Search Engine Optimisation. PageRank best Practices

Graduand:
Neus Ferré Viñes
neusferre@gmail.com
UPC Barcelona
RWTH Aachen
Business Sciences for Engineers and Natural Scientists

Supervisor:
Florian Heinemann
heinemann@win.rwth-aachen.de
RWTH Aachen
Business Sciences for Engineers and Natural Scientists

July 2008

Abstract

Since the explosion of the Internet age the need of search online information has grown as well at the light velocity. As a consequent, new marketing disciplines arise in the digital world. This thesis describes, in the search engine marketing framework, how the ranking in the search engine results page (SERP) can be influenced.

Wikipedia describes search engine marketing or SEM as a form of Internet marketing that seeks to promote websites by increasing their visibility in search engine result pages (SERPs). Therefore, the importance of being searchable and visible to the users reveal needs of improvement for the website designers.

Different factors are used to produce search rankings. One of them is PageRank. The present thesis focuses on how PageRank of Google makes use of the linking structure of the Web in order to maximise relevance of the results in a web search. PageRank used to be the jigsaw of webmasters because of the secrecy it used to have.

The formula that lies behind PageRank enabled the founders of Google to convert a PhD into one of the most successful companies ever. The uniqueness of PageRank in contrast to other Web Search Engines consist in providing the user with the greatest relevance of the results for a specific query, thus providing the most satisfactory user experience.

Google does use PageRank as part of their ranking formula. Although it is not as important as many believe, it is nevertheless a measure of a web page's popularity, and gives a certain indication on how "important" Google considers a page to be.

The goal of search marketing is being visible to the end user. Two different fields within search marketing can be pointed out: Search Engine Optimisation and search engine marketing. This study focuses on the first one, Search Engine Optimisation, which refers to all types of initiatives and actions taken by website designers in order to increase the relevance for the Search Engines. It is about design, optimising content, linking structure (internal and external) and other page specific factors.

Because of the predominance of Google, this thesis looks at which steps can be taken in a certain website when trying to be optimized for Google's algorithm PageRank. Moreover, other factors which also have an influence are analyzed.

Index

Index of Figures and Tables

Put a search box in front of just about anybody, and he'll know what to do with it.

John Battelle, The Search (2005)

1. Introduction

Just ten years ago, bandwidth was scarce and storage was expensive. Use of internet was comparatively sparse, files were small, and Internet companies, for the most part, did not keep their log files –storing that data was too expensive. In the past few years, a good portion of our digitally mediated behaviour – be it in e-mail, search or the relationships we have with others- has moved online. Nowadays more people have access to a broader bandwidth and this fact has facilitated the transition to the online world.

The introduction of the web has had implications for the development of online commerce. The unique characteristics of the Web for ecommerce and online retailing is fundamentally transforming the way in which consumers and vendors interact (Jansen and Resnick December 2006).

Searching for information in Internet has moved from a useful service on the edge of most Internet users' experience to the *de facto* interface for computing in the information age. John Battelle comments in his book *The Search* that "as the amount of information available to us explodes, search has become the user's interface metaphor. There is now all this information that is possible to get into your hands. Search is the attempt to make sense of it" (Battelle 2005).

In the summer of 2004, the Pew Internet & American Life Project released a research paper on the American usage of the Internet. It concluded that of all Americans who use the Internet, about 85 percent use search engines, or more than 107 million people in the United States alone (Battelle 2005), and more than two-thirds of them are active users of search. That means nearly 4 billion queries each month only in the most popular internet search engines. This study also concluded that the younger people are or the higher their educational attainment is, the more they search. The same study carried on in February-March 2007 reported that 71% of the American adults use the internet and that 91% of them used a search engine to find information.

Latest updated results coming from the World Internet Usage Agency show that in the first quarter of 2008 the 21.1% of the world population are internet users. That means a target group of 1.407.742.920 customers.

Simultaneously with the transition of the users to the online world, companies have done the same. As a proof of the importance of being visible in the online world results interesting to observe the ad-online spending in Internet. The President and CEO of IAB said in the last results of the *Internet Advertising Revenue Report* (InteractiveAdvertisingBureau and PricewaterhouseCoopers May 2007) that interactive advertising revenues continue to show solid growth as advertisers and agencies recognize that it is a medium that can uniquely exert influence on customer behaviour from product awareness, to purchase intent, to actual purchase and then brand equity. Figures coming from this study outline the evolution in ad spend and ad revenues in U.S. during 2006. Within these revenues, search marketing continues leading the revenues accounting for 40 percent of 2006 full year revenues. That means a nearly $7 Billion business.

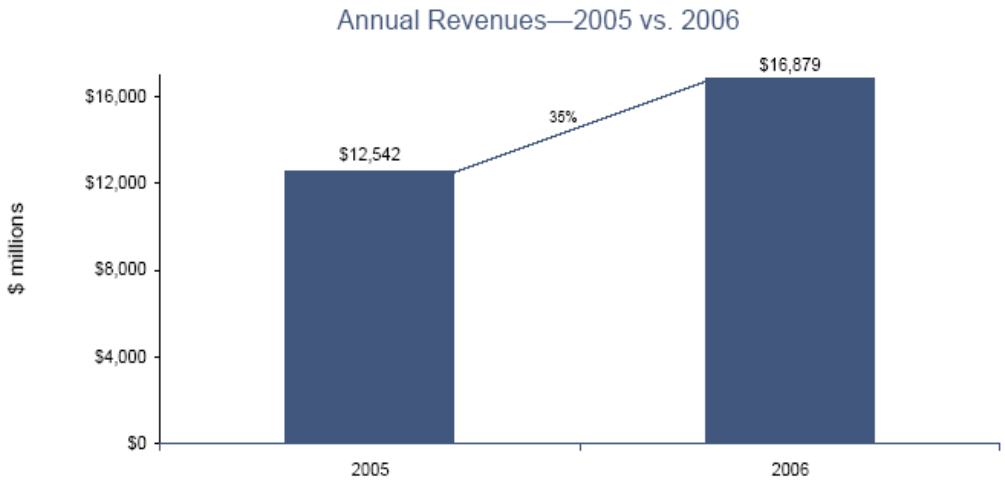


Figure 1-1: Annual ad online spending in Internet 2005 vs 2006

1.1. Why is search so hot?

Now more than ever, users rely on the Internet for information and news. Users have an information need associated with some task (Broder 2001). And this need is translated into a query posed to a search engine. Finding high-quality content is increasingly challenging. Providers of information and services know that their website is a key factor of their business and that, in a crowded information marketplace; searchers must be able to find it using search engines. Search engine advertising has grown tremendously in the past recent years, and prospects for continuing growth are strong.

Search is a big deal, and some figures are the proof: Americans conducted 6.9 billion searches online in February 2007 and nearly half of those were on Google. Google has 48.3% search market share in the U.S. compared with the 27.5% of Yahoo (News.Com April 2007).

Information providers and marketers know that Web users seek information on the Web prior to making a major purchase or information decision, and that users rely heavily on commercial search engines for most of their searches (Vine February 2004). Change is a key component of search marketing and it's what makes this industry so interesting and challenging. Search drove the Internet and continues to do so (Battelle 2005).

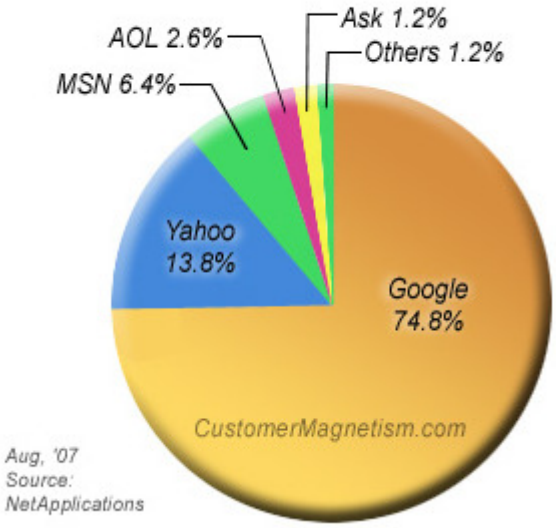
Daily Internet Activities		
According to our December 2006 survey, 65% of American adult internet users, about 92 million people, use the internet on an average day. Here are some of the things they do on a typical day:	Percent of internet users who report doing this "yesterday"	Most recent survey date
Use the internet	65	December 2006
Send or read e-mail	54	December 2006
Use a search engine to find information	41	December 2006

Source: Pew Internet & American Life Project

Figure 1-2: Daily Internet Activities

Web search represents a significant portion of Web activity as Figure 1-2 presents. At least a portion of searching is for products or services that the searcher will eventually purchase.

It is certain that we do ask a lot of the same question, but we ask far more that are unique, and therein lies the power of search. In other words, there are a few queries that have very high frequency, but quickly the graph flattens out into a massive tail, a tail that is extraordinarily long. And the power of search lies in that tail: no matter what the word is, somewhere on the Web there is most likely a result that contains it. In 2005, Google claimed that nearly 50 percent of the searches coming in on any given day – more than 100 million – are unique. And what is more, Google processes 75% of all the queries made each day on the Internet (Battelle 2005).



Source: http: //www.customermagnetism.com/

Figure 1-3: Market share owned by the different search engines

This new scenario where searching is the core of the World Wide Web offers far more advantages to both customers and marketers than previous models of activities like advertising, purchasing, getting information about products, etc. Technological advances have enabled advertisers to track the success of their Web-based ad placements. Moreover, the availability of this technology –along with specialized ad-buying programs where payment is made only if a link is clicked –has enabled advertisers to ensure greater return on investment of their Internet ad purchases in ways not offered by traditional media (Vine February 2004).

There is another relevant factor that contributes to the Search blow up during the last few years. This very important aspect is the Customer Acquisition Cost. Internet Retailer reported on a Piper Jaffray study, which finds that search is the most cost-efficient customer acquisition tool. In this study the approximate Customer Cost across various channels is calculated. Getting a new customer through search cost 8.50$, through Yellow Pages it costs 20$, through Online display ads 504, via email 60$ and direct mail 70$. The benefit of acquiring

users via search engines has a great potential mainly because these users are high potential users and they will maximize the ROI.

The Internet – in its various forms of websites, search engines, advertising, email, and professional and consumer reviews – is highly influential at every stage of the purchase decision making process, from first awareness to final decision making. In fact, the web influences purchase decisions – online and offline – more than any other factor. And corporate websites have more impact than any other kind of sites and services (DoubleClick November 2006).

Source: (DoubleClick November 2006)

Figure 1-4: Influencing factors when making online purchase decisions

In DoubleClick Touchpoints IV Report from November 2006 a consumer survey was made to identify and qualify the impact of various media and other influence factors on consumers' purchase decision making process. Its results are presented in the previous table and demonstrate the high influence that the online marketplace has on the searchers.

