords are in somehow relevant in both the search engines but with a different weight, thus what may be a good keywords optimization for Yahoo, will probably be over optimization for Google.

I have just described how the so called crawler based search engine works; actually there are at least other two kinds of source of information: the human-powered directories and the Hybrid Search Engines. In the first case the page are inserted in the database by humans, consequently the SEO rules has no power to increase the ranking which is driven by the quality of the websites; an example is Yahoo directory. In the second case, like suggested by the name, both the mechanisms can be used. For instance in the 2002 MSN Search presented human-powered listings from LookSmart but it also used a crawled based search for the more obscure queries[3]. However this hybrid category is now disappearing.

Nowadays the three main search engines are Google, Yahoo and MSN and they are all crawler based. During July 2006, it was calculated that about 50% of the global searches were run by Google; about the 25% by Yahoo and 10% by MSN. Therefore thus the space for the other dozens of search engines is really thin. Google is considered the outstanding best engine, no surprise if since it was born its ranking algorithm influenced forever the ranking algorithms of all the other search engines, therefore no dissertation about the working mechanisms of a search engine can avoid to speak about it.

Google

Google runs over a distributed network of thousands of low cost computers. The parallel execution is the key of the fast processing of the queries. The Google's spider is called Googlebot and it is made up several computers requesting continuously pages, which are downloaded and then passed to the Google's indexer. There are two different kind of crawling techniques: in the deep crawling, Googlebot fetches a page and then it creates a list of all the links connected to that page, it is a good way to retrieve all the pages of the same website; on the contrary, during fresh crawling, popular frequently changing pages, for example those from websites of newspapers, are crawled at a rate roughly proportional to how often the pages change.

The indexer store all the text that was crawled in the database; then the index is sorted alphabetically by search terms, with each index entry storing a list of documents in which the particular term appears and the location within the text where it occurs. Like in all the major search engines, the common words, called stop words, are ignored to save

[3]How Search Engines Work" (2002) By Danny Sullivan http://searchenginewatch.com/webmasters/article.php/2168031

disk space and speed up search results. Since they would be present in the majority of the pages, they are not useful to narrow the list of results. Some stop words are for instance: "how", "when", "where", "the", "a" and any characters.

In the Google ranking system there are hundreds of factors involved but the main one is based on the so called PageRank algorithm which considers inbound and outbound links, I will analyse in deep this algorithm further on. The ranking is one of the functions of the Google's query processor; another is to apply the spelling correcting system. This is one of the ways Google uses machine learning techniques to automatically learn relationships and associations within the stored data. In the spelling correcting system, the search engine checks if you have typed an orthographically wrong word or if there is a similar sentence with a much longer number of results[4].

Approximately once a month the Google's index is updated: the PageRank of all the crawled pages is calculated and since this requires on average 40 iterations of the algorithm, the calculations take several days to complete. During these days the ranking positions of the pages fluctuate; sometimes minute-by minute, that is the why it is nicknamed Google's Dance.

There is another circumstance during which Google dances: the Updates. It is called Update every time the Google's ranking algorithm is changed to improve the service to the users. In particular every change has the aim to spot better and better SEO spam techniques. Those are particularly dangerous for a search engine when are used in unethical and unscrupulous ways to mislead the search engine and let it thinks a website has a different relevant content.

The first Update was the world famous "Florida Update" (November 2003). The Austin, Brandy, Bourbon and Jagger updates followed it. But the huge one was the first one. It took eight weeks to re-establish stable listings; a quick search over the Internet is

[4]"How Google works" (2006) Google Guide http://www.googleguide.com/google_works.html

enough to see how many upset webmasters were involved in this change. The update consisted in the introduction of semantic contextualization software and a variation of the Hilltop Expert Document Algorithm.

The semantic components were quite mild, it is general belief that Google is moving to a semantic-type system but also nowadays you still cannot write "glass, nails, and wood" hoping to find windows.

The Hilltop was more important; it cut the importance of the reciprocal links. Since it was understood the importance of the inbound links for the PageRank, lots of website started to exchange links. After the Update the weight of those links was dramatically reduced.

Perhaps the greatest effect of the Florida Update was the sudden rise in the popularity of blogs. The reason is that all the owners of a blog have a link on the website to their blog entry, and these are not reciprocal links. Every link increases of a bit the PageRank, when the blogs are thousands, the ranking of the website will go sky high. By the way, Google owns the main blog software developer, Blogger[5].

The last change in the Google search engines is called Bigdaddy (February 2006) but it was not a real update, like the Google engineer Matt Cutts said: "in this case the changes are relatively subtle (less ranking changes and more infrastructure changes). Most of the changes are under the hood, and this infrastructure prepares the framework for future improvements throughout the year". Anyway in other threads of his forum Cutts said about a site: "Some one sent in a health care directory domain. It seems like a fine site, and it's not linking to anything junky. But it only has six links to the entire domain. With that few links, I can believe that out toward the edge of the crawl, we would index fewer pages." It means that links into and out of a site are being used to determine how many pages of the site should be included in the index. Since the Web is

[5]"Florida Update One Year Later - The Year Google Grew Up" By Jim Hedger (December 2, 2004)
http://www.search-this.com/google/florida_update_one_year_later_the_year_google_grew_up_Print.aspx

exponentially growing day after day, not all the pages can be indexed and kept in a distributed database. In order to keep in the index more and more websites, the idea is to index much more pages for the major websites and just few pages for the minor ones. Those will be at least represented by their home page and some other major pages. This is a solution to a problem that may cause to have only big websites indexed and small or young websites not indexed at all. In this change a new attempt to fight the reciprocal links was made, in particular links which connect not related websites are now considered junk, Cutts makes the example with free ring tones website and Omega 3 fish oil site.

Although it was from February that Google was working on BigDaddy, during the summer 2006 the new system did not seem stable yet, there were still lots of complains about duplication of results, non existing links and different results only after few seconds with the same search term[6].

Now let us analyse the most important aspect of the optimization from a Google standpoint.

Links

When you think to the Web you think to a jungle of pages interconnected one another by links. If you create a website with few links to the external world, not only it will be hardly found by a spider but even when it will be eventually crawled for sure the assigned rank will be pretty poor.

You can distinguish two different kinds of links in a page: an inbound link, aka backlink, is a link from another page to this one; an outbound link is on the contrary a link that points from the page to another one. Backlinks are the important ones because the webmaster has little control on them, there are some way they can be faked but it is

[6]"Feedback on Bigdaddy data center" Matt Cutts, January 2006 http://mattcutts.com/blog/bigdaddy/

definitely more difficult than faking a metatag or keywords; in particular inbound links are fundamental for the PageRank[7] algorithm used by Google.

When Google arrived with its link-based PageRank, link popularity took off and became an absolute essential ingredient in achieving top rankings. In particular the idea behind PageRank is that is not only important how many inbound links a page has but where these links come from, in particular a website which is pointed only from other websites with a poor PageRank will have a low PageRank as well, but if it is pointed by even a single link which comes from the NASA website it is likely it will have a great rank. Actually you can consider an external link like a vote to the pointed page quality, this is the reason that an external link to a website guilty of web spam is punished with a drop in the ranking: it may be not your fault if you have a backlink from a bad website but you have to control the websites you link to. Every page has a PageRank, which is a numeric value to represent the importance of that page; the higher the PageRank is, and the more important are the votes to the other websites.

To calculate roughly the estimated value of a page you can use the following equation:

$PR(A) = (1-d) + d(PR(t1)/C(t1) + ... + PR(tn)/C(tn))$

This equation was published when the PageRank was being developed the first time. Obviously Google uses a variation of it but anyway this one is good enough to understand the underlying mechanism.

In the equation PR(A) means the PageRank of the page A; t1…tn are the backlinks of the page A, C is the number of links that the correspondent page has and d is a damping factor usually set to 0,85. Like you can see from the equation, the more links a page has the less value this vote gives to the pointed page, hence, when you check if

[7]"Google's PageRank explained and how to make the most of it" by Phil Craven

http://www.webworkshop.net/pagerank.html your page has a fair number of inbound links, you should also consider the number of other links in the page the backlinks start from.

If you check the Google toolbar you will find a value from 0 to 10 for the PageRank but this is like a label, which says which interval the PageRank of the page is in. In fact the billions of pages on the web average out to a PageRank of 1.0 per page, the difference between the page with the highest rank and the one with the lowest one is divided in ten intervals and these are the ten values which are shown in the Google toolbar. How these ten intervals are divided is unknown, but there is confidence among the experts that divisions are based on a logarithmic scale or something very similar, one of the possible proofs of this is that it is much harder to move up a toolbar point at the higher end than it is at the lower end. This means that is much better having a backlink from a page with PageRank of 8 where there are other 20 or 30 links than having an inbound link from a page with PageRank of 4 where there are only other 5 links.

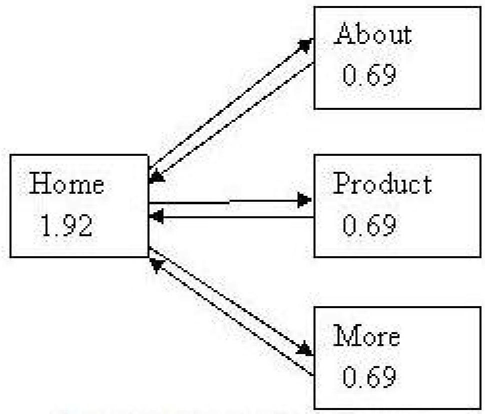
So far the lesson is that every inbound link will increase the ranking of your website.

Unfortunately this is not always true. In case your website has backlinks from a link farm you risk be punishing or banning from Google. We generally refer to link farms when there are a group of web pages which are created only to provide links to a group of websites, so that a single webmaster, who was paid, or a group of webmasters, who agreed to promote each others, put one or more link pages in their website. Since search engines cannot understand the reasons of the great number of links with any content in the page they usually exclude the websites from the index or decrease their ranking. An example is the website "LinksToYou.com".

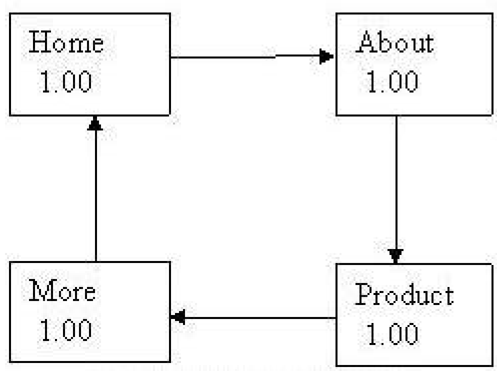
Going back to the algorithm behind the PageRank, it is interesting to notice that the calculation of the PageRank must be done iteratively. You can imagine two page A and B, which are connected each other and no other page points to them. The problem is that every time a PageRank is calculated the old value is thrown away, accordingly, if you want to calculate the value of A, you need to consider its backlinks, but its unique backlink is from B and the new B's PageRank has not be calculated yet so, the new value of the A's PageRank will be inaccurate, and when the B's PageRank will be calculated, for the same reason, also its value will be inaccurate. Like Phil Craven showed in his famous paper "Google's PageRank explained" the solution is possible when you calculate iteratively the value of the PageRank using the above equation and a starting value of 1 for a page, which has no PageRank yet. It was showed that 40 or 50 iterations are enough to reach a point where any further iteration does not change the solution in a meaningful way. This is the reason of the Google's Dance: it takes a lot of time and work to calculate the final value of the PageRank, during the first iterations the value is not stable and subsequently, even for the same search terms, the result list may change after few minutes.

Another thing to be remembered is that the results you get are proportion among the different values of the PageRank of the different pages, the real value of the PageRank is found using a scale, known only to Google. However the results are interesting to consider the effects of inbound and outbound links in the overall ranking of a page. Every website can have a maximum PageRank equals to its number of pages. That value can be only increased with backlinks from external websites. Using the equation we have seen before you can calculate the effects of the internal links to the PageRank of every page. There are tools on the Web, like the Phil Craven's PageRank Calculator, that can be used to simulate how the PageRank changes adding or eliminating links from the page [A]. A link to an external website will usually decrease the overall average PageRank of your website. In fact the vote of the page is divided among the links; if you have, on the same page, links both to other pages of the same website and to another website, only part of the page vote weight is kept inside the site.

It is obvious that if the number of pages increases also the maximum PageRank will increase, but you should be careful adding new pages with a similar content or you can be penalized. In general a good idea is trying to concentrate the PageRank in a few pages, to have at least an "entry point" to your website which is well ranked, leaving all the other pages with a poor rank. Let us consider the situation of Picture 1 and Picture 2.



Picture 1: Hierarchy structure



Picture 2: Loop structure

Like you can see from the reported value of the PageRank under the name of the page, in the hierarchy structure the rank of the home page is much higher. Although in both cases you achieve the maximum PageRank for the site (4 since there are four pages), only in the first structure you will have an highly ranked entry point to the website. To see the effects of outbound links, think that in the hierarchy structure, in that case if you added a outbound link from "Product" to an external site, the home page's rank would drop to 1,18.

You can waste PageRank also using dangling links, which are links to a page with no outbound links or to a page that Google has not indexed yet. In this case Google removes the link shortly after the start of the calculations and consider it again shortly before the calculations are finished as a result the link affects only a little the PageRank of the other pages. In the case of picture 1, if "Product" had no link to the home page but only a backlink from it, the PageRank would become:

Home	1,46
About	0,77
Product	0,15
More	0,77

This is a good reason to always link a page at least with another one. It is also important that any inbound link points to the pages where you want a high value of PageRank and not to any random page because, in the latter case, the benefit will be shared among all the pages of the website.

You have just seen how much the outbound links are a drain for on a site's total PageRank. In general also for all the other search engines, it is not such a good idea having a page where the outbound links outnumber the inbound links. So you can decide to hide the outbound links to the spiders. You can do this using a JavaScript or using a particular attribute of the anchor tag.

An example of the first case is:

```
<a href="javascript:goto('http://www.hiddensite.com')"> link text </a>
```

To be totally sure that the spider will not see the link, the JavaScript code may be loaded from a file which is in a directory that is barred to Google's spider by the Robots.txt file instead of being inside the tag.

An example of the second solution is:

link text

Google and other search engines recognize this attribute since January 2005 and it tells to the spider to not consider the link at all for the ranking of the destination page, or the page itself.

Another reason you may want Google not consider a link is that it is connected to a “bad neighbour”. In fact if there is another website which you want to link to, but it has a bad reputation, for example is banned from the Google’s index, you should use a link inside a JavaScript; the reason is that search engines do not penalize your website if it has inbound links from suspicious pages, maybe over optimized ones, because you do not have much control on the other webmasters’ actions but it is up to you avoiding outbound links to bad sites. The concept is similar to real life if you choose bad guys for friends, you are considered to be one of them. For this reason all the outbound links which can be view by a spider should be carefully analysed also by the webmaster to check if they link to a banned website. Verifying this is trivial, you just need to write “site: nameOfTheSite” on a search and if the website does not appear in the list of the results it means that it is banned. In case your website has too many links to manually verify all of them, automatic tools exist which can help you [B].

Now let us consider the case where you want that the spider will not read a whole page, or directory, and not just it will not index the page using a link from another page, what could you do? You can use the “Robots.txt” file. This is a file where is explained which pages are supposed to be analysed and which ones are not. It is not a mean to hide sensitive data since it does not protect the website like a firewall or a password; in fact the crawler may decide to ignore the file but, since it usually will not, this file is useful every time you have a page in two different version, for instance one for printing and one for visualization, to avoid to be banned for duplicate content by the search engine. However you should understand the difference between "to index" and "to read": using the Robots file you can prevent a spider from reading, or analysing, a page so that it will not, for instance, calculate the keywords of the page but you cannot prevent to index the page, if it has a backlink from a page which is not protected by a Robots file. In fact, in the latter case, the address of the page will be present in the database of the search engine among the pages connected to the one that was analysed.

To use a Robots.txt you have to follow some rules. First of all it has to be stored in the main directory, because the crawler will look for it only there. If nothing is found at that address, which can be for instance "http://mydomain.com/robots.txt", the crawler will assume that the Robot.txt is not used and it will read all the pages of the website. Then you have to follow a protocol: it is a text format where is written for every kind of crawler which pages are to be ignored. The structure is the following:

User-agent:

Disallow:

Where the "User-agent" is the name of the crawler (i.e. Googlebot for Google, the character * means everyone), and "Disallow" is the directory or the page that should not be read.

Since the smallest grammatical mistake is enough to prevent the spider to understand your indications, there are several tools to validate the file or even to automatically generate it using a graphical approach where you have just to point and select which files and folders have to be excluded [C].

Like we have already seen, the inbound links are fundamental for the ranking position of your website: the more backlinks a page has, the higher ranking it will receive. However, the quality of the links matters even more than the quantity. A link of high quality is always more important than a bunch of links from a link farm. The quality of a link is determined by different factors: the ranking of the page it comes from, the relevance of the content of the two connected websites and how the anchor text of the backlink incorporates keywords relating to your site. In particular about the last point you may need to contact another webmaster to suggest the best anchor text that he should use to link your website, in particular it is a bad habit using "click here" instead of a meaningful sentence. It is enlightening the "Google bombing" practice. A group of people linked many pages to particular target page and all used the same link text phrase. The target page was nothing to do with the link text phrase. It did not contain it and it was not about it. Nevertheless, the page was at the top position in Google's result list for the link text phrase[8].

You can retrieve good inbound links in some different ways.

First of all, you can use a forum on a website with a good ranking and with a content which is relevant to your website. You need only to write a post in a topic and to place a link to your site in the signature line. However before spending time posting in lots of website, you must check the Robots.txt to see if the forum is readable by the spider and if the links in the posts are not hidden, for instance displayed inside a JavaScript.

Then you may submit an email to a directory like Yahoo and DMOZ. Unfortunately these directories are managed by hand instead of automatic software like the spiders, thus you will have to wait a while before your pages will be included. Furthermore DMOZ is an open project run by volunteers where every website and page that is added has to be manually reviewed. The editors are divided in categories and each editor can only edit in

[8]"Google and Themes" Phil Craven - http://www.webworkshop.net/google_themes.html

his, or her, own categories, which means that for minor categories the queue can be much shorter then for the major ones. Since the Web grows faster than the editors can review, it seems there are some categories where there are literally hundreds of thousands websites waiting for a review; if your website falls under one of those categories, it can take your website years before being included! For this reason you have to carefully submit to the right category, if you did not it would be forwarded to the right category only after an editor would start to review it, hence it would have to queue twice.

Another way is links exchange. You can join a link exchange centre like LinkPartners.com. There you can find other websites, with a relevant content to your site, which want to exchange links. In this kind of services it is usually good only sign up with centres where you can approach other sites personally, and where they can approach you personally. However you have to always avoid link farms and avoid links exchange with websites which are not relevant to the content of your website or you risk being penalized. You have also to pay attention in case you have more websites hosted on the same server and not content related. In fact some search engines, like Google, verify if the IP address of the connected websites, therefore try to keep the number of links between not highly content related websites on the same host to a bare minimum.

You can always buy inbound links from high ranked websites. You may find a good site with a content related to your website and buy the links directly from it or you can use a third part, usually called broker. Obviously you can choose the topic of the website where the links will appear and you can also sell links through a broker.

However, even if links are pretty important, they are not the only way to optimize a website. Keywords are another important thing to consider, in particular for Yahoo.

Keywords

Keyword optimization, the art of choosing the correct keywords, is one of the most important things related to SEO; they are what search strings are matched against.