The following figure 1-5 compares the number of users among the main search engines in U.S. What can be extracted from this table is that search is centralized within the big four portals: Google, Yahoo, AOL and Microsoft. Several studies have also demonstrated that search users are loyal to the search engine they use Among them, 82% of search engine users report that they re-launch an unsuccessful query using the same search engine, but using more keywords (iProspect April 2006). This same figure was just 68% in 2002. If we have a look at the latest data provided on Februar 2007, the 6.9 billion searches conducted during this time in the US account for a 1 percent increase over January 2007 and a 19 percent increase over February 2006. Web surfers used Google for 3.3 billion search queries. Yahoo served 2 billion, MSN garnered 730 million, Ask.com served 348 million, and Time Warner sites, including AOL, served 338 million searches in February (SearchEngineWatch March 2007).

U.S. Online Searches by Engine, January 2007 and February 2007 (%)			
	January 2007 (B)	**February 2007 (B)**	**Change**
Total Internet population	100	100	N/A
Google	47.5	48.1	0.6
Yahoo	28.1	28.1	0.0
Microsoft	10.6	10.5	-0.1
Ask.com	5.2	5.0	-0.2
Time Warner	5.0	4.9	-0.1
Source: comScore Networks, 2007			

Source: (comScoreNetworks 2007)

Figure 1-5: U.S. Online Searches by Engine, Jan-Feb 2007

Taking in account the Global Search Market Share (Global, Excluding Canada & US), the figures from last February 2007 show that there is a polarization between the use of the different commercial search engines. Google concentrates 73.8% of the online searches, while Yahoo remains in a 10,8% followed by Google Images with a 6,9%. MSN and Ask.com have both 1,2% of the market share (Traffick 2007).

Online marketing spend is growing rapidly, but companies are realising that pay-per-click is only half the picture and strategic search engine optimisation also makes difference (Croft April 2006). Moreover, the search market has turned out to be increasingly competitive over the last two years and the art of good ***search engine optimisation (SEO)*** has become more valuable than ever as the paid search market has got more crowded (Bigmouthmedia September 2006).

From the point on time when internet became present in our everyday lives, online marketing has become also a basic element of the marketing mix. However, what was know as online marketing in the last decade has evolved in different ways until reaching the point of focusing online marketing activities around what is know as search marketing. This remains one of the most effective forms of online marketing. The main reason for this is that more consumers and businesses search the Internet when they intend to purchase products and services. Potential customers first search and then execute transactions. So, in order to maximize the ROI (Retunr on investment), the main goal is to drive as many potential customers as possible to the company's website.

With the high grow of information available at the World Wide Web, search becomes a necessity for people. And it is at this point when the marketing practises of enterprises extend their activities within the search engine world, trying to make them visible to the final users and drive traffic towards their sites. To accomplish this goal, there are basically two different ways of implementing

search marketing. The first one is search engine optimisation (SEO), which will improve the standing in the normal, or "organic", search results that search engines return when users introduce a query. The second one is search engine marketing (SEM), the keyword-based pay-per-click form that allows companies to bid on keywords and make sure their company comes higher up the search engines' paid-for listings (Brooks January 2006).

Regarding what online marketing is, we realize that is a very dynamic world, nearly changing everyday. There is a past, a present and a future and the companies must be aware and active in following the new trends for keeping their leadership online.

1.2. What exactly is a Search Engine?

A Search Engine is a programme that connects words a user enters (queries) to a database it has created of Web pages (an index). It then produces a list of URLs (and summaries of content) it believes to be the most relevant for the query of the user (Battelle 2005).

A Search Engine is basically composed of three major constituents: the crawler or spider, the index, and the runtime system or query processor, which connects a user's query to the index. The crawler is a particular software programme that hops from link to link on the World Wide Web, gathering as many pages as it finds and sending them back to be indexed. Crawlers are executed massively, parallel and simultaneously on many different servers. These requests bring back Web pages, which the crawler then conveys to the indexer. It also takes note of any links it has found on the page, and queues these links in its request file. Though the science behind crawlers is complex, what they do is pretty simple: they set off on an endless foray of dialling for URLs, and then they report back what they have found. Crawlers have long been the least visible of the search engine's components, but they are arguably the most important. The more sites they crawl, and the more frequently they crawl them, the more complete the index is. The more comprehensive the index

is, the search results pages (SERPs) that are returned for a particular query have a greater chance of being relevant.

Once the crawler sends its data back to a massive database which reads these documents and creates an index based on the words contained in each document, making it ready for the usage by searchers.

After the crawl data had been analyzed, indexed, and tagged, it is dumped into what is called a runtime index –a database ready to serve results to users. The runtime index forms something of a bridge between the back end of an engine (its crawl and index) and the front end (its query server and user interface).

It is important to note that each existing search engine base their search, and therefore their indices, on an own developed algorithm. This algorithm is what makes the difference among search. Moreover these algorithms should provide only meaningful results to answer the query entered by the searcher.

1.3. Traffic, Relevance and Monetization

Rita Vane explains in her article “The Business of Search. Understanding how Web advertising, partnerships, and the race for market dominance affect search tools and search results” (Vine February 2004) that commercial search engines require three key elements to ensure ad placements success. Traffic represents the flow of Web users to a search engine site. The main idea is to attract the maximum traffic as possible in order to maximize the probability that some of this traffic will turn into revenue for the search engine or will generate activity. Search drives clickstreams, and clickstreams drive profits. Current techniques to maximize this flow of users will be displayed in the next chapter.

The second element is relevance and represents the capacity of the search engine to deliver meaningful results to satisfy the user’s keyword query. The particular algorithms each commercial search engine uses define how search results are ranked for presentation to the user. These algorithms are regularly adjusted in order to improve the user experience.

For example, PageRank is the algorithm created by Google and ranks the results tanking into account the number of links to those pages. Moreover, PageRank apply different weights of relevance to some linking pages in the raw links analysis. The last step in achieving ad placement success is to monetize it. In other words, to convert the all-important traffic into revenue for the search engine. This monetization can occur in many different ways. When search engines deliver ads to search results pages, advertisers pay fees to the search engine or their designated ad-feed partner for every ad impression that is delivered. If the searcher clicks on the ad link –clickthrough –additional revenue may accrue to the search engine.

1.4. Relevance of topic

"Consumers are spending far more time online than anywhere else, so it makes sense to target them there" says Scott Gallacher, Sky's director of affinity acquisition (Jones March 2007). The ability of online sites to rank (i.e. have a good Search Engine Optimisation of the site) well will lead directly to more visibility and more clickstreams to the website.
In today's online world, a Website does not mean anything to anyone unless it can be found by its customers. Optimizing a site so that it appears high in the search engine seems to be as much an art form as it is a science (Trollinger November 2006).

The simple white little search box is the place where nearly 400 million people per month start on the internet, it is the No.1 gateway to the Net's vast commercial potential (Hof April 2007). With the clear leadership of Google, it can offer the most targeted and relevant advertisements alongside the results, drawing more clicks, more cash, more users.

This Master Thesis focuses on the optimisation of sites. Concretely in one of the factors that affects the ranking made by Google: PageRank. This algorithm, one of the most difficult to be manipulated artificially, determines for the most part the ranking served by Google when executing a query. To conduct this study

about how to optimize website optimisation practises focusing on PageRank's criteria, an introduction of the two search engine techniques (search engine optimisation and search engine marketing) will be done. Afterwards a detailed analysis of the criteria applied by PageRank will be given. These best practices should be implemented when optimising a website for the long term and with high quality.

2. Basic Principles of Search Engines and Search Marketing

2.1. Current situation

After having reviewed the situation of online marketing and the great importance of searching due to the broad adoption as the interface where both users and companies do meet in the virtual world, the next words will be focused on the major player, Google. Best practices that can be adopted in order to be relevant when a search is conducted will be analyzed. Concretely, the scope of the Master Thesis will be focused on Search Engine Optimization.

A brief description about what a Search Engine does when resolving a query and a broad view about search marketing will be given.

2.2. The evolution of Search Engines

In the evolution of search engines, three stages are identified in the evolution of web search engines (Broder 2001):

First generation: uses mostly on-page data (text and formatting) and is very close to classic IR. supports mostly informational queries. This was state- of-the art around 1995-1997 and was exemplified by Alta Vista, Excite, WebCrawler, etc.

Second generation: use off-page, web-specific data such as link analysis, anchor text, and click-through data. This generation supports both informational and navigational queries and started in 1998-1999. Google was the first engine to use this analysis as a primary ranking factor and DirectHit concentrated on click-through data. By now, all major engines use all these types of data. Link analysis and anchor text seem crucial for navigational queries.

Third generation: emerging around 2001, attempts to blend data from multiple sources in order to try to answer "the need behind the query". For instance, on a query like *San Francisco* the engine might present direct links to a hotel reservation page for San Francisco, a map server, a weather server, etc. Thus, third generation engines go beyond the limitation of a fixed corpus via semantic analysis, context determination, dynamic data base selection, etc. The aim is to support informational, navigational and transactional queries. This is a rapidly changing environment.

2.3. Exerting influence on placement in Search Engines

2.3.1. Search Marketing

From all the online marketing activities present in the Web during the last decade, search marketing remains one of the most effective forms (Brooks January 2006). Search is growing faster than any other sector of online marketing in terms of spend and accounts for 40% of the total ad spend online in the UK, for example (Brooks January 2006).

There is a reason why Google has become a household name so rapidly: because consumers like it and use it more than almost any other tool. And if consumers are there, then brands need to be there, too. It is no longer a question of whether or not to have a search strategy; it is simply a necessity for any brand genuinely attempting to reach its customers online.

There are essentially two methods of appearing in search engine results pages: paid search and natural search. Paid search results typically show up at the top or right of a search results page, while natural search results are displayed in the body of the search window. The big difference between the two is that paid search costs advertisers money generally on a per-click basis. Natural search is free and is regarded by many customers to be more credible than paid results.

The main purpose of search marketing is to focus on creating sites that will rank high in the long term and that will attract qualified visitors that will convert into buyers. It is quite typical to apply what is called “non solid marketing techniques” like hidden text, link farms, cross-linking and quite a few other techniques that are applied to get higher rankings at one time, increasing visibility on the Web. However, all these practises fail once the algorithm of the search engine is adjusted in a way that makes sure that these tactics will not work anymore.

2.3.1.1. SEO

Search Engine Optimisation is also known as Natural Search or Organic Search. The main focus of Search Engine Optimisation is looking at the technical design of a site, its written content and its structure and information architecture in order to ensure that the various semi-intelligent software packages that are constantly indexing the World Wide Web rank it as highly as possible against the search terms that the site’s target audience are using in the search box.