Consequently, when you care about optimizing your website, you must sit down, take time and read all the pages of your website. For every page is good to write down a list of keywords that you suppose are relevant enough to the content; it does not matter if there are some keywords for more then one page but it should not be exactly the same list for two different pages, in that case you would have two different pages that have to be merged or you chose the wrong keywords.

Then you must verify which search terms are frequently used by the users, there are plenty of free tools to do that and so, you will be able to drop the words, which are not used, from the list [D]. On the contrary if other words look popular, and they are relevant to your page content, then, it might be a good idea to consider adding them.

At this point you must also get rid of the very competitive single words: it would be very difficult achieving a good result with words like “dog”, “computer”, “holiday”, “book” and the like; there are hundreds of thousands of sites targeting them and even with excellent SEO skills, they are very tough to conquer, in particular if you think to use just ethical SEO techniques.

It is a more realistic goal trying to achieve a top ten ranking in a two or even three word search strings; as a result you must add to the list a second word to the single keywords. Furthermore if you manage to describe better the argument of the page you may get users who are more interested, for example if you chose “adopt a dog” instead of simply “dog” you might have few visitors but the target audience will be probably more interested in your information.

Finally you can add to the list some synonyms; since in the major search engines they are automatically added, you have not to exaggerate.

Now that the list is complete you must choose two or three main keywords and other three or four minor ones. Then you have to use again free tools to verify the current density of these keywords in the page [E]. You want to have a density of 3%-7% for the major keywords and 1-2% for the minor ones, so you will need to add them to the page or eliminate according to the current percentage.

Obviously it is not the same wherever the keywords appear in the page: the main keywords have to appear in the first paragraph of the page, better if in the first few lines. In particular the main keyword has to appear in the title of the page, which cannot be left empty or the page will drop in the result list; the title should be long about 5-6 words and the keyword has to be at the beginning.

You must also consider the headings: in particular <h1> and <h2> carry a lot of weight. A good compromise is needed, headings are useful for displaying information, but stuffing them with too many words may disorient the reader. You can try to reduce the headings size with CSS but extreme measures, such as the use of {display: none}, may get you banned. I will analyse this technique in the chapter 4.4 "CSS tricks".

In particular Yahoo considers important the main keyword be included in the domain, directory and file names. You should carefully decide the name of the website evaluating which words to use and trying to achieve a trade-off between the usability and the SEO principles, hence a five word domain name will be a nightmare to remember but using just one word will not be good in the ranking list. File names and directory names are also important. If the domain name cannot include all the main keywords for all the pages of the website, you must choose for the file of every page accurately the name. If having a general name, instead a keyword rich one, for the domain gives to the website a better flexibility, the name of the HTML file should be the main keyword for that particular page.

Finally also bold text is given more weight than ordinary text but not as much as headings.

Metatags

A metatag is a line of information, which may be in the head part of an HTML document. There are different kinds of metatag but the most important are the Keywords tag, the Description tag and the Robots tag. Other kind of tags, like the author tag, not only are useless, from the SEO point of view, but can damage your rank since they put in the top part of the page not content related information, whilst we saw that for a keyword the closer is to the top of the page, the better is.

The metatags are not displayed to the readers of the page, but they are analysed by the crawlers.

They used to be very important in the ranking scores but since they are easily faked they are no more. However it does not take much afford to use them and even if the boost to the ranking is minimal, it would be a shame missing the chance.

The most important tag is the description one. It looks in this way:

<META name="description" content="A search engine shows the content of this tag below the title of your site when it appears in the results.">

Although Google and other search engines do not support it, you should choose an appealing description for the page incorporating the main keywords, since this tag appears at the beginning of the page. For the search engines which still show the description under the title of the page in the result, it would be a pity being at the top of the list but not being clicked by the users just because the description looks uninteresting. So you must keep the description about 13-15 words long and try to hit the point of the page in the first few words, in case the search engine cut the description because it is too long.

The Keywords tag is supposed to include a list of keywords that are relevant to the page. Once more the major search engines will not take in account this indication but it is a way to emphasize the keywords of the page and even their more common misspellings. If it is not a good idea to include typos in the content of the page because that can waste the reputation of accuracy of the whole website, here in this tag they can be useful with no side effects.

Even if it is easy and quick developing these two tags, you may decide to use one of the free tools which are available to automatically build those [F].

The Robots tag looks in this way:

<META name="robots" content="parameters">

The word 'parameters' should be replaced with two valid commands to the spider. The possible commands are: index, follow, noindex, nofollow. In a similar way to the Robots.txt this parameters are used to indicate to the spider if the page has to be indexed and if the links of the page are supposed to be followed. Anyway not all the spiders support this tag and the Robots.txt file is a better alternative.

A particular attention is required by the Refresh tag used to automatically redirect visitors from one page to another. You have to carefully use this tag since some search engines may ban you for it. When it is necessary to redirect a user it may be better to use a link that they have to click or a JavaScript redirection, preferably in an external file. I will deepen these issues in the chapter 4.2.

Content

Obviously the main mean you can use to get the highest rank is having a good, clean web page design with interesting and original content. The search engine is looking for the best content that corresponds to the user's needs, if you have a website with tons of pages of original content you will achieve the top rank. The design is important but users are more disposed to forgive many design issues if the content is good, than surfing in a very appealing website with junky information.

Although from a strict SEO point of view these are useless, you must follow some general common sense rules to create a website where the users, who get there from a search engine or elsewhere, will like browsing: the theme of the website must be something that the surfers are interested in; the employees who write the content have to be interested in, and are familiar with, the topics they are talking about; typos and dead links have be detected and corrected as soon as possible; the pages should not be filled with too much information, keywords, or "stuff" (images, animations, banner ads and the like).

It is also important that your website is up to date: if you update it every day, probably the spider will visit it at least once a week, but if you update it once a year not only the spider will rarely visit your website but you will drop position in the result list. The search engines want to provide their user with the most updated information. This is one of the reason the website of the newspapers have a very good ranks, they have lots of content and they are updated daily.

If you run a website for a big company there are two way to accomplish this task: the first is to have a news section which is update regularly, but unfortunately not all the website are suitable for a similar kind of section; the second is to allow and encourage the employees to keep a blog in a separate area of the website, it will furnish fresh, regularly updated content with an informal point of view about what is going in the company, in the industry as a whole, or in the world in general.

When you add content to your website you need also to care that the content is not duplicated by other sources in the Web because if spotted by a search engine, it hurts. Thus, in case you need to duplicate content from some other website, you have to spend enough time to modify it a bit, maybe with comments with your opinion or the like.

Even if you do not duplicate from other sources it may happen that another website copies yours, in this case you can spot this with appropriate tools [G] and ask to the webmaster to remove the information he copied from you. Sometimes it may be the case to modify your own information to avoid risking to be banned for duplicate content.

It is really worth spending time to avoid duplicate content since the major search engines are getting better and better in the detection.

There are four main kinds of duplicate content.

Websites with identical pages are the main target of duplicate content filters. If two website have common pages because they are affiliate, maybe commercializing the same product with the same content, they probably will see a drop in the whole website ranking and in particular for that page. The reason of the duplication does not matter; the search engine cannot understand the hidden reason of the duplication.

Scraped content is another flavour of duplicate content. It is taking content from a web site and repackaging it to make it look different, it starts to affect blogs and their syndication becoming really difficult to detect for the search engines.

E-Commerce Product Descriptions are also dangerous. It may happen that different websites sell the same product and use the same technical description, maybe taken from the producer's website. Avoiding the penalization involves commenting the description but it could also require writing your own original descriptions.

Finally there is a risk also in the distribution of articles. Let us think what may happen if an article is so good that many websites decide to publish it. In case the author does not allow that the article is modified, all the websites with the article will risk the duplicate content ban. In this case you have to consider how relevant the article is to your overall web page and then to the site as a whole, the more is relevant the less it is likely that just a comment to it can be enough.

Once you have detected a duplicate content in your website, consider that the search engine will look at the entire web page and its relationship to the whole site to decide if someone has to be punished and who is the guilty. Since this is not a trivial task there are cases of many legitimate websites who where unfairly banned because other bigger websites scraped theirs. Bigdaddy, the new change in Google, was supposed to fix the problem but it seems that it could not be going worse. During the summer 2006, more than six months after its introduction, it still had problems to decide the website which originally published the content, and to detect the legitim cases of duplicate content. Think to news feeds for instance, in that case there is a duplicate content because lots of users are actually looking for the same information. So it may still happen that you are banned because a bigger website copied your original content or because you speak about a very common recent news. We will have to wait some updates in the futere to see these problems fixed[9].

SEO Copywriting

Looking for information about the ethical SEO techniques that I have presented in this chapter I found another kind of SEO: SEO copywriting. I confess I got confused, at the beginning, between common ethical SEO and this second kind of SEO.

Eventually I realised there are two definitions of SEO copywriting: some SEO experts, like Phil Craven author of http://www.webworkshop.net/ or John Scott [10], consider SEO copywriting just page-elements SEO, more or less all the techniques I talked about in this chapter except inbound links; other experts, like Karon Thackston[11], claim that SEO

[9]"Google SEO algorithm problems" By Rodney Ringler (24/05/2006)

http://www.sitepoint.com/article/google-seo-algorithm-problems

[10]"SEO copywriting" (March 29th, 2006) John Scott - http://blog.v7n.com/2006/03/29/seo-copywriting/

[11]"What SEO copywriting is and isn't" (December 06, 2005) Karon Thackston - http://www.searchengineguide.com/thackston/2005/1206_kt1.html

copywriting is the process of writing content to appeal to your visitors, while including elements to help the search engines and your visitors understand what the page is all about.. In this second definition, it is still true that only ethical page-elements SEO are used but you try to achieve also much more.

To understand the difference, it is useful thinking about common copywriting.

Copywriting is the skill to use the perfect combination of powerful, persuasive, and motivating words to shift your goods and services; it involves finding the right words in order to sell, because it is knowing what, how and who to speak to, which is essential to sell[12].

Therefore in SEO copywriting the final goal is a marketing one: you want that the users will buy your product, or use your service. To achieve this you need that the website is human appealing and that is well ranked, or none will ever visit it.

This second definition is proved also by website like http://www.searchenginewriting.com/ where is offered a service which include both the aspect of the SEO copywriting. In fact it is fundamental for a company that all the SEO efforts will not have the effect to weaken the commercial message of the website. Therefore it is impossible for a commercial website think to SEO without thinking also to copywriting, SEO copywriting is not written exclusively with the search engines in mind.

SEO copywriting involves also marketing and philological considerations which are beyond the aims of this paper. I will analyse more the technical elements that can be appreciated by a search engines and it is up to you avoiding that they will transform your website in a not human appealing one. So in the next chapter I will speak about all the adjustments that you have to do for the particular characteristics that a spider has; there are in fact several elements that it cannot see or appreciate, like a human does, those have to be reduced, or modified, to avoid rank penalisation.

[12]Joe Robson - http://www.adcopywriting.com/index.htm

How a spider views your website

Images, Flash content and JavaScript

When the World Wide Web was born, it was mainly a text-based medium. Sounds, images and complex animations were either very rare or completely unheard of. Not surprisingly, the first major search engines that came around a couple years later were built to classify and rank pages largely based on textual content.

Unfortunately if the things did change for the webmasters, they did not for the search engines.

Nowadays it is almost impossible to find a website which does not use images or Flash content; if the text is still very important, in some cases you can find websites which are wholly developed in Flash or images are used to display a fancy font. This is extremely bad from a SEO point of view, if it is true that the spiders try to get the much information they can from also these graphical elements, a massive use of images or Flash content may really hurt the ranking of your website and since there is no meaning to have a very fine website with no visitors, you should find a good trade-off between search engine and graphical appeal.

First, when you use images they must have a descriptive name and you must use the "alt" tag attribute to give a complete description. This attribute is also important in case of Braille readers based on speech synthesisers, these kinds of browsers use the alt text, and thus you might want to make them usable whilst including the search term. For instance the following tags are a bad and a good example of an image tag:

<img src="pic1.jpg" alt="Tom and Sarah">

<img src="wedding_party_Tom_Sarah.jpg" alt="A picture of Tom and Sarah at the Naismiths' wedding party at Pollock Park on 25th June 2006">

Then it is good to use some images for illustration and decoration, but they must be never used for navigation or for displaying text.

A similar approach must be followed for Flash animations. A Flash introduction to a website is a top ranking killer; in fact spiders cannot index a Flash movie directly, as they do with a plain page of text. Spiders index file names, which have to be chosen carefully like for the images, but not the contents inside. This is true for all the search engines, although there are differences in the way the handle Flash content, all of them index only a minimum part of it.

There are also other reasons to not use Flash animation: it is proved that users tend to skip flash introductions to go ahead to the information they are interested in; they are bandwidth consuming and in case of a user with a slow connection this may lead to lose a potential visitor; they are expensive since require expert designers. The right rule should be: Flash is good for enhancing a story, but not for telling it. So if you want to use it, you can do it but with care.

You must also avoid using Flash for navigation: text links are the only SEO approved way to build site navigation.

However every time you decide to use Flash, there are some workarounds to limit the damage. Using metadata in this case is important, Flash development tools allow easily adding metadata to your movies; it is a chance to describe them to the spider.

It is better providing an alternative page made only of HTML content if the user wants to skip the Flash content. In general, it is a good habit, not only for SEO but also for usability, giving to the users the chance to choose between Flash and normal HTML. It requires double afford but you will be paid back.

Using an ad hoc tool for translating from Flash to HTML can save lots of working hours. This is the job of one of the handiest applications in the Flash Search Engine SDK, called "swf2html". Obviously you still have to check the automatically produced output, verifying for instance possible mistakes in the colour: for instance if the font was of the same colour of the background, you might be banned for hidden text. You need also to remove possible duplicate context, adjust the links to put the keyword-rich content in the title, in the headings and at the beginning of the page.

You can use also tools [H] which are useful to see how different search engines actually see your page with Flash content; in particular if you are using SDK, they provide one more check of the accuracy of the extracted text. There are also more general tools [K] which can be used to analyse what a spider sees of a page to verify if there are any JavaScript part which includes important keywords or links.

JavaScript is in fact another content which is not analysed by the spiders. It means that if you have, for example, a JavaScript menu it will not be read by the crawler. To solve the problem you should always us the <noscript> tag including all the links of the JavaScript menu. And you should also avoiding put the JavaScript instructions at the top of the HTML file, area were the most important keywords are supposed to be found; it is better put them in a separate file, otherwise you will forces the spider to wade through something that it is not at all interested in, before being able to read the text. While the major search engines can handle quite well such unfriendly pages, you can say that filling your pages with non-HTML code is more likely to hurt than to help you. Furthermore the less the search engine knows what kind of CSS and JavaScript you use, the better. This is true in particular if you use SEO spam techniques, but also in case of mild ethical SEO there is no guarantee that in the future, what today is allowed, will be strictly forbidden. Putting the JavaScript file in another directory and protecting it with the Robots.txt, is more likely that the spider will not analyse it.

Dynamic pages

A dynamic website is a website where interactive technologies are used. Languages like PHP, CGI, JSP or the like allow connection with a database, authentication of the users, storage of the user's session and generally to manage more dynamic pages with a single file. For instance with the file "product.php", according to the parameters that you will use, it is possible to present all the product of your company to the users. You will have different distinct dynamic pages, for example "www.mydomain.com/product.php?id=1" and "www.mydomain.com/product.php?id=2", managed by the same PHP file. On the contrary with static files you would need to have as many static HTML files as pages of the websites, since no connection to a database is possible using just HTML, all the information will be hard coded in the files instead of stored in the database. This is the reason dynamic websites are increasingly popular, they are flexible, fast developing and efficient.

Unfortunately most of search engines do not like dynamic content. They usually will not go deep indexing links with parameters; they will give such files much less PageRank or even refuse to crawl dynamic pages at all. Static URLs are typically ranked better, and they are indexed more quickly than dynamic URLs. Usually a spider cuts off the URLs after a specific number of variable strings (e.g.:?&=), if, for instance, it cuts everything after a question mark, instead of indexing two different URLs, in the example above, it will index just one. This can be a serious drain on the overall ranking of the website.

Another issue is that dynamic pages generally do not have any keywords in the URL. You have already seen that this is one of the key places where the search words are to be found. A research about the top ten results in very competitive words showed that Google has 40-50% of those top ten with the keyword either in the URL or the domain; Yahoo shows 60%; and MSN has 85%.

Anyway there are solutions for this problem. The easiest and less powerful solution is the creation of traditional, static pages. The correct way to use these newly created static pages is to place links to the dynamic pages on the static pages and then submit the static pages to the major search engines according to each search engine's recommended guidelines. This technique is easily implemented with a site map that fully displays all the links to the dynamic pages across the website. While the crawlers cannot index the entire dynamic pages, they will index most of the content[13].

A bit trickier, and powerful, solution, if you are hosted on a Linux server, it is using the Apache Mod Rewrite Rule, which converts the dynamic URLs in static ones in a transparent way for the crawler. Every time the static address will arrive to the server, the module will convert it back to the dynamic one. For example:

http://www.mydomain.com/product/dishwasher_blue

Every time is converted back to the original dynamic version:

http://www.mydomain.com/product.php?id=2345

To implement this solution, you need to install the module on your server and to edit the .htaccess file adding, for each dynamic URL you want Apache to convert, a rewrite command, which is the rule used by the module to translate every static link to the original dynamic one. Since those rules are written in a particular syntax that the module can recognize, you can avoid learning it, using the URL rewriting tool [J]. With this tool for each dynamic page, you just need to enter the URL into the box, press submit, and copy and paste the generated code into your .htaccess file on the root of your website.

However also this solution has a drawback: you have to modify all the links of your website to the static address version in order to avoid penalties due to have duplicate

[13]"Dynamic SEO tips and hints" Paul Bruemmer (November 28th 2002)

http://www.sitepoint.com/article/dynamic-site-seo-tips-hints

URLs. In fact the old dynamic address still works; the only difference is that it does not require the translation of the server, but obviously there is no difference in the content between the two versions. It means that if a crawler is smart enough to read a dynamic page, or this is particularly simple, and within the website there are both, the dynamic and the static link, to the same page, you can be penalised for duplicate URLs. The solution is to hide the dynamic page with the Robots.txt file.

In case your website is not hosted on a Linux server, or in case you want a completely automatic solution, there are commercial URL rewriting tools. For instance LinkFreeze is a fast and easy tool with support the main scripting languages. XQASP from Exception Digital Enterprise Solutions is specially minded for ASP websites. There are also other good solutions to the problem. The main difference is the flexibility, the price and the provided assistance.

Structure of the website

When you care about SEO, you should also consider the website as a whole. Every website has a particular structure which is reflected in all its parts and it is important to choose a design that is search engine friendly. The choices you make are going to be with you for a long time and errors will be very time-consuming to repair at later stages.

For example we have already seen that it is better to keep CSS and JavaScript file in external files. It is also better to restrict the use of PDF and DOC files. The crawlers of the major search engines can handle these formats; but if a document is long it may be translated in several pages of good content instead of a single file. Since every page has its own PageRank, from a SEO point of view, it is much better this solution than using the file.

Another important question is if it is fine to use frames.

Generally frames are not as search engine friendly as tables. That is not to say that it is impossible to build a site that uses frames and does well in the ranks, it is just harder to do than with tables.

The problem is that there are more frames for the same URL, which is equivalent to say that there are more pages for just a single URL. The spider handles much better when there is just a page for every URL, that is the why it does not appreciate frames.