Exactly how the different search engines come up with their natural search rankings is a closely guarded secret – if the world knew exactly how Google’s algorithms worked, every site would be number one. So the algorithms are well kept secrets. However, what is know is that most search engine algorithms place major emphasis on the number of active links which a website has, the reason being that the more times a website is referenced by another site, the more relevant it must be to its topic. Yet, just knowing the number of websites

that link to a client's site is not enough: the content of the site must be optimised to appeal to the target market. Nowadays, the search engines are trying to index only real content visible to visitors and trying to avoid fraudulent practices which are only readable for the search engine's spiders.

The process of improving natural search performed in six steps (Trollinger November 2006):

1. **Know the terms**. It is important to optimize the "right" keywords – those words and phrases that people are actually searching for and that therefore offer opportunity for traffic. There are practical and very useful tools which help when looking for the "right" keywords. For example, there is a software available that offers a free service for evaluating the popularity of potential keywords.

2. **Get to know the competitors**. After having isolated the most important terms in a concrete site, it is critical to know who else is also there. It is good then to check in the most popular search engines who ranks higher and to analyze what the competitors are doing better in those terms.

3. **Crawl the site**. There are some Spider Simulators that will unveil how the spiders of the search engines will "see" the content of the site and also the content of the competitors. When entering a valid URL, the tool returns a listing of content as seen by a search spider – page title content, meta-tag content, internal and external page links and description data. The terms optimized in the first step should be prominent within the content.

4. **Knowing the density**. Keyword density is an important aspect of natural search success. There is a fine line between dense keywords and "being banned from search" keyword stuffing. If the density of the site is too low,

the rank will not be high; if it is too high, the site can be banned. A good target for keyword density is 2% - 3%.

5. **Adjust the content**. Having compared with the competitors and optimized the right keywords it is time to start making adjustments to the content of the site.

6. **Manage results**. The results must be monitored over time.

2.3.1.2. Search engine marketing: SEM

Search engine marketing, SEM, is used mainly for tactical and campaigned marketing initiatives that are specific to either time, product, geography or other such variables. A campaign focused on maximizing SEM is easy to control and as Peter Brown, lead maximiser at Google, says: “You can stop campaigns and analyse them, you can target a myriad of options depending on things such as geography or keywords, and you can have a high level of control” (Brooks January 2006).

SEM has to be integrated with all other channels to make it successful. It is important to link any activity through to all other marketing channels if a company wants to maximize the possible benefit. Then it is crucial to make SEM relevant for the customers. That is not to deliver users to the targeted business online but to convert them into customers once they are there. Most multichannel marketers today understand that search engine marketing is becoming a vital part of the marketing mix (Trollinger November 2006).

Running search marketing campaigns is normally accompanied by tracking referral sources in order to know which advertising or marketing campaign is driving visitors to the website. This method shows up exactly how much revenue comes from a specific campaign for calculating ROI.

There are different modes of SEM:

- Paid Inclusion Programs: in this modality, search engines and their adfeed partners guarantee that their search engine will list pages from the advertiser's website in its index. However, paid inclusion typically does not guarantee that the advertiser's pages will rank high.
- Paid Placement Programs: by contrast with Paid inclusion, generally this modality guarantee that a link to the advertiser's URL will be delivered in the search results on a matched keyword or keywords. Location of the delivered link generally governs the fees, so advertisers will pay more to be placed higher up the page in the search results.
- Pay-per Options: after the advertiser's link is delivered to the page, additional gradations of monetization are possible based on whether the link is clicked on or otherwise processed by the searcher. Advertisers have a variety of pay-per options which escalate in price as the deliverable moves closer to an actual sale.
    - Pay-per-impressions enables advertisers to pay based on how many users were served their ads. In this model, users do not have to click on the ad for monetization to occur.
    - Pay-per-click: it is an easy and low-risk way to find new business. Is based on choosing keywords, and when a user searches for that keyword, the ad appears as a sponsored result; the company then pays an amount, based on the keyword, for each click-through to the end site. Rates for keywords vary based on demand. This modality allows marketers to start small and then evaluate. At the beginning, the advertiser pays a certain amount of money for each click-through to the site in advance, and when this money runs out, the search engine will stop serving the ad. An analysis of how many customers the ad is driving to the web site can determine whether to invest more money for the original keywords or changing them. Nor is it enough just to decide on a list of the most important keywords for a certain brand. It is also

important to know when the target audience is most likely to go online to search. So keywords may be a few pence during one time of the day, and pounds at others. Managing PPC campaigns is a complex and time consuming process (Croft April 2006).

- Pay-per-lead enables advertisers to pay only for each "sales lead" generated. For example, an advertiser might pay for every visitor who clicked on an ad or site and proceeded to complete a form.
- Pay-per-sale allows advertisers to pay based on how many sales transactions were generated as a direct result of the ad.

All this activities start with keyword bidding. The more an advertiser bids, the higher his site will rank in the search engine selected. Google uses keyword bidding as a partial determinant of the placement of ads in its right sidebar (Vine February 2004).

- Contextual Search. As the search marketing arena becomes more crowded, advertisers are seeking ways to improve relevancy. This modality is a process that drives selected paid search results by user behaviour and perceived relevance as opposed to strict keyword matching. Google's AdSense is an example of contextual search: the programme places ads on pages of the websites that sign up for the program, and the ad selection is contextually based on what Google believes the page to be about. Overture's ContentMatch is a similar program.

3. Google

Google is a search engine composed basically of two important features which help to produce high precision results. The first one makes use of the link structure of the web and is called PageRank. The main goal of PageRank is to assign a quality ranking for each web page that is being crawled and indexed

within the web. The second feature makes use of the linking text or anchor text in order to improve search results.

The website where Google explains its technology claims that the heart of its software is PageRank. They continue describing it as a system for ranking web pages. PageRank was developed by Google founders Larry Page and Sergey Brin at Stanford University. In the website, they point out that while dozens of engineers work to improve every aspect of Google on a daily basis, PageRank continues to play a central role in many of their web search tools. This technique actually improves as the web gets bigger, as each new site is another point of information and another vote to be counted (SearchEngineLand 2007).

On the papers written by Brin and Page describing the early design goals of the Google Search Engine (Brin and Page April 1998), they emphasize the particularity of Google in comparison to other Search Engines: Google makes heavy use of the structure present in hypertext in order to produce better search results. It was actually designed in 1998 to crawl and index the Web efficiently and produce much more satisfying search results than existing systems.

Brin and Page were aware about the fact that the amount of information present on the Web was growing rapidly. In such a scenario, the former search engines based only on keyword matching usually returned too many low quality matches. Moreover, some web designers took measures meant to mislead automated search engines.

Google was designed to overcome many of these problems, both in quality and scalability. As the amount of information grows, tools that have very high precision about what is considered human notion of relevance becomes a need. Using the hypertextual information can improve search quality. In particular, link structure and link text provide a lot of information for making relevance judgements and quality filtering. Google makes use of both link structure and anchor text.

Other important factors in designing Google were the ability to handle queries quickly, at a rate of hundreds to thousands per second and offering a great deal of usage. Ease of use was important as long as a main design goal was to build systems that reasonable numbers of people can actually use. A final design goal was to build an architecture that can support novel research activities on large-scale web data. However, this last goal became blurred as soon as Google became a company and no longer a University research project.

3.1. Link Structure of the Web

The current structure of the crawlable Web has roughly 155,583,825 million nodes (pages) and much more edges (links) (Netcraft 2008). Every page has some number of forward links (outedges) and backlinks (inedges). It is impossible to know whether all the backlinks of a particular page have been found but if this page is downloaded, it is possible to get to know all of its forward links at that time (Page, Brin et al. 1999). Web pages vary greatly in terms of the number of backlinks they have. Generally, highly linked pages are more “important” than pages with few links. In this way, PageRank provides a sophisticated method of conducting citation counting and what is more important is that PageRank also considers the fact that in many cases simple citation counting does not correspond to our common sense notion of importance. For example, if a web page has a link off the Yahoo home page, it may be just one link but it is a very important one. This page should be ranked higher than many pages with more links but from obscure places. PageRank is an attempt to see how good an approximation to “importance” can be obtained just from the link structure.

3.2. PageRank

3.2.1. What is PageRank

PageRank is one of numerous methods Google uses to determine the relevance or importance of a page. PageRank might be the best method used by Google when ranking pages. PageRank has the task of condensing every page on the World Wide Web into a single number, its PageRank value which aims to be a global ranking of all web pages, regardless of their content, based solely on their location in the Web's graph structure (Magazine 2007). It evaluates two things. The fist one is how many links there are to a web page from other pages, and the second point analyzed regards the quality of the linking sites. This brings up the possibility to order search results so that more important and central Web pages are given preference (Page, Brin et al. 1999). PageRank is based on incoming links, but not just on the number of them. Also relevance and quality of them are important: not all the links weight the same when it comes to PageRank.

An accessible definition of PageRank tells that a page has high rank if the sum of the ranks of its backlinks is high. This covers both the cases when a page has many backlinks and when a page has a few highly ranked backlinks.

The Google principle is based on a voting system. Basically it says that if Page A links to Page B, then Page A is saying that Page B is an important page. PageRank also factors in the importance of the links pointing to a page. If a page has important links pointing to it, then its links to other pages also become important. The actual text of the link is irrelevant when discussing PageRank (Ridings 2002).

When all other factors such as title tag and keywords are taken into account, Google uses PageRank to adjust results so that sites that are deemed more "important" will move up in the results page of a user's search accordingly. To

summarize in a simplified way, the path that Google follows when a search term is launched are:

1. Find all pages matching the keywords of the search.
2. Rank accordingly using “on the page factors” such as keywords.
3. Calculate in the inbound anchor text.
4. Adjust the results by PageRank scores.

It is important to note that PageRank is a multiplier and is not just simply added to the score. Thus, if a page had a PageRank of zero, it would rank at the very end of the SERP.

The actual amount of effect that Google’s PageRank has on the ranking of a sites is debated, and it is probably prudent to say that Google will not unveil it. What is sure is that a web page’s PageRank does play a role in Google’s indexing and Google’s ranking. The higher a web page’s PageRank, the more frequently it will be crawled and refreshed. While in most cases, a higher PageRank will accompany a higher-ranking site, but this is not always the case.

How significant is PageRank?

The significance of any one factor in search engine algorithms depends on the quality of information it supplies (Ridings 2002). A factor’s importance is known as its weight, i.e., how much we trust the information provided by that factor.

But with human nature being what it is, the easier something is to influence, the more it is manipulated. PageRank is, without doubt, one of the hardest things for a Webmaster to manipulate ethically. However, it is possible to generate links to a website from other sites fairly simply through the use of link farms and guestbooks. Google frowns upon this kind of abuse, and many sites that have tried this have had their PageRank influence blocked. Though, it must be said

that the abuse is still rampant, and that it can have an influence on PageRank. So, whilst not easy to do, PageRank is still subject to manipulation.

The influence of PageRank over the results is substantial. However, its usage and capabilities must not be over-estimated. The final ranking in Google is due to a mix of factors, of which PageRank is only one.

To examine the worth of PageRank, we need to first look at its premise, and how accurate it is. Basically PageRank says:

> *If a page links to another page, it is casting a vote, which indicates that the other page is good. Following the same idea, if lots of pages link to a page, then it has more votes and its worth should be higher.*

The basic implication here is: people only link to pages they think are good. It should not be hard to realize that this premise is wrong. A few of the reasons people link to pages other than ones they think are good are:

1. Reciprocal links: “link to me and I will link to you”.
2. Link Requirements: “using our script requires you to put a link to our page” or “we will give you an award solely because you link to our page”.
3. Friends and Family: “my mum’s site is here, my dad’s site is there. My dog’s site is here”.
4. Free Page Add-ons: “this counter was provided by www.linktocountersite.com”.

Furthermore, anybody who has a top-ranking site will tell to a webmaster that it tends to get links from new sites. This is not necessarily because it is good (although they generally are). Assume a Webmaster is setting up a new site and they are looking for some outbound links. Nowadays, one of the first things they do is a Google search for similar sites. The links they end up with may not necessarily be the best sites, but merely the easiest ones to find. If PageRank

influences rankings, and if they subsequently link to those pages – the new Webmaster will be adding to the inaccuracies in the judging of the quality of a page. The same is true when these new Webmasters use the Google Toolbar PageRank indicator to choose whom to link to.

To put this in another way: PageRank is determined by the links pointing to a page. But if PageRank itself has an influence on the number of links to a page, it is influencing (in a circular way) the quality of that page. The links are no longer based solely on human judgement. If a Webmaster picks their outbound links by searching on Google or by looking at the Google Toolbar (even if this is only part of their judgement), then there is a corresponding increase in a page's PageRank. This increase is not solely because it is a good page, but because its PageRank is already high. So the basic question is when PageRank is not worthwhile.

Performing any broad search on Google, it will appear as if the searcher would have found several thousands results. However, it is only possible to view the first 1000 of them. Understanding why this is so explains why efforts should always concentrate on "on the page" factors and anchor text first, and PageRank last.

The answer to speeding up the search is to get a subset of documents that are most likely to be related to the query. This subset of documents needs to be larger that the number of search results Assume that a search on Google returns 200.000 results (that is the number that actually gets shown). To select the subset of results, the search engine does a query to the whole database using 2 or 3 factors, finding the 2000 documents that rank highest for them. Then the engine applies all the factors to those 2000 results and ranks them accordingly. Because there is a drop in the quality of the results (not the pages) at the bottom on this subset, the engine shows the first 1000. PageRank is almost certainly not one of those factors chosen when selecting the subset. The search engine is looking for pages that are on-topic in creating the subset of

2000 pages. If PageRank were included in that list we would get a lot of high PageRank pages with topics that are only slightly related (because of the second factor applied), but this is not the final objective.

So the important point that can be extracted from here is to do enough "on the page" word and/or anchor text work to get into that subset of 2000 pages for the chosen key phrase, otherwise the high PageRank will be completely in vain. PageRank means nothing unless the site has a high enough ranking from other factors to make it into the first subset.

The difference between PageRank and other factors

To assess when PageRank is important and when it is not, we need to understand how PageRank differs from all other ranking factors. Considering that the final rank score is a multiplication of the score for all non-PageRank factors per the actual PageRank score it is interesting to know the difference between these two groups and how we can influence them.

To do this, here is a quick table that lists a few other factors and how they add to the ranking score:

Factor	Observations
Title tag	Can only be listed once.
Keywords in body text	Each successive repetition is less important. Proximity is important.
Anchor text	Highly weighted, but like keywords in body text, there is a cut off point where further anchor text is no longer worthwhile.
PageRank	Potentially infinite. It is always possible to increase the PageRank significantly, but it takes work.

Figure 3-1: Influencing factors on ranking score

From this table we can extract that all other ranking factors have cut off points beyond which they will no longer add to the ranking score, or will not add significantly enough for it to be worthwhile. PageRank has no cut off point. This means that improving either side of the equation can have a positive effect. However, because non-PageRank factors have a restricted maximum benefit, the actual PageRank score must be improved in order to compete successfully.

The worth of PageRank in different strategies

With ranking factors other than PageRank, there is a score beyond which the slow down in the rate that any factor adds to this score is so insignificant that it is not worthwhile. This is called the non-PageRank factor threshold (Ridings 2002). This threshold defines when high PageRank is worth striving for and when it is not. To illustrate this, one can put an example figure on this of 1000 (actually, this threshold is a hypothetical line, it has no value).

If we have a query where the results are Page A and Page B, then Page A and B have scores for that query which are the total scores for all ranking factors (including PageRank). Assuming that Page A's score is 900 and Page B's score is 500, obviously, Page A will be listed first. These are both below our hypothetical non-PageRank factor threshold (fixed to 1000), thus without any change in PageRank, it is possible for Page B to improve their optimization to beat Page A for this particular query. These queries are commonly thought of as less competitive queries.

Presupposing that Page A raises its score to 1100, suddenly Page B cannot compete in the SERPs without increasing its PageRank. In all probability, Page B must also improve in all the other ranking factors, but an increase in PageRank is almost certainly necessary. There are also lots of queries like this in Google, which are more commonly though of as more competitive queries.

The conclusion to draw from this non-PageRank factor threshold is that in order to be competitive it is necessary for sites to get scores beyond the Threshold. Failing to do so means that these sites can be easily beaten in the search results for their respective query terms. Optimizing “on the page factors” is the quickest way to approach the non-PageRank threshold. However, it is not feasible to move above the non-PageRank factor threshold without PageRank.

If we compare two different ranking strategies, one that considers PageRank to be unimportant and the second one which takes into account PageRank and considers it very important, one will see that there are advantages and disadvantages for both strategies and that it is possible to use them to different degrees, depending on the strategy that best suits the site’s style.

- Strategy A – “PageRank is unimportant”

  These sites are optimized by improving their pages through “on the page” factors. The designers of these websites understand the basics of anchor text but they do not care at all about PageRank.

  In this situation, they are reaching the non-PageRank Threshold very quickly because they are maximising the “on the page” factors. They focus on carefully choosing keywords. As long as their content is good, high-ranking sites (over time) will tend to get linked to. Whilst they did not directly ask for it, as slow trickle of sites will begin to link to them and give them PageRank, which helps to consolidate their position.

- Strategy B – “PageRank is important”

  This strategy could correspond to those pages in the result that have no content, but great rankings (often occurs with big brands). This strategy understands a lot about PageRank and concentrates heavily on it.

  In this second strategy, in the opposite way from the first one, one bears in mind the PageRank factor and finds oneself getting non-PageRank

factors. The reason for this is that increasing PageRank requires links, and links have anchor text. Thus, by carefully choosing the anchor text linking to his page, this strategy automatically increases the non-PageRank factor scores whilst obtaining the high PageRank score.

These situations are two extremes but they are useful to extrapolate some advantages and disadvantages of each approach:

	Advantages	Disadvantages
Strategy A	• Quick entry into the results pages. • Self-generating links limit the amount of work needed.	• Securing the results to make it more difficult for competitors to compete takes more work. • Slower to react to new competitors.
Strategy B	• Solid position. Can easily modify "on the page" for a quick boost if required. • Probably higher traffic from non-search engine sources.	• Slower entry to results pages. • Difficult to do well. • Increased likelihood of tripping spam filters.

Figure 3-2: Ranking Strategies

It is important to remark that for some keywords there is a heavy competition and in this case it is necessary to do everything possible to maximize the ranking score. In such situations it is impossible to rank highly through non-PageRank factors alone (mainly because initially the site will not be listed high enough to be noticed and linked to).

Under really heavy competition it is true that a site cannot rank well unless the actual PageRank score is above a certain level. In this way a "Minimum PageRank level" exists. For queries that do not have heavy competition, this

level is easy to achieve without even trying. However, where heavy competition exists, non-PageRank Factors are just as important until they reach the non-PageRank factor threshold. This is why careful keyword choice can help to avoid the extensive work associated with highly competitive search phrases.

3.2.2. Algorithm

Quoting from the original Google paper (Brin and Page April 1998), PageRank is defined like this:

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.
PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.

In the above cited equation T1 ... Tn are pages linking to page A. C( ) is the number of outbound links that a page has and d is a damping factor.

When analyzing the iterative algorithm in detail we can observe the following factors:

- PR(Tn): it refers to the PageRank assigned to each individual page. That means that each page has a notion of its own self-importance. That is PR(T1) for the first page in the web to PR(Tn) for the last page.

- C(Tn): each page spreads its vote out evenly among all of its outgoing links. The count, or number, of outgoing links for page 1 is C(T1) and C(Tn) for page n, and so on for all pages.
- PR(Tn)/C(Tn): this factor indicates that if the page in which PageRank is being calculated has a backling from Page n the share of the vote that the page gets is PR(Tn)/C(Tn).
- d: this is a damping factor to avoid the other pages having too much influence. Google is thought to apply a damping factor of 0.85 (that is the value mentioned in the Stanford papers). This factor assures that the convergence of the recursive procedure occurs (remember that the final values of one stage of the calculation become the starting values of the next stage) and sets the stop of the calculation (Ridings 2002). This convergence basically means that whatever the starting values are, after running the calculation a number of times we will end up with the same final values and that these values will no longer change if further iterations are done. These final values are known as limiting values and Google no longer needs to expend processing power on calculating the PageRank (Ridings 2002).

Some facts can be extracted from the algorithm of PageRank and its functionality. The first of them is that PageRank can be seen in a simpler way as: PageRank equals 0.15 + 0.85 * (a “share” of the PageRank of every page that links to it). The “share” that certain page passes is its PageRank divided by the number of outbound links, and this value is shared equally between all the pages that it links to (Magazine 2007). In this way, the PageRank algorithm distributes its established PageRank across all of the outbound links. For example: a web page with PageRank 8, PR8, and one link on it, the site linked to would get a fair amount of the PageRank value. On the contrary, if the web page had hundred links on that page, each individual link would only get a fraction of the value. From this can be concluded that a link from a page with PR4 and five outbound links is worth more than a link from a page with PR8 and

hundred outbound links. The more links there are on a page, the less PageRank value is passed to the linked pages.