If you are determined to utilize frames, use the tag <noframes> which is useful for both the users that are using a browser which does not support frames, and for crawlers. Inside the tag you will write all the content of the page your frameset points to and links to all of your other content pages.

However there is still a problem due to the design itself of a website developed by frames. Usually the navigation menu and the content are in different frames, it means that if you use the tag <noframes> to describe the content of the page, it will not contain the navigation menu. When eventually a surfer will arrive to that page, he will not be able to see the navigation frame since only the frame correspondent to the searched content will be loaded. But with no means to navigate inside the website he will soon look for another website.

A partial solution is using the following JavaScript code:

```
<script type="text/javascript" language="javascript">
<!--
if (top == self) location.replace("FILENAME OF YOUR FRAMESET PAGE");
-->
</script>
```

You have to write this script in an external file and include it in all the html files. Every time a page is loaded it will check whether the frameset is loaded and if not, it will load it.

This is quite good; the main problem is that it points to your entry frameset page that can be for instance the homepage of the website. It is not great that a surfer coming to your website will find himself at the homepage instead at the researched page.

Also this problem can be fixed. You have to use this script in all the <head> of your HTML pages:

```
<script type="text/javascript" language="JavaScript">

<!--

if (top == self || (parent.frames[1].name != myframeset))

top.location.href = 'frameset.html?' + location.href;

//-->

</script>
```

It works like in the precedent example but it passes the location of the current page (location.href) to the parent frame. The page with the parent frame is "frameset.html", in its code you will add:

```
 <script type="text/javascript" language="JavaScript">

<!--

document.write('<frame src="' + (location.search ?
unescape(location.search.substring(1)): 'default.htm') + '">');

document.write('<frame src="rightframe.htm" NAME="myframeset">');

document.write('<\/frameset>');

//-->

</script>
```

The important line is the third one where is read if it is already set the value of the location of the sub frame, if it is not the default page is loaded.

Unfortunately this solution has some compatibilities problems with the browser Opera and, for the browsers where JavaScript is disabled, no relocation can be made.

Search engines generally do not have any trouble reading a table-based page. Anyway it is still true that for a spider is more important what is at the top of the page than in the middle. If your website has the very common layout where the navigation menu is on the left side of the page and the content is on the right, the HTML code will present all the instructions needed to build the menu at the top of the page. It would be much better to have all those instructions at the bottom.

This is still possible without changing the layout of the page using the 'Rowspan' attribute of the <td>. You have to divide the page in a table of four main areas with two rows and two columns:

- You create a very small cell in the first are on the top-left and leave it empty;
- Using rowspan=2 you merge the two area on the top and on the bottom of the right side of the page, then you put the main content text in this big cell;
- In the area on the bottom-left you put the navigation menu;

I have represented the result in the Table 1.

<table>
<tr><td></td><td rowspan="2">Main Content</td></tr>
<tr><td>Navigation menu</td></tr>
</table>

Table 1: Structure of the page

Using a background colour the cell in the top-left area will be invisible. The code of this table is the following:

```
<table border="0" cellspacing="0" cellpadding="0">
   <tr>
        <td height="1"> </td>
        <td rowspan="2" valign="top">
                Main Content
        </td>
   </tr>
   <tr>
        <td>
                Navigation Menu
        </td>
   </tr>
</table>
```

However you still have to give up having a header at the top-right of the page, like most of the websites.

Finally the best solution to create your layout is to use CSS. They are flexible, efficient, recommend by the W3C, easy to modify and search engine friendly. A large part of the code needed to display in a nice way the information is separated by the content so that the spiders always read the most important and well-optimized part of the page first.

A sitemap is just a map of your site: on a single page you show the structure of your site with all its sections and the links to get there. Sitemaps are important both for users, who will use it to navigate directly to the section they are looking for, and for spiders, which will know where to go, and if there are new added sections to the website. It will take much less the spider to find all the new pages of the website, in particular if it is big. Sitemaps can also solve temporary problems in case of broken internal links, avoiding orphaned pages that cannot be reached in other way.

Unfortunately, although standard HTML sitemaps are fine for Yahoo and MSN, they are not enough for Google. To have your sitemap analysed also by Google, you need to write a first HTML file with the sitemap for human readers, and a second XML one for Googlebot. Obviously since this is a specific Google's indication, you do not risk to be penalised for duplicate content.

To create a XML sitemap there are two ways: you download a tool from the Google website, install, configure and run it, which will then produce a Python script which is the actual XML sitemap generator; the alternative is using an on-line free sitemap generator, you can find a list of products that support Google sitemap at http://code.google.com/sm_thirdparty.html. The first choice is more difficult but you have more control over the output.

After you have produced your sitemap you can submit it to Google. Currently Yahoo! and MSN do not support XML sitemaps: Yahoo allows submitting a text file with a list of URLs; MSN indexes the sitemap which is available on-line. Anyway it is general belief that they will catch with Google supporting XML sitemaps since they are such a powerful SEO tool.

Search engines optimization spam

We may discuss a lot whether SEO spam is right, is moral or have not to be used. The fact is that these techniques are often the most powerful trick to increase your ranking; if your are competing for a single competitive word and the twentieth position is not enough the only alternative to SEO spam is buying the position from the search engine.

Anyway these techniques are forbidden, if you are caught using them, although you do not use them with unscrupulous ways, you may be penalized or banned. Actually guilty websites are not often penalized. It depends on the nature of the 'offence' and in the importance of the website; the final goal of a search engine is to present the relevant content, if your company is famous and the website has lots of customers, obviously your website will never banned from the index or the visitors will start doing searches with other search engines. But if your website is about animals and you use SEO spam to achieve a good ranking in the search term "car", it is likely that sooner or later you will be banned.

While it is true that search engines are very poor at spotting unwanted techniques, they mainly rely on people reporting them, it is definitely better that if you are a beginner in SEO spam, you buy a new domain and experiment with it first; after gathering some confidence and experience, you could expand your techniques to your serious website.

Cloaking

Cloaking is the technique of returning different pages to search engines than to people. The reason is that good SEO often requires sacrificing some of the visual attractiveness of the page and changing the textual content into somewhat that may look unattractive to human visitors. Furthermore if you apply very good SEO and you are worried that someone may steel your pages, cloaking is the solution. The spiders will see the optimised page and the users the graphical appealing one.

Identification is usually done either by checking the visitors' IP address, or their user-agent string.

The first technique is better but it requires maintaining an up to date database with all the IP addresses of the crawlers, which change often. You may have to periodically buy the list.

Given that there is the risk to be banned, you need to cloak carefully.

First of all, since often Google save a copy of the page in a cache, you need the tag:

<META NAME="GOOGLEBOT" CONTENT="NOARCHIVE">

This is necessary to avoid that the human users will see the over optimized version of the page.

Secondly, it is better that your title, Meta description and the first row of text are the same with both your search engine optimized and human visitor pages. It is better than also the sizes of those pages are close to each others.

Cloaking can be useful also to solve the problem of session id. Indeed some spiders of the major search engines do not spider pages that have session ids in their URLs. If fact every time the spider would arrive to the website its session id would be different, since the URLs would contain the session id, all the pages, which it would spider, would be new pages, for it. Hence it would run the risk of spidering a potentially infinite number of pages. However by spotting page requests from the spiders, and delivering modified pages without the normal session ids in the link URLs, you would allow all the crawlers to spider your website.

Sometimes the word "cloaking" is misused.

Cloaking is not simply "IP delivery", because that means that you deliver different pages according to the IP address, on the contrary cloaking makes difference only between search engines and human users, it involves hiding the normal pages of a website from search engines.

Cloaking is neither hidden text since it assumes that there are two different pages, not just one which looks different to humans or to spiders, like hidden text.

Cloaking is finally not auto-redirection based on the IP address, like Google does when it sends the users to its local search engine version when they type the .com address into the address bar of the browser.

Auto-Redirecting

Auto-Redirecting is the technique of automatically sending surfers to a different page. It means that just after a page is loaded in the user's browser, he is redirected to a new page. This technique may be used to redirect people when there is a browser-specific page versions or when a domain has moved. Since these are legitimate reasons, search engines do not punish every kind of auto-redirecting but they try to spot misuse of it. In particular when a user arrives to a website with a content he is interested in, and he is redirected to another website on an entirely different topic.

There are four main methods to redirect a user. "301 redirect" is the most efficient and Search Engine Friendly method for webpage redirection. It is interpreted like "moved permanently". It is implemented in different way according to the platform.

If you are using an IIS server, you have to use the Internet services manager. After a right click on the file or folder you wish to redirect, select the radio titled "a redirection to a URL"; once entered the redirection page, check "The exact URL entered above" and the "A permanent redirection for this resource". Finally click on 'Apply'. You can use "301 redirect" also in all the main server side script languages like is shown in the following table.

ColdFusion	PHP	ASP	ASP.NET

<.cfheader
statuscode="301"
statustext="Moved
permanently">
<.cfheader
name="Location"
value="http://www.new-
url.com">
	<?
	Header("HTTP/1.1 301
	Moved Permanently");
	Header("Location:
	http://www.new-url.com"
	);
	?>
		<%@
		Language=VBScript %>
		<%
		Response.Status="301
		Moved Permanently"
		Response.AddHeader
		"Location", "
		http://www.new-url.com"
		>
			<script runat="server">
			private void
			Page_Load(object sender,
			System.EventArgs e)
			{
			Response.Status = "301
			Moved Permanently";
			Response.AddHeader
			("Location","http://www.new-
			url.com");
			}
			</script>

Table 1. Auto-redirecting in some common programming languages

"301 redirect" allows redirecting all the files and directories of an old website to a new location. Add the following lines of code to the .htacess file, present in the root of the old website.

```
Options +FollowSymLinks

RewriteEngine on

RewriteRule (.*) http://www.newLocation.com/$1 [R=301,L]
```

And if you need to ensure that all the requests to the address "http://myDomain.com/" will be redirected to the address "http://www.myDomain.com/" you add this code to the .htaccess file:

```
Options +FollowSymlinks

RewriteEngine on

rewritecond %{http_host} ^ myDomain.com [nc]

rewriterule ^(.*)$ http://www.myDomain.com/$1 [r=301,nc]
```

A second way to redirect a user toward a new page is using a Meta Refresh tag. It has to be placed in the <head> section, it can be detected by the search engines and it is allowed as long as there is a reasonable delay, usually five seconds, between landing on a page and being redirected from it.