In the original Google paper, Sergey Brin and Lawrence Page gave an intuitive justification to explain the calculation of PageRank. This model will be used in other sections of this Master Thesis in order to give another more approachable point of view to the results given.

The mentioned model assumes that there is a “random surfer” which is given a web page at random and keeps clicking on links, never hitting “back” or eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. The damping factor d is the probability that at each page the “random surfer” might get bored and requests another random page. One important variation is to only add the damping factor d to a single page, or a group of pages. This allows for personalization and can make it nearly impossible to deliberately mislead the system in order to get a higher ranking.

The founders of Google cited another intuitive justification. This second one is based on the fact that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank. Therefore, pages that are well cited from many places around the web are worth looking at. For example pages that have perhaps only one citation from high quality human maintained indices such as Yahoo! or DMZ are worth looking at. If a page was of low quality, or was a broken link, it is quite likely that Yahoo’s homepage would not link to it. PageRank handles both these cases and everything in between by recursively propagating weights through the link structure of the web. In fact, Google has a place where pages are listed because human editors have selected them, rather than being crawled. It is called the Google Directory and it is based on work done by editors at the Open Directory Project (SearchEngineLand 2007).

The PageRank of each page depends on the PageRanks of the pages pointing to it. But it is not possible to know what PageRank those pages have until the pages pointing to them have their PageRank calculated and so on. This problem however is easily solved as long as PageRank can be calculated without knowing the final value of the PageRank of the other pages. That seems strange but, basically, each time the calculation is run a closer estimation of the final value is obtained. So all that is necessary to do is to remember the single values of the calculation and repeat the calculations many times until the numbers stop changing much. In order to estimate how many iterations are needed in order to converge to the final value for big networks is needed the damping factor. The damping factor contributes to control the number of iteration needed until the values convergence. For example, if the damping factor is too high then it takes long for the numbers to settle, if it is too low then the numbers swing about the average like a pendulum and never settle down (Rogers 2002).

3.2.3. Influencing Factors

Having control over PageRank can be seen from two different perspectives: on the one side, the one of Google; on the other, the one of the Webmasters.

Google has shown that they can and will modify the data on which PageRank is based. The primary example for this is what has become known as PageRank Zero (PR0). Rumour has it that when Google wants to penalize a page, it is assigned a PageRank of zero. As PageRank is a multiplier, this will obviously always list PR0 pages as the very last entry on the SERPs. The second way Google has to penalize is the same that they do apply to link farms. Google has shown that they are capable of ignoring links they believe have been artificially created removing this link entry from the matrix (Ridings 2002).

Webmasters do also have some control over PageRank. It is, however, the most difficult of ranking factors to control (Ridings 2002). However, a well-

worked PageRank added to optimization mix techniques can make a difference in front of the competitors.

There are three fundamental areas to look at when trying to optimize the PageRank of a site:

The links the site chooses to have a link to it, i.e., which ones it chooses and how much effort it put on getting them.
Who the site opts to link out to, and from which page of the site does it place the link in order to maximize PageRank Feedback and minimize PageRank leakage.
The internal navigational structure and linkage of the site's pages, in order to best distribute PageRank within the site.

3.2.3.1. Inbound Links

External linking is the gravest factor in determining PageRank, and is the place where webmasters have the least control. There is no way to force another web master to link to a certain site, especially when they already have a high PageRank. However, it is certain that PageRank influences the frequency Google updates the site. In other words, the higher a web page's PageRank, the more frequently it will be crawled and refreshed (SearchEngineGuide 2002).

There is an important point when contacting other websites to link to a web page in the attempt to build and increase the PageRank of it. When Google is determining the PR of Page A by page criteria and title tags of Page B and others, it will also take into account the on page criteria and title tags of Page B and other links that are pointing to Page A. If Page A's target keywords and theme are about a certain topic, then this is what Google will look for in the external links that are pointing to the site. As long as Google use on page criteria and title tags when determining PageRank, these web pages from

where we request links from should be relevant and of the same theme and market of the page that someone is requesting they link to.

Another important part of the external linking campaign is the actual links and the way that they are formatted. This topic will be explained more detailed in the Anchor Text chapter. Only point that to maximize results from the links pointing to a page, it is necessary that the link text pointing to the site includes the keyword and is rich and descriptive (SearchEngineGuide 2002).

There is another important point that must be made that will be crucial to the successful building of PageRank. One common mistake in creating PageRank is that webmasters or search engine optimization professionals will contact other webmasters and request that they only link to the homepage or the top level of a section. This can have two effects, which will be explained shortly, but the pages that are below these pages linked to will not encounter the full effects of being linked to. Requesting a link to a top-level page will have a positive effect on this page. If the PageRank of this page increases, then the page below it will increase (if the internal structure is correct), but it will still be one PageRank number below the page above it that was linked to. In the case that the pages that are deep in the structure of the site have quality content it will be necessary to go further in order to increase the PageRank for those pages. These sub-categories and pages that are below are the niche keywords and the ones that are going to bring the targeted traffic that is easily converted (SearchEngineGuide 2002). Hence, the importance of increasing the PageRank of these pages and sub-categories is understandable.

One thing that should also be taken into account is the fact that multiple votes to one link from the same page cost as much as a single vote. It is reasonable to assume that a page can cast only one vote for another page, and that additional votes for the same page are not counted (Magazine 2007).

As mentioned before, it is not always the best strategy to get links from pages that have the highest Toolbar PageRank. The PageRank from an individual page is shared out amongst the links on that page (as it is clearly indicated in the PageRank calculation). Consequently, links from pages that have the same PageRank are not always created equal. It depends on how many other links a link is sharing the links page with. For instance, a link from a page with PR of 4 might be better than a link from a page with a PR of 6 if there are less total links on the PR4 page. There is another reason why that type of linking strategy might be the best; sites with high PageRanks are often fussy about which sites they will link out to, making them harder to get linked from than lower PageRank sites.

Regarding the effect of the linking feedback can be useful when choosing from which pages links are to obtain from. E.g. there are two separate pages on other people's sites which both have a PageRank of 4. Both of these have ten links to other pages. In the case that a page in the home site where the webmaster would like to be linked already has a link to the page on the second site. By getting a link from the second site one generates feedback and receives a higher PageRank than if one had gotten a link from the first site. That is an oversimplification but it is important to remember that the number of links on the page linking to ours will alter the amount of feedback.

These facts have an influence on PageRank but their factual results are complicated to work out completely for a given page's situation. The advice that should be taken into account is to get links from sites that seem appropriate and have good quality, regardless of their current PageRank. If they are relevant to a site, and are high quality sites, they will either help the site's PageRank now, or will do so in the future. To really get the PageRank humming, a good strategy is to achieve a listing in DMOZ and Yahoo Directory and enjoy the artificially enhanced PageRank that they provide – however, a listing in DMOZ and Yahoo does not give a site a special PR Bonus but increases its popularity (what is know as social bookmarks). Google uses Open Directory Project (DMOZ.org) to

power its directory. Coupling that fact with the observation that sites listed in DMOZ often get decent and inexplicable PageRank boosts, has lead many to conclude that Google gives a special bonus to sites listed in DMOZ. This is simply not true. The only bonus gained from being in DMOZ is the same bonus a site would receive from being linked to by any other site. However, DMOZ data is used by hundreds of sites (Magazine 2007).

3.2.3.2. Outbound Links

The main rule to follow when considering links out from your site is to keep PageRank within your own site and reduce the amount of PageRank let out of your site. This does not mean that you will lose PageRank from your site by linking out, but that the total PageRank within it may be lower than it could have been had you not linked out. From this, we can derive that the best outbound link PageRank scenario occurs when the outbound link comes from a page that has both a low PageRank and a lot of links to pages on your site (Ridings 2002).

One way to achieve this is to write reviews of the sites a site link out to on a separate page of your site, and by providing a link to those reviews along with each hyperlink to the external site. Optionally, it would be right if these pages open in another window but do not open these pages using Javascript because the spiders cannot yet follow this. Make sure that the review page also links back to a page in the own site that is high up its structure.

Moreover, it is very good if someone links to a site's review page, so in addition they may let the site they have linked to know that they have reviewed them.

The result achieved by appliying the mentioned techniques is due to a mix of factors: the power of adding extra pages which increase the whole site's PageRank, the feedback produced and of drawing PageRank away from the outbound links pages. Fundamentally, however, adding extra internal links to

your outbound pages is one of the most (if not the most) important internal factors to improve your PageRank. This boost may not be as much as from getting a new outside link to your site, but it is also much easier and more beneficial for your site's readers.

Pages linking to each other can create a feedback effect that may increase the PageRank of these pages. An interesting thing about PageRank feedback is that it can be used as an advantage via the internal navigational structure of a site (HighRankings 2007). PageRank Feedback is the principle that explains how, sometimes, it is good to link to outside pages. The fact is that linking out to a page that links back to it causes its own PageRank to rise. This has occurred, not because we are generating PageRank, but because it is taking it from the system as a whole.

It is also possible to find links that point nowhere. They are called dangling links. Dangling links are simply links which point to any page with no outgoing links. They do not have an impact on PageRank but affect the model because it is unclear where their weight should be taken into account, and there are a large numbers of them (Magazine 2007). Because dangling links do not affect the ranking of any other page directly, we simply remove them from the system until all the PageRanks are calculated. After all, the PageRanks are calculated they can be added back in without affecting things significantly (Brin and Page April 1998).

3.2.3.3. Internal navigational structure and linkage

Internal linking is a factor in the PageRank of the pages within a site. It is most common to see the homepage to have the highest PageRank of the website. The linking structure within the site should be optimized. It is a fact that digging deeper into the structure of a website, the PageRank reduces. There are several hypotheses that discuss why this phenomenon takes place. Some think that Google does this as a result of the deep structure as it does not prefer it.

Others think that this PageRank reduction takes place as a result of the smaller amount of internal linking that takes place. The question that arises from here is how the PageRank of these pages that are deeper in the site can be increased. Since the internal linking of a site plays a factor, not in increasing PageRank, but in sharing the PageRank of the site, it becomes important to review the internal linking structure so the pages that are targeting the refined keywords within the site are becoming the PageRank necessary to maintain its importance (SearchEngineGuide 2002).

The main reason when optimizing the internal linking of a website is to keep the visitors within the site as long as possible. This situation increases the chances that the visitor buy the product or will click to some ads (Pandia 2007).

Most enterprise-size websites have many pages with no or very few incoming links and fewer pages that get a lot of incoming links. This problem can be alleviated by linking to link-poor pages from link-rich ones manually or restructuring the website. Log files are a good tool for collecting information about a website: links, clicks, search terms, errors, etc. In this case, they can be of great use to identify the pages that are getting a lot of links and the ones that are getting very few (Batista 2007).

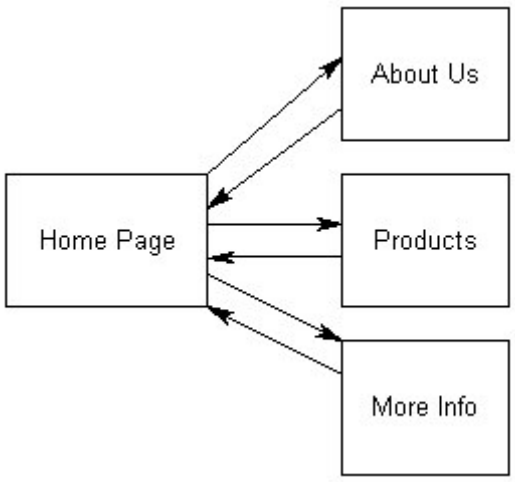
As Rogers claims in his paper (Rogers 2002), linking to a site map on each page increases the number of internal links in the site, spreading the PageRank out and protecting the site against the vote “donations”.

As long as PageRank is really about pages having a vote, the more pages a site has in the index of Google, the more overall vote it is likely to have. Or simply, bigger sites tend to hold a greater total amount of PageRank within their site (as they have more pages to work with). However, to get a high PageRank it is not enough to have tens of thousands of pages. These pages must also be in the index of Google. To achieve this, they must contain enough content for Google’s algorithms to consider them worthy of being added to the index

(Ridings 2002). By developing more content for a site, one also increases the PageRank for this site. This is hard work and a slow process. Nonetheless, creating pages that people will want to link to is killing two birds with one stone as one creates PageRank from both directions. Or, in basic terms, the best internal thing to do to build PageRank is to write a lot of good content. It should be ensured that pages are not overly short or excessively long and breaking the content into several pages where necessary.

There are three different ways in which pages can be interlinked within a site. In practice, sites might use a combination of these. Applying a combination is fine and normal as long as it takes into account the different sections and how they affect the PageRank. The first type is the hierarchical structure, looping is the second type and extensive interlinking the last one.

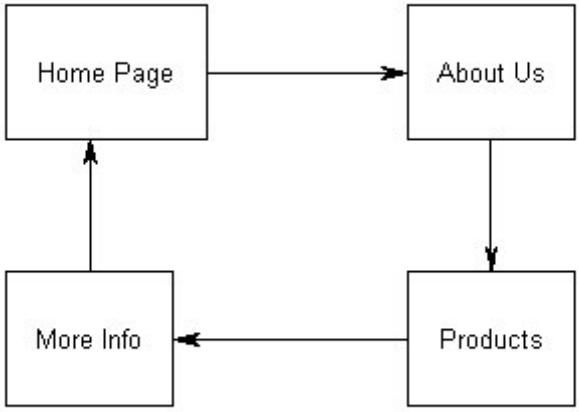
Hierarchical



Source: (Ridings 2002)

Figure 3-3: Hierarchical Internal Linking

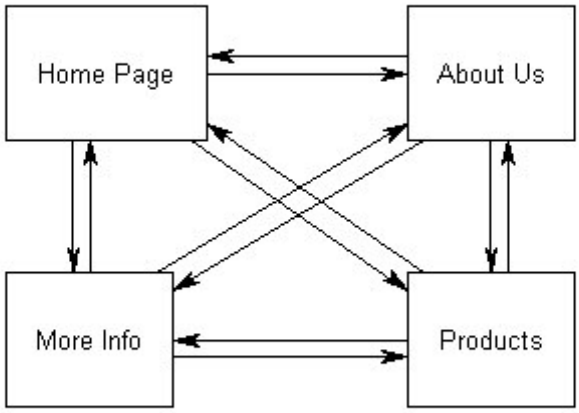
Looping



Source: (Ridings 2002)

Figure 3-4: Looping Internal Linking

Extensive interlinking



Source: (Ridings 2002)

Figure 3-5: Extensive Internal Linking

Because there are no inbound or outbound links, these sites are restricted to have the same total PageRank across all pages in these different structures, PR4. However, the hierarchical structure can alter the distribution of the PageRank. It can be also achieved with other combinations, but they must

contain more of the properties of the hierarchical structure than the looping or extensive interlinking. The ability to shift PageRank to designated pages is the webmasters' easiest way to manipulate PageRank. This methodology can be applied to pages that want to rank high for extremely competitive keyword phrases, or to pages that must compete in a large number of keyword phrases.

If one considers the non-closed systems, that means sites that have got inbound and outbound links, one conclusion is extracted: the extensive interlinking strategy retains the most PageRank within the site. Following this is the hierarchical strategy and lastly the looping strategy.

This is just a principle and in practice each site must choose the appropriate structure for sub-sections of the site. E.g. as more pages are added into the hierarchical structure, it becomes more successful because more links are being added to the page that contains the outbound link. However, as more pages are added, the PageRank of the home page is also diluted.

3.2.4. Visualization of PageRank: the Google Toolbar

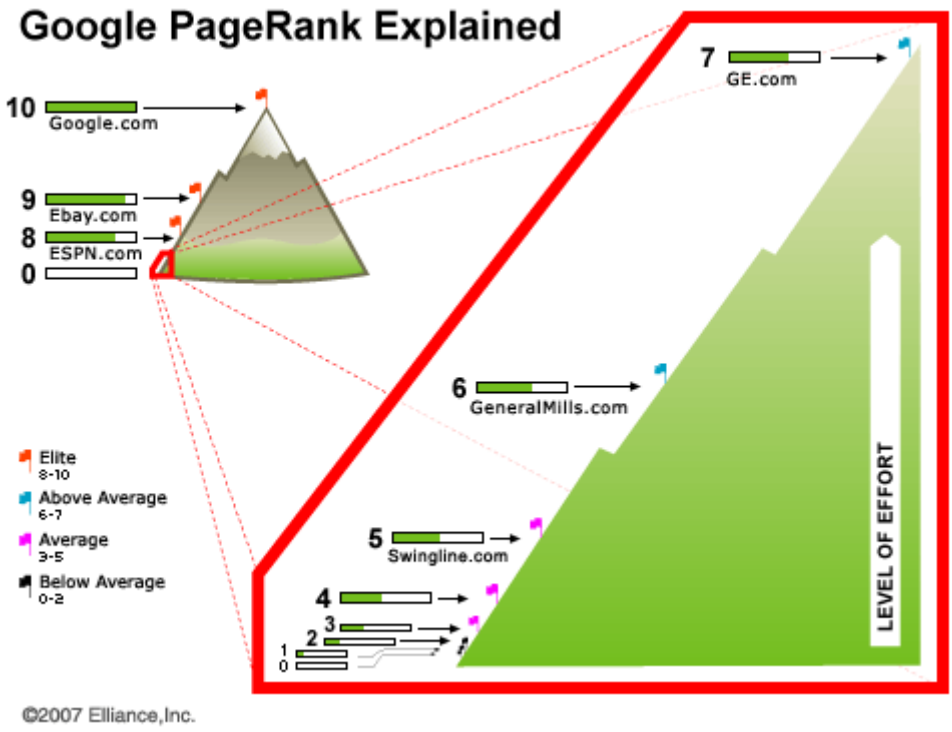
PageRank can be displayed on the toolbar of the web browsers if the Google Toolbar software is installed. It will show a bar graph at the top of the browser displaying the PageRank value of the page being browsed. Holding the mouse over the bar shows a number from zero to ten which tells the only thing right now about the value of PageRank. This value is not very accurate as long as it has some limitations (Ridings 2002). The first limitation is that the toolbar sometimes guesses. If a page which is not indexed is entered, but a page that is very close to one indexed by Google, then it will provide a guesstimate of the PageRank. However, this guesstimate is worthless because it is not featured in any of the PageRank calculations. The only way to tell if the toolbar is a guesstimate is to type the URL into the Google search box and see if the page

shows up in the SERPS. If it does not appear, it means that the toolbar is guessing. The second limitation is caused by the fact that the toolbar is just a representation of actual PageRank. Than means that whilst PageRank is linear, Google has chosen to use a non-linear graph to portray it. So on the toolbar, to move from low PageRank values (for example 2 to 3) takes less of an increase than to move from higher PageRank values (from 4 to 5, for instance).

Despite the fact that this Toolbar only goes from 0 to 10 it is guessable that the original scale is a logarithmic scale (Magazine 2007). PageRank is a floating-point number, so the scale going from 0 to 10 is a linear approximation of the real scale. The fact that the each PageRank level is progressively harder to reach strengthen the idea of the logarithmic scale. Another reason that supports the assumption of the logarithmic scale is the fact that because the rank of various different documents can vary by orders of magnitude, it is convenient to define the logarithmic rank (Page 2001).

In conclusion, it is impossible to know the exact details of the scale because the maximum PR of all pages on the web changes every month when Google does its re-indexing.

What seems to be happening is that the toolbar looks at the URL of the page the browser is displaying and strips off everything down the last "/" (i.e. it goes to the "parent" page in URL terms). If Google has a Toolbar PR for that parent then it subtracts 1 and shows that as the Toolbar PR for this page. If there is no PR for the parent it goes to the parent's parent's page, but subtracting 2, and so on all the way up to the root of your site. If it cannot find a Toolbar PR to display in this way, that is, if it does not find a page with a real calculated PR, then the bar is greyed out (iprcom 2007).



Source: (Magazine 2007)

Figure 3-6: Google Pagerank Explained

There are several online tools that make use of the Google Toolbar. For instance, iWebTool has now developed an online PageRank viewer that adds PR-bars to all links on a particular web page. External links are marked with EX, and links with a no follow tag with a red X (iwebtool 2007).

However, highlighting PageRank in search results does not help the searcher. That is because Google uses another system to show the most important pages for a particular search you do. It lists them in order of importance for what you searched for. Adding PageRank scores to search results would just confuse people. They would wonder why pages with lower scores were outranking higher scored pages (SearchEngineLand 2007).

In contrast, if someone is looking at a single page, such as when someone is surfing the web, the searcher no longer wants the search ranking but rather an idea of how important or reputable that page might be. This is where PageRank makes more sense.

Of course, SEOs and others may want PageRank in search results (Google does not offer this option for searchers). In order to visualize PageRank together with search results, there are several tools as the one discussed before.

However, these PageRank scores that are visible through the toolbar is something different from what is called “internal” PageRank.