The code looks like:

```
<meta http-equiv="refresh" content="5;url=destinationPage.html">
```

The content parameter contains the seconds of delay and the destination page, separated by a semicolon. Since it is annoying for the user waiting some seconds, you may think to set the delay to 0 but unfortunately search engines can read HTML and the website would be penalised for that.

To redirect in a faster way, you may use a JavaScript. Use the code:

```
<script language="javascript">
<!--
location.replace("destinationPage.html")
//-->
</script>
```

Using the "replace" function instead of the "href" one, causes the new page to replace the current page in the history of the browser. If a visitor clicks the Back button he will go directly to the page before the redirecting page. It avoids the annoying trap of going back to a page which immediately redirects you to the current page.

Finally you can use a form. A crawler cannot fill any forms, thus it ignores them. But if you use a JavaScript which submits a form as soon as the page is loaded, the user will be redirect to the page which is in the "action" parameter. Here the code:

```
<head>
   <script language="javascript">
   <!--
        document.myform.submit();
   //-->
   </script>
</head>
<body>
   <form name="myform" action="destinationPage.html" method="get"></form>
```

</body>

The destination page can be any absolute or relative URL.

It is not auto-redirecting when the server does the entire job, directing two different addresses to the same page. For example, it may happen that you have more possible addresses for the same home page:

http://www.mydomain.com/index.html

http://mydomain.com/

http://www.mydomain.com/

Anyway this is not good from a SEO point of view. If some search engine can handle this situation, Yahoo and MSN for instance, for others, like Google, it causes the so called "Canonical Issues". In fact Google flags the different copies as duplicate content, and penalizes them. Otherwise, it may happen that the different versions are not penalised but the PageRank is split among them so that the overall rank of your homepage will be lower that it should be. At the moment, standardizing the URL of the home page is the only way to ensure that the PageRank is not shared among ghost URLs[14].

Doorway pages

If a website has a very appealing graphical layout, it will be probably able to keep the surfers inside the website whenever they will get there. However we have already said that spiders can understand only text therefore they may rank badly the website dropping the number of visitors. Restructuring the website to add text and remove some of the design features costs time and money. Cloaking is cheaper but it requires a database with the IP addresses used by the spiders, requiring some kind of regular

[14]"Google SEO algorithm problems" By Rodney Ringler (24/05/2006)
http://www.sitepoint.com/article/google-seo-algorithm-problems

maintenance. The easiest and fastest solution is to keep the old design and add external pages that are specifically designed to perform well in the search engines. Those are doorway pages.

The concept of doorway page changed a bit in the last years. Since nowadays the inbound links are usually a fundamental element of the ranking algorithm, it is no more possible just creating pages full of keywords, optimising them for spiders and with an auto-redirection to the main site. But you still can use doorways pages in two main methods.

A first elegant way of using doorway pages is creating a dozen of well optimized pages. These doorway pages are supposed to not be seen by users; consequently they can be generously optimised. Every one of the doorway pages has to be linked only to the relevant content page in order to transfer the maximum weight, according the PageRank algorithm. Finally you have to create a hallway page, liked from the relevant content page and linked to all the doorway pages. The relevant page will lose a bit of rank for the outbound link but it will receive a dozen of backlinks from the doorway pages hence it will gain rank instead of losing it. The spider will find the hallway page from the relevant content one. The link to the hallway page has to be hidden so that no human user will follow it, but in this way you can avoid any kind of redirection in the doorway pages. This is a solution that can be easily automated.

The second solution cannot be automated but, on the other hand, it cannot be spotted by a search engine. In this case you will choose a page with relevant content, and you will create a group of pages that have outbound links to that page. Every page is linked to that main content page and it has a link to all the doorway pages. Those pages are optimized for the search terms and every one for a different term. A strategy to create the doorway pages is to split the main page in more pages, to create "further notice" pages, or add some content to a page which presents a larger picture after the user has clicked on the smaller one, included in the main content page. Since the content which is added is natural there is no chance of a penalty, unfortunately this solution requires much more time than the first one.

CSS tricks

CSS, like JavaScript, has a legal side and a dark one. The legal side is when you use CSS to set the layout of a page or the used font for the website. The dark one is when you hide in various ways the text in the page; you make the italic or bold text look normal or resize the <h1> heading. The reason you should use CSS tricks is the same that leads to use cloaking: what is appreciate by search engines often is ugly for the human reader. Instead of having two different versions of the same page, as using cloaking, you will have just one version where some optimizations are invisible to the readers.

The most famous way to use CSS is to hide text. The text contains important keywords that may be difficult place inside the page producing a meaningful content. In order to hide the content, the background colour, or image, has to be of the same colour of the text. This is an example really difficult to spot:

<BODY BGCOLOR=black TEXT=red BACKGROUND="white.gif">

You set with an HTML tag the colour of the background, the colour of the text and an image. The image will have a white background but its name will be something different from"white.gif", I called it in that way to help to understand the example.

Then you use a CSS to turn the text colour to white:

.adjust {color: white}

Finally add a paragraph where you use the defined style:

<P class="adjust">Keywords invisible for readers</P>

If a spider is clever, it may compare the background colour with the text colour set in the CSS, in this case white with black, but comparing the colour of the background image with the CSS-defined text colour is much harder.

Another way to hide text is using layers. In this example you will display out the chart the keywords rich text:

.position {position: absolute; width: 200px; height: 95px; z-index: 3; left: -250px; top: -110px; visibility: visible }

Then put in the HTML page:

<DIV class="position">A very keywords rich text</DIV>

In this trick since the negative value is superior to the dimension of the chart, the text is positioned to a place where users, with graphical browsers that support CSS, will not be able to see it.

CSS can also be used to resize the dimension of headings. In the text you use lots of <h1> headings so that the spider considers more important the words inside, and then you make them imitating normal text.

H1.type { font-size: 100% ; font-weight: normal; font-style: normal }

<H1 class="type">write here keywords which appear of normal dimension</H1>

The same can be done with bold or italic.

.style { font-weight: normal; font-style: normal }

<B class="style">fake bolded text</B>

<I class="style">fake italic text</I>

General topics

Reinclusion in the index

Optimizing your website you always have to worry about being excluded from a search engine's index. The majors search engines use techniques to detect over optimized ethical SEO or SEO spam. It may happen that is acceptable for a search engine what it is over optimized for another one. For example Yahoo may appreciate a keyword density about 8% or 9% but Google may consider already too much 7%.

This is one of the reasons, for which is tricky to optimize for all the possible search engines.

Following only the suggestions of the first three chapters of this paper you will probably able to achieve good results with no risks. But if you need a special boost, you may use the SEO spam techniques or exaggerate with ethical optimization. This can sometimes lead to the nightmare of the SEO specialists: being excluded from the index of a search engine. So you will have the problem to be re-enlisted in its index.

The first step to do is to analyse why it happened: it is useless to submit a re inclusion request when you did not change anything in your website, this might compromise the opportunity to be again in the index in the case the search engine keeps a statistics of the misbehaving websites. In this phase the webmaster guidelines, which are usually present in the help section of the search engines, can be useful to understand how you are supposed to behave.

Then you will have to send an email or fill a form with the re inclusion request. For instance, in the case you were excluded by the Google's index you have to fill the form at the address: http://www.google.com/support/bin/request.py. Filling the form you should try as much as possible to be polite and admit the errors you did.

Finally you will have to wait for an answer. It is a very poor idea to send more then a request for the same website until you will get the answer.

Age of site and Google sandbox

So far for all the major search engines there is no penalty for a website if it is new. Google is an exception. Since lots of websites are created just to furnish backlinks to major websites, Google waits from a minimum of one month to a maximum of eight, to assign the PageRank to new comers. This barrier is called Sandbox, the analogy is from a child who is put there when is too young to stay in the world of the adults. It is quite reasonable that a new website is not so influential like an old one, also in the real world a new store in town brings a short burst of initial business because people tend to trust a business that has been around for a long time over one that is brand new.

In particular after the Jasper Update the age became one of the parameters considered in the calculation of the PageRank; age of incoming links, age of web content, and age of the domain [I].

Although there is nothing that you can do to make older your website but wait, there are some tips you can apply in order to make the effect of the young age less destructive.

Make sure you register your domain name for the longest amount of time possible. Having a domain which is registered year by year, gives the bad idea that at the moment the site is there but in a close future it may disappear. Often you can choose a contract which lasts from just one year up to five or ten years.

Consider registering a domain name even before you are sure you are going to need it. While you will develop the website it will get older. Then as soon as the homepage will be ready, you should put it on-line so the clock for the Sandbox will start. This is good also because search engines prefer a website which grows day after day than a full website with 500 backlinks which appears overnight.

Think about purchasing a domain name that was already pre-owned. This is not only a way to have a domain name already old but also to inherit a possible good PageRank for the homepage and for the pages with the same URL. On the other hand you must pay particular attention that the domain is not blacklisted.

It is also better hosting on a well-established host. Also the Sandbox restriction is less severe in this case. For instance if you are using a free host, you can aspect that you will never rank very well, these domains usually have a short life, consequently they are penalised by search engines.

Finally a bit less ethical trick is to put some pages of the new website on an older one with links to the younger. Anyway you should avoid auto-redirection or the search engines may consider it SEO spam.

Yahoo!

Like I have already said, Yahoo covers about the 25% percent of the searches. It is the second major search engine therefore it is worth spending some time to learn a bit more about its algorithm.

Its spider is called Yahoo! Slurp and it works in a similar manner to Googlebot. It is also important to say that the WebRank does not have a relevant importance in the ranking algorithm. WebRank was introduced by Yahoo, following the example of the Google's toolbar PageRank, but it worked collecting data in order to calculate the popularity of the websites not to indicate the rank of a website. And while it is true that often a website which had a high WebRank, it ranks also well, there are several reasons to explain this. The simplest is that if a website has a high WebRank, it is popular; but if a website is popular it means that lots of information can be found, as a result it will be probably made by lots of pages. We saw that the dimension of the website is a fundamental element in its rank. Anyway the WebRank is no more included in the Yahoo toolbar even if tool with the aim of evaluating it are available online [L].

There is not a lot of information about the Yahoo's ranking algorithm, in particular compared to the Google's one. Also the official website does not really help, but from some experiment based on searches and the analysis of the results, it is possible to understand the key factors[15].

First of all, keywords are the main factor. If the Google's algorithm is particularly concerned about the backlinks, the Yahoo's one is more about "on page" factors.

For Yahoo the most important place where a keyword has to be included is the title. This is fundamental, it looks very hard to rank well if the search term is not on the title of the page. On the other hand, it is not necessary that the keyword is repeated more times.

The second more important place for a keyword is the file name or the directory where it is stored.

Then it is also important that the keyword is in the domain name, and possibly it is not a derivative from another word. Hence "www.potbellied_people.com" is not as good as "www.pot.com" for the search term "pot" even if it is better than "www.belly.com" [M].

Obviously the keywords should have a high density also in the page content. Here to achieve a great result with Yahoo, you may risk to be punished for keywords stuffing by Google. In fact a density above 7% can be good for the former and bad for the latter.

Since Yahoo trusts very much its own directory where the website are inserted in the relevant categories by human employees, it is a great boost to your rank having your website in the Yahoo directory under a category which is relevant to the search terms. This is more than just an important backlink for your website, like in Google.

Then it is relevant the PageRank. It looks that the weight to the links is mainly given through the importance that the PageRank has in the algorithm. It is less important than having a keyword in the file name or in the title but it does a part.

[15] http://www.apromotionguide.com/yahooalgo.html

Location optimization and themes

Search engines try not only to bring the most relevant content to the searchers; they also try to bring them to pages written in the same alphabet, possibly in the same language, and get the user as close to home as possible in the realm of their search results. That makes sense since if a user is looking for information about the environment he will be probably more interested to information about the part of the country where he lives, if he is doing a generic query.

Search engines will determine country not only based on the domain name, but also the country of a website's physical location based upon IP address. This is a good reason to check where is physically located the host service that you are using. In particular there is a free tool [N], which helps determine the Country in which the specified website is hosted.

"Themes" are another factor to keep in mind during optimization. A theme is the topic of a website, what the web pages are about. For Google it is particularly important.

Let consider the example of a site about jewellery. Probably there are pages which speak about bracelets, necklaces, rings and so on. The overall theme is jewellery and, if it is search engines optimized, it will rank quite well about the search term "jewellery"; however it will rank worse than a website about just rings with the search term "ring". Even if this second website is much smaller since it is specialised in that subject its keyword "ring" is not diluted with other search terms.

Anyway if the jewellery website divides its website in different sections, every one specialized in a different item, those pages will be strongly themed towards a different item and the correspondent keyword will not be diluted among the others. Consequently the jewellery website will rank well for the search term "ring" with one section and for the term "bracelets" with another section.

Click popularity

Click popularity is a ranking system that is used by some search engines. The users' behaviour is traced in order to determine which pages they found to be relevant to their queries. In particular is analysed how many users come back to a search engine, using the "Back" button of the browser, after clicking on a search result. If a website in the period of case analysis had 32 visitors out of 35, who came back to result page of the search engine, after visiting it, probably it is not a very useful website, perhaps it expired or its content is not about the search terms.

Google does not use click popularity in its algorithm but it seems Yahoo does at least in a minimal part. MSN used to receive reports of click popularity from Direct Hint. Right now Direct Hint no more exists as stand alone search engine; it is just a part of Theoma, a minor search engine.

It is not without any reason that it is only one of the minor components of the ranking algorithms.

First sometimes people behave strangely and click on results that have nothing to do with what they have searched for. It may happen that even if a user searched for "computer science", he finds an interesting website about digital camera, which sells also personal computer nevertheless, is specialized in photography. If the user is also interested in digital cameras he may not go back to the search engine, although the content was not really relevant to the search term.

Secondly only the first ten or twenty results in the list will receive visitors. That means that the search engine does not have information to evaluate the others; if it was the only parameter of the ranking algorithm the pages with a low PageRank would perform worse and worse for their insufficient visibility.

Finally in some cases a prompt return to the search engine might mean that the user did find the answer to his query, like the score of a soccer match, before continuing to search for information on other topics.

There are mainly three methods you can use to improve the click popularity.

First of all, you need to write good content and use the right keywords so that all the visitors who will arrive to your website will find what they were looking for.

Second, if your website is not too big and it has just some dozens of visitors a day you can use a non-transparent proxy server to do a search with the search term you are trying to get a top ten ranking on. Every time you get to your website, you close the browser, delete the cookies and repeat the process with a new proxy server. It will mask your IP-address, which fools the search engine into thinking that you are someone else than you really are, just as long as you remember to delete any cookies possibly used by the search engine and change the proxy server every times.

A third trick can be done using JavaScript to disable the Back button of the browser. This method has the drawback that it may be annoying for the users who may want to use the button just to navigate inside the website. This may push them to leave the site.

Promoting your website

Generally speaking, the more famous your website is, the more backlinks it will have. Since these are the main PageRank factor in Google, and are important also in all the other major search engines, it is worth promoting your website.

We have already seen the importance of submitting it to the major directories. It is important also to submit your website to the search engines. For sure you have to submit your website to the major search engines: Google, Yahoo and MSN. That will speed up the process of finding your websites for their crawlers. Anyway the world of the search engines is not only these three big ones, even if they provide the majority of the visitors; it means that you should not ignore all the minor general purpose search engines. Maybe, since they are less important, it is better to automate the process of sending them your website, in particular you may use an automatic tool [O].

Then there is another particular category of search engines. These are the specialised search engines; their peculiarity is they are used only to retrieve information on a precise topic. Some of them are specialized on law, others on educational content and so on. The best thing to do is consult a website like Pandia (http://www.pandia.com/powersearch) or Webquest (http://webquest. Sdsu.edu/searching/specialized.html). There you can find list of specialised search engines, identify those which can be interested in your website, and submit it to them.

If you cannot get the top ranking with your efforts there is still a solution: paid advertising. It consists paying a search engine for displaying at the top of the results the URL of your website.

Usually you have to pay only for the users who actually click on your link, for this reason are sometimes called pay per click (PPC) advertising, and the cost depends upon the keyword and the position where it appears. Really competitive words, as "hotel", are several time more expensive than cheaper ones, as "hotel facility animals", but creating an opportune network or less competitive words can be more effective and less expensive than buying a single really competitive. In fact the less competitive, like in the example above, can attract more interested visitors. On the other hand, you may achieve a good ranking in these search terms with a relatively little SEO effort.

Unfortunately PPC advertising has an important hitch: users generally do not trust paid links as much as they do with the normal ones. They often ignore sponsored links.

Tools

Here I report a list of the Web tools I spoke about previously to carry out quickly lots of tedious work connected with SEO. They are all free so sometimes their functionalities are limited but that does not mean they are useless.

[A] http://www.webworkshop.net/pagerank_calculator.php

PageRank calculator: this is a useful page, developed in JavaScript which will help you to calculate the PageRank value of your website. The page presents a grid which represents all the internal, inbound and outbound links of your website and you can play changing the link structure to see the PageRank's behaviour in the different pages.

[B] http://www.bad-neighborhood.com/

This free service can be used to check if any of your website outbound links can be damage your PageRank pointing a website which is banned by the Google's index. In fact whilst it is usually not dangerous having inbound links from poor website, since it is not under your control who links to you, it is very harmful linking to a link farm of to a spam website.

[C] http://tool.motoricerca.info/robots-checker.html

This tool can be used to validate the Robot.txt files, so to be sure that it can be correctly understood by the spider. In fact even if the syntax of this file is really simple, the smallest mistake is enough to prevent the spider to understand, and follow, your instructions.

[D] http://www.webconfs.com/website-keyword-suggestions.php

http://www.webconfs.com/keyword-playground.php

Since it is important not only choosing the keywords which reflect the content of your website but also the keywords which generate lots of traffic since are popular, the first tool will analyze the content of your pages and indicate which keyword have to be chosen. The second tool will indicate alternative keywords if the selected ones have a poor popularity, it will present also the average expected number of searches over the Web with the suggested keywords.

[E] http://www.webconfs.com/keyword-density-checker.php

This tool use two mean to show the density of your keywords in the page: the "Keyword Cloud" which is a visual representation of keywords used on a website, where larger fonts are used for keywords having higher density; the "Keyword Density" which is the percentage of occurrence of your keywords to the text in the rest of your page.

[F] http://www.submitcorner.com/Tools/Meta/

http://www.apromotionguide.com/metagene.html

Since Metatags are no more a fundamental SEO factor, you can decide to automatically build them using one of these two free tools. The first one is an advanced tool which allows you to select among several options, the second one is more basic.

[G] http://copyscape.com/

Here you can find a tool to spot other website who copied your own website. In fact for a spider is often impossible to understand which website is the owner of the rights above the duplicate content, so they usually penalize the smaller. For this reason is important to detect as soon as possible the problem and ask to the other webmaster to remove our copyrighted content.