Internal PageRank are the PageRank scores that Google uses as part of its ranking algorithm. The scores are constantly being updated. In contrast, the PageRank scores that Google allows the world to see – Toolbar PageRank - is a snapshot of internal PageRank taken every few months.

The important point here is that if there is a brand new site, it will likely have a low or no PageRank score reported in the Google Toolbar. That might concern the webmaster of the site, even though its main consequence only is that it will get crawled regularly (the higher the PageRank, the more likely Google will revisit the pages of the site). It also has an impact on the ranking ability.

It is likely that after a few weeks, this site will have gained some internal PageRank. More traffic can be seen as a result. But outwardly, the Google Toolbar PageRank meter will still show the same old and low score. Then a snapshot will be taken, and the better score the site gets will reflect what is already been happening behind the scenes.

3.3. Other factors that affect ranking

3.3.1. Onsite factors (on the linked site)

3.3.1.1. Topic relevance: keywords, title tag.

PageRank has only ever been an approximation of the quality of a web page and has never had anything to do with the measuring of the topical relevance of a web page. Topical relevance is measured with link context and on-page factors such as keyword density, title tag and everything else. Google combines PageRank with sophisticated text-matching techniques to find pages that are both important and relevant to the user's search. Google examines all aspects of the content of a page (and the content of the pages linking to it) to determine if it is a good match for user's queries (Magazine 2007).

3.3.1.2. Page Specific Factors

Headers (h1, ..., h6), strong tags and semantic content are important but they do not improve the PageRank. As long as search engines will become more semantic in the coming years, the importance of the content is a precondition. Keeping the content readable and useful as well as being aware of the text surrounding the keywords is a key (Magazine 2007).

3.3.2. Offsite factors (on the linking site)

3.3.2.1. Anchor text

Google treats the text of links in a special way (Brin and Page April 1998). Instead of associating the text of a link with the page that the link is on, Google associates it with the page the links points to. This fact has several advantages. The first one is that anchors often provide more accurate descriptions of web pages than the pages themselves. They frequently state precisely what is

significant about the Web page and furthermore they are written by people other than the author of the Web page, so they are more resistant to malicious tampering to move a page to the top of the search results for commercial gain. The second one is that anchors may exist for documents which cannot be indexed by a text-based search engine, such as images, programmes and databases. Thus, it is possible to return web pages which have not actually been crawled. In addition, the engine can compare the search terms with a list of its backlink document titles. Accordingly, even though the text of the document itself may not match the search terms, if the document is cited by documents whose titles or backlink anchor texts match the search terms, the document will be considered a match. In addition to or instead of the anchor text, the text in the immediate vicinity of the backlink anchor text can also be compared to the search terms in order to improve the search. (Page 2001).

The idea of propagating anchor text to the page it refers to helps to search non-text information, and expands the search coverage with fewer downloaded documents. Using anchor propagation helps to provide better quality results (Brin and Page April 1998).

Search engines not only look for the text on the Web page itself when they determine what the page “is about”. They also look at the anchor text of the inbound link. That leads to content relevancy for the search engines. For competitive search queries this will not make much of a difference, but for sites focused on the long tail (i.e. trying to generate traffic from more obscure search queries) these links may be able to top the scales in their favour. An additional bonus is that such links also make it easier for the search engines to find and index the page. Moreover, they may pass along some PageRank or link juice to that particular page (Pandia 2007).

A good illustration of the importance of anchor text is the following. E.g. imagine the Nike website. They want to rank for the “word” shoes. Probably they will get hundreds of PR9 pages linking to the Nike’s webpage this way:

Nike

All these pages are going to send lots of PageRank to Nike's site. Nike will be seen as important. Google is going to look at the word in the link itself as a key signal to determine that Nike is an important place for Nike. Nike will rank for its name.

Now imagine that another shoe fabricant, less known, called Zappos, does not get links from all those PR9 sites. Instead they receive a mix of links from PR4, PR5 and PR6 sites. They all link to Zappos like this:

Zappos for Shoes

However the importance of this link is smaller, they do carry some weight. Moreover, the anchor text used are key words are adding information and therefore, value. They are pointing to the final site and saying that the word "shoes" in the links. That is going to help to rank better for the word "shoes", almost certainly much better than all those links Nike has (SearchEngineLand 2007).

4. Conclusion

As search engines evolve, so must the methods of the marketers. The main focus of search engines is providing a better understanding of the user intent. In this direction, rather than delivering billions of pages, search engines are working to deliver the one to three pages users are really looking for.

This Master Thesis has analyzed in depth state-of-the-art of search marketing, focusing afterwards on search engine optimisation and concretely on the optimisation of PageRank factors.

Because of the fact that the most influential factor in the purchase decision is the website, providers of information or services realize that being visible and having good quality is not an options but a requirement. The website is indeed a key factor for success. Driving as many potential customers as possible to a company website will increase the possibility of increasing the ROI. This effect relies heavily on the visibility of website and its ability to attract clickstreams.

In order to lead the traffic to a website search marketing offers two different techniques. search engine optimisation is based on organic search. This means optimizing the technical design of the website and enhancing the PageRank strategy. The second one is search engine marketing and the main difference is that this approach costs money and is based on keyword pay per click in its different versions.

The interface where both users and companies meet in the virtual world are the search engines. Each search engine applies a proprietary algorithm when delivering the search engine results page, SERP. The goal of this ranking is to order the results on a relevance basis.

Google is the most used search engine. Since a large number of customers use Google, brands need to be there as well. This is not an option, it is a necessity.

Google bases its ranking on several factors. This Thesis focuses on the core of Google: PageRank. The main characteristic of PageRank is that it makes use of the link structure of the web and assigns a “quality ranking” for each page crawled and indexed. Furthermore, PageRank makes use of the linking text - or anchor text - in order to improve search results and to serve as a quality filter.

The algorithm of Google, PageRank, was designed in 1998 and was aimed at improving the low quality matches of the search engines of that time. Moreover, it was designed as to overcome scalability problems in a virtual world which was by then starting to grow tremendously. At the same time, the system should handle the queries quickly in order to ameliorate the user experience. PageRank is based mainly on inbound links, outbound links and internal navigation. These concepts have been reviewed and discussed in Chapter 3.

PageRank is only one of the many factors that are taken into account when Google delivers the SERP. However, PageRank could be seen as the pivot of search engine optimisation. The complements of this primary but essential structure are the inbound and outbound links as well as the anchor text. All these factors together with other techniques determine the success of a search.

PageRank started as a university project and has evolved into one of the most powerful companies ever. The reason for this achievement is that the Google search engine guesses in a satisfactory level “the need behind the search”. This is the fundamental matter. During the composition of this Master Thesis many changes have happened in the online marketing field. And much more will take place. This is proof of how fast this field evolves and of the necessity of being constantly updated in order to be relevant to the final user, who at the end represents the goal.

Understanding the function of linking in the internet is a basic prerequisite in order to adapt the strategy of each company. As seen in the Chapter 3, different strategies to appear in the first results of a SERP are compared. When optimizing a website it is also important to take into account the intensity of the

keyword competition. Depending on the case, focusing on PageRank optimization will make more or less sense.

The future brings more techniques which try to fulfil in a higher level the requirements that hide behind each query. Vertical search is a new typology of search engines. Other trends that are developed and scarcely implemented at the moment but that will be widely adopted in a near future will make use of vertical search engines. The main areas where these specialized search engines focus on are health, video games, finance, travel, politics and automobiles. They will also combine content from several sources in the search results. Image search is the fastest growing type of vertical consumer search.

Another important field that was discussed in the Search Insider Summit Conference was about video search optimization. Online video is the emerging media best received by consumers with a ratio of 43%. If we take into account that in the U.S. more than 80% of the households have got broadband Internet access at the moment, publishers should provide as much quality streaming video content as they can to meet the growing demand of consumers and advertisers alike. One of the proofs that video is getting in the search business is the recent acquisition of YouTube by Google for $1.65 billion.

5. Bibliography

Batista, H. (2007). "Log based link analysis for improved PageRank." from http://hamletbatista.com/2007/05/29/mining-your-server-log-files.

Battelle, J. (2005). The Search. How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture.

Battelle, J. (April 2007). Search, Media, Technology. It's worth the attention (videoconference).

Bigmouthmedia (September 2006). Best Use of Search Marketing. New Media Age. **Supplement:** p25-25.

Brin, S. and L. Page (April 1998). "The anatomy of a large-scale hypertextual web search engine." Computer Science Department, Stanford University: 20.

Broder, A. (2001). "A taxonomy of a web search." IBM Research.

Brooks, G. (January 2006). Search Marketing. New Media Age. **Special Section:** p5-8.

comScoreNetworks (2007).

Croft, M. (April 2006). Optimising the search mix. Marketing Week. **Vol. 29:** p41-42.

DoubleClick (November 2006). DoubleClick Touchpoints IV: How Digital Media Fit into Consumer Purchase Decisions.

eMarketer (May 2007) "Social Network Marketing to Reach $2,5 Billion in 2011." **Volume**, DOI:

Google. (2007). "PageRank Explained ", from http://www.google.com/technology/.

HighRankings. (2007). "Search Engine Optimization." from http://www.highrankings.com/archives/issue070.htm.

Hof, R. (April 2007). Is Google Too Powerful? BusinessWeek Online. B. Online.

InteractiveAdvertisingBureau and PricewaterhouseCoopers (May 2007). IAB Internet Advertising Revenue Report.

iprcom (2007). "PageRank shows up in the SERP."

iProspect (April 2006). iProspect Search Engine User Behavior Study**:** 17.

iwebtool (2007). "Visual PageRank View."

Jansen, B. J. and M. Resnick (December 2006). "An examination of searcher's perceptions of nonsponsored and sponsored links during ecommerce Web searching." Journal of the American Society for Information Science & Technology **Vol.57**(Issue 14): p1949-1961.

Jones, G. (March 2007). Top 100 online advertisers. Marketing**:** p36-38.

Laycock, J. (February 2007). "Is Effective SEO Always Good SEO?" from http://www.searchengineguide.com/laycock/009489.html.

Magazine, S. (2007). "Google PageRank: What Do We Know About It?" from http://www.smashingmagazine.com/2007/06/05/google-pagerank-what-do-we-really-know-about-it/.

Netcraft (2008). "January 2008 Web Server Survey."

News.Com. (April 2007). "Google rises at Yahoo's expense." from http://news.com/Google+rises+at+Yahoo+expense/2100-1038_3-6178164.html.

O'Reilly, T. (September 2005). "What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software." from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Odden, L. (May 2007). "Gord Hotchkiss on Connection and Community." from http://www.toprankblog.com/2007/05/gord-hotchkiss-on-connection-and-community/.