[H] http://www.se-flash.com/

This is a useful tool to analyze how a spider can actually see your page with Flash content. It is particularly important when you use an application for translating from Flash to HTML.

[K] http://www.webconfs.com/search-engine-spider-simulator.php

This tool simulates a search engine by displaying the contents of a webpage exactly how a search engine would see it. It also displays the hyperlinks that will be crawled when it visits the particular webpage. It is useful to analyze how much information we lose for JavaScript, Flash and graphical elements.

[J] http://www.webconfs.com/url-rewriting-tool.php

Here you can find a tool which produces the instructions that have to be included in the .htaccess file to allow Apache to automatically convert static address to dynamic ones. In this way you can use dynamic pages without losing ranks.

[I] http://www.webconfs.com/domain-age.php

Since the age of a website is one of the ranking factor for Google, you can check your competitor website age and compare with yours using this tool. It displays the approximate age of a website on the Internet and allows you to view how the website looked when it first started.

[L] http://www.webconfs.com/check-yahoo-webrank.php

Although the WebRank is not directly related to its ranks, it is a measure his popularity so it can be interesting calculate its value. Unfortunately it is no more included in the Yahoo toolbar, so we can use the tool above in order to evaluate it.

[M] http://www.webconfs.com/keyword-rich-domain-suggestions.php

This tool suggests keyword rich domain names analyzing the keywords which you are optimizing for. In fact having a keyword rich domain name is an important SEO factor. Choosing the right domain could boost your search engine rankings. It checks also that the suggested domain name is not already registered.

[N] http://www.webconfs.com/website-to-country.php

Search engines determine the country of the website based on the physical location of the IP address of the website thus is fundamental to know where the server where your website is hosted is physically located to understand in which country you can achieve the best rank. This is possible using the tool above.

[O] http://www.selfpromotion.com/

Googlebot is really efficient to find new website over the Web. Unfortunately the crawlers of the minor search engines are not the same, so to rank for them it may be necessary that you manually send your website. Since there are literally dozens of minor and specialized search engines it is convenient using this tool which automates the sending process saving time and efforts.

Other tools

http://www.webconfs.com/anchor-text-analysis.php

Since the anchor text is a key factor in the weight that an inbound link has, this tool help you to determine the inbound links of your website and link text used in them

http://www.webconfs.com/domain-stats.php

This tool helps you get all kind of statistics of your competitor's domains. The information include: Age of the domains, Yahoo WebRank, count of backlinks and number of pages indexed in Search Engines like Google, Yahoo, Msn.

Reference

Writing this paper I found three very useful websites full of information, free tools, suggestions and examples about search engine optimization. These are:

http://www.apromotionguide.com/

http://www.webconfs.com/

http://www.webworkshop.net/

Even if they overlap in the explanation of the main optimization techniques, I suggest consulting the articles of all of them since part of the information is presented only in one of those. In particular the first one is the less up to date but it has a more complete approach to all the main information about SEO with beginners and advanced articles. The second one has a tutorial that can be considered a very good starting point and the third one has very detailed information about Google, SEO spam techniques and ethical issues.

Together with these papers you can find other useful information in the articles I cited times by times along the relation.

AgentService Website

Introduction

AgentService is a website named from a framework which offers a complete support to the development of multi-agent systems, it can in fact exploit the underlying support of the Common Language Infrastructure to provide the agents with advanced run-time services and be portable over all the operating systems hosting applications designed for the same infrastructure.

Information about the releases of the framework, technical support, general information and links toward other topic related websites are the main contents of this website. At the moment of my optimization there were also some features under development, like the forum, or with no resource available yet, like the download page.

The main keywords I optimized for were "AgentService", "Programming Framework", "Distributed Systems", ".NET", "Common Language Infrastructure", "Ontology" and "Distributed Systems".

In the following section I will analyze the means I used to optimize the website for these words.

Links

Like I said during my relation on SEO, links are the most important SEO factor for Google, and since it is the outstanding most used search engine, their optimization is really important.

First of all, I reorganized the structure of the backlinks along the website. I used a "Hierarchy structure" (see Picture 1 at page 16 of the second chapter) to make the home page particularly relevant. So all the pages of the website have spiderable links toward the home, and the home has links toward all the other pages. Then all the other internal links among the pages are ignored by the crawlers. This is due to the use of the attribute "rel", set at the value "nofollow", inside the anchor tags. Like I explained during the relation, this is one of the three main methods to achieve the same goal. I though it was better that using JavaScript because this is the practice encouraged by search engines; because it may happen that in the future the crawlers will be able to understand simple scripts like the ones used to open a new window at the anchor address; and because JavaScript is often disabled for security reasons. The third way is using a form but is much less convenient and it may generate obscure HTML code if heavily used as I did with the "rel" tag.

Since the navigation menu was included in every page with a simple asp command, I renamed its file to "side_menu.inc", included now only in the home page ("index.aspx"), and then I created another menu file ("side.inc") which is included in all the other pages. The former has links which are spiderable; the latter has only one link toward the home page which is spiderable, all the others are protected by the "rel" attribute".

Another drain of PageRank is represented by outbound links. Therefore I added the "ref" attribute also in all the links toward external sites except for the links to the **`l.i.d.o.`** website, since this is owned by the same research group of AgentService and since a website needs at least an outbound link to not lose rank as well.

Finally I posted on some forums meaningless posts to get inbound links. Unfortunately I did not find a website with a forum based exactly on the same topic of AgentService since the theme is quite specialized. However I posted on forums of the Computer Science topic, still related, and I was able to select almost always a target keyword as anchor tag, increasing the relevancy of the link.

Also posting inside the forums in not that easy since three topics I posted in the forum "knoppix.net" were erased few days after I posted them, probably because the webmaster spotted the trick. For the other five links I checked out that all the forum pages were spiderable and the links not inserted inside JavaScript commands.

Keywords

Keywords are the main SEO factor for Yahoo, which is the second major search engine, so I carefully optimized them as well. To achieve the best results I used three means: tags, CSS and metatags.

Tags

I added comments full of keywords to all the anchor tags of the website and to the image tags. This is possible using the "alt" attribute for the "img" tag and the "title" attribute for the anchor one. I tried to not exaggerate doing this, in fact the content of the "alt" attribute is displayed when the file is not found or the browser is a textual one. So you have to pay attention to not put keywords absolutely not content related, for instance in a "img" tag for a picture which substitutes a bullet, or the effects of visualizing them could be at least disorienting. So I added some comments in all the places where it was not only possible but also opportune (all the main images, all the internal links, the majority of the external links, some of the bullet images, the images of the pdf and ppt in the publication page).

CSS

Crawlers consider very relevant the text inside headings and bold tags. Unfortunately having a page full of headings and bold text is usually not compatible with having a graphical appealing page. For this reason I modified the file "style.css" and inserted two new classes: "h1 special" and "b special". The former looks exactly like the h3 tag used in all the website, the latter like normal text. So I added bold tags, of the class "special", in the content of the pages to underline the main keywords, and I changed the heading, which included the main keywords, from "h3" to "h1 special". The result is that the website looks exactly the same than before for a human visitor but not for a spider which will see the main keywords much more underlined.

Search Engines are not very good spotting CSS tricks and the file I modified is already pretty complicated to be easily understood. Anyway I added some useless attributes to make the definitions of the "h3" and "h1 special" look different. In particular I added inside the "h3" definition the attribute "font-variant" set to inherit, and so consequently to the value it already had; and inside the "h1 special" definition the "font-style" set to "normal", its default value. Now, for the spider, even if it was able to compare the two tags, it would be more difficult to spot that they actually look the same.

Metatags

Like I underlined during my relation, metatags are no more a SEO key factor as they used to be. However it does not take a long time to create them and it can be a shame wasting the little ranking boost they can furnish.

So I added the "description" and "keywords" tags to all the pages of the website. Writing the former I tried to choose concise descriptions, which were able to capture the attention of the reader with few words and which included one or two main keywords. Writing the latter I tried to underline the main keywords of the website specifically present on the specific page.

Further optimizations

I changed the name of the page with the list of publications of the research group from "pubs.apsx" to "publication.aspx" (and consequently from "pubs.inc" to "publications.inc" for the included file) since Yahoo looks for keywords also in the filenames. So although it is not a very relevant word for the general content of the website, this new file name can be anyway useful since it helps to find this particular page if interested in the publications about the AgentService framework.

I added the robots.txt file where I specified which directories are not to be analyzed by the spiders. In particular those are:

- App_Data
- aspnet_client
- Bin
- css
- scripts
- wiki/App_Code/
- wiki/App_Data/
- wiki/App_Themes/
- wiki/Controls/
- /wiki/js/

I specified all the directories which do not content textual or anyway informative content about the topics of the website. This file is a further protection for the CSS files, since, although is not guarantied that the crawlers will not look inside the directories above, it is very likely.

I did not delete the DOCTYPE declaration inside the pages, even if it is not keywords rich, since this may produce some visualization errors in the modern browsers. In fact it tells the web browser that what follows conforms with a certain specification, for instance XHTML or HTML 4.01. If it is present the web browser can then take advantage of this knowledge but if it is not I will try to use a browsing modality called "quirks" which tries to emulate the behaviour of the older web browser versions. In this modality errors occur more easily than in the normal one.

Conclusion

There is not a clear limit to SEO optimization. Generally speaking the more you optimize the more likely you will be banned by a search engine for SEO spam. Paradoxically you can be banned even if you did not optimized at all, if some other major websites copy your content and a search engine considers yours the "duplicate content".

To avoid the risk of being banned I did not used unethical, although very powerful, optimization tricks like doorway pages, heavy CSS tricks or cloaking. With the mild SEO means I used it is very unlikely to be banned, but the results can be already good. In fact "AgentService", the main keyword, is not very competitive so that the website can already rank very well, while for the others I think it can rank well for a website of these dimensions. I think also that getting older, adding some good content pages and some other backlinks will help much more than the risky SEO spam techniques.