Page, L. (2001). Method for node ranking in a linked database.

Page, L., S. Brin, et al. (1999). "The PageRank Citation Ranking:Bringing Order to the Web."

Pandia (2007). "Online tool gives a visual presentation of the PageRank of links."

Pandia. (2007). "The role of insite linking in search engine marketing." from http://www.pandia.com/sew/403-linking-2.html.

PronetAdvertising. (May 2007). "Fox Interactive Media Research: Social Networks Are A Good Advertising Platform." from http://www.pronetadvertising.com/articles/fox-interactive-media-research-social-networks-are-a-good-advertising-platform10019.html.

Ridings, C. (2002). "PageRank Uncovered."

Rogers, I. (2002). "The Google PageRank Algorithm and How It Works." from http://www.iprcom.com/papers/pagerank/index.html.

SearchEngineGuide. (2002). "Understanding and Building Google PageRank." from http://www.searchengineguide.com/orbidex/2002/0207_orb1.html.

SearchEngineLand. (2007). "What is Google PageRank? A Guide For Searchers and Webmasters." from http://www.searchengineland.com/070426-011828.php.

SearchEngineWatch. (March 2007). "U.S. Search Engine Rankings, February 2007." from http://searchenginewatch.com/showPage.html?page=3625336.

Sen, R. (2005). "Optimal Search Engine Marketing Strategy." International Journal of Electronic Commerce **Vol. 10**(Issue 1): p9-25.

Traffick. (2007). "Global Search Market Shares: AOL insignificant ", from http://www.traffick.com/2007/04/global-search-market-shares-aol.asp.

Trollinger, S. (November 2006). Free (Yes, Free!) Tools for BETTER SEO. Multichannel Merchant. **Vol. 2:** p45-46.

Vine, R. (February 2004). "The business of search engines : understanding how Web advertising, partnerships, and the race for market dominance affect search tools and search results - 1 - Cover Story."

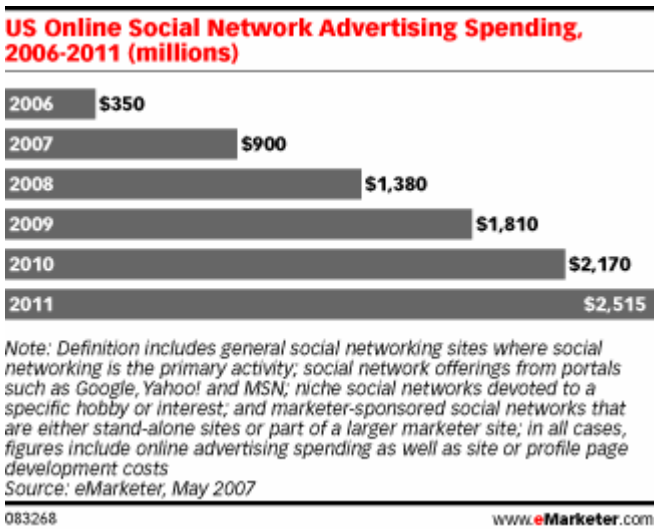
Wikipedia. (2007). "Semantic Web." from http://en.wikipedia.org/.

6. Annexes

6.1. Definitions of common terms in Search Marketing

6.1.1.General definitions

Community: The definition of "community" is changing. It's gone from what we have in common geographically to communities of ideology. The internet allows communities of interest having nothing to do with geography to exist and thrive. It's not physical proximity, it's mind space proximity. People move in and out of communities as they like. We need to think of search differently. Rather than a box we enter words into to get links back, search is a function that drives connections between communities. Things like personalization and social search are what's becoming important. Don't think of it as searching, think of it as connecting (Odden May 2007). One example of what is involve users and make them active within the community is the fact that on average, teenagers in America spends two hours per day on their Facebook site (Battelle April 2007). The impact of social network marketing is growing rapidly. Only in US, the ad spending on social networks will grow 180% from 2007 to 2011 (eMarketer May 2007).



Source: (eMarketer May 2007)

Figure 5-1: US Online Social Networking Ad Spending

Interestingly enough, the research also found that sports apparel manufacturer Adidas and video game developer Electronic Arts attributed more than 70% of their marketing return on investment to the "Momentum Effect", which is marketing-speak for how your brand propagates through a social network beyond the extent to which you advertise (i.e. word-of-mouth) (PronetAdvertising May 2007).

Mashups: A mashup is a website or application that combines content from more than one source into an integrated experience. It has become a common used term within the Web 2.0 environment.

PageRank (Google 2007): “PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at considerably more than the sheer volume of votes, or links a page receives; for example, it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important." Using these and other factors, Google provides its views on pages' relative importance. “

Semantic Web: The semantic web comprises a philosophy, a set of design principles, collaborative working groups, and a variety of enabling technologies. It derives from W3C director Tim Berners-Lee's vision of the Web as a universal medium for data, information, and knowledge exchange (Wikipedia 2007). As the vast corpus of information now available to us grows, the necessity of somehow tagging the information is becoming essential. Many in the search industry believe search will be revolutionized by what is called metadata. Clickstreams are a form of metadata –information about where you go and what you choose as you browse the Web. But to get to more perfect search, we need to create a more intelligent Web. That means tagging the relatively dumb Web pages that make up most of the Web as we know it today with some kind of code that declares, in a machine-readable universal lingo -called Resource Description Framework, RDF- what they are, what they are capable of doing, and how they might change over time. The semantic web is a vision of information that is understandable by computers, so that they can perform more

of the tedium involved in finding, sharing and combining information on the web. In a paper published by Berners-Lee, he explains the impact that the Semantic Web might have on search: *"If an engine of the future combines a reasoning engine with a search engine, it may be able to get the best of both worlds... It will be able to reach out to indexes which contain very complete lists of all occurrences of a given term, and then use logic to weed out all but those which can be of use in solving the given problem... I also expect a strong commercial incentive to develop engines and algorithms which will efficiently tackle specific types if problem... Though there will still not be a machine which can guarantee to answer arbitrary questions, the power to answer real question which are the stuff of our daily lives and especially of commerce may be quite remarkable"* (Battelle 2005). One of the pioneer companies that harness from the collective intelligence is Flickr. Users are able to categorize sites using freely chosen keywords -tags. Tagging allows for the kind of multiple, overlapping associations that the brain itself uses, rather than rigid categories. In the canonical example, a Flickr photo of a puppy might be tagged both "puppy" and "cute"--allowing for retrieval along natural axes generated user activity (O'Reilly September 2005).

Web 2.0: According to John Battelle, Search is driving business in the Web 2.0 world (Battelle April 2007). The term Web 2.0 emerged during a brainstorming session in O'Reilly Media in 2004 and it refers to a perceived second-generation of Web based communities and hosted services —such as social networking sites, wikis and folksonomies (tagged spaces)— that facilitate collaboration and sharing between users (Wikipedia 2007). In this new collaborative Web, the service automatically gets better the more people use it (O'Reilly September 2005). According to a report written by O'Reilly about Principles and Best Practices the reasons that have caused this change are a combination of some raw demographic and technological drivers –for example: one billion people around the world have access to the Internet, mobile devices outnumber desktop computes by a factor of two, nearly 50% of all U.S. Internet access is now via always-on broadband connections- with the fundamental laws of social networks and lessons from the Web's first decade –such as: in the first quarter

of 2006, MySpace.com signed up 280.000 new users each day and had the second most Internet traffic, by the second quarter of 2006, 50 million blogs were created (new ones were added at a rate of two per second), in 2005 eBay conducted 8 billion API-based web service transactions-. A set of best practices are standardized within the Internet Community. The first, release early and release often has been taken to the extreme by eBay, which deploys a new version of its service approximately every two weeks. Flickr photo-sharing service took this even further, deploying hundreds of incremental releases during an 18 month period from February 2004 through August 2005. Another principle is engaging users as co-developers and real-time testers. Amazon.com runs multiple feature tests on its live site every day. To instrument your product in order to see how users are using the product is also a must. What users do often tells you more than what they say. Some other best practices advice to incrementally create new products, make operations a core competency, use dynamic tools and languages (O'Reilly September 2005).

The Long Tail: Overture and Google's success came from an understanding of what Chris Anderson refers to as "the long tail," the collective power of the small sites that make up the bulk of the web's content. DoubleClick's offerings require a formal sales contract, limiting their market to the few thousand largest websites. Overture and Google figured out how to enable ad placement on virtually any web page. What's more, they eschewed publisher/ad-agency friendly advertising formats such as banner ads and popups in favour of minimally intrusive, context-sensitive, consumer-friendly text advertising (O'Reilly September 2005).

6.1.2. Search terms

SEM - SEARCH ENGINE MARKETING

Click-through rate (CTR): The percentage of those clicking on a link out of the total number who see the link.

Conversion rate: The percentage of visitors to a Web site whose actions are considered to be a “conversion”, such as a sale or a request to receive more information.

Cost per click (CPC): An advertiser pays an agreed amount for each click someone makes on a link leading to their Web site.

Landing page: The Web page that a visitor reaches after clicking a search engine listing.

Paid listings: Listings that search engines sell to advertisers.

Pay for performance: Synonym for pay per click, stressing to advertisers that they are only paying for ads that “perform” in terms of delivering traffic.

Pay per click: When advertisers pays a fee to the search engine for every user who clicks through on their ads.

SEO - SEARCH ENGINE OPTIMIZATION

Meta tag: Information placed in a Web page that is not intended for users to see but which passes information to search engine spiders.

Meta keyword tag: Allows page authors to add text to a page to help with the search engine ranking process. Not all search engines use the tag.

Organic listings: Listings that search engines do not sell. Sites appear solely because a search engine has deemed it editorially important for them to be included, regardless of payment.

Search engine optimization: Altering a Web site so that it does well in the organic, spider based listings of search engines.

Spider: A component of a search engine that gathers listings by automatically "crawling" the Web. A search engine's spider follows links to Web pages, making copies that it stores in the search engine's index according to a set of rules known as an algorithm.

6.2. Results from different studies

Several studies have been realized within the Internet community in order to know how do users behave when searching, which impact have the paid results contra the organic or natural results, which type of ads have a good reaction on the user and which ones are rejected by them. From this studies, interesting key factors can be extracted.

- iPropstect carried out one of this studies (iProspect April 2006). They find out some interesting results. Being on the first page of results, or at minimum within the first three pages of search results is vital for search engine marketers to ensure that their websites are found as long as 10% of the users click on results beyond the third page and 62% do not go further than the first page. Regarding this group, what they do later is to switch query and/or search engine. Another study reported that the conversion rate (i.e., customers who actually buy something) drops nearly 90% between the 1st and the 10th positions (Jansen and Resnick December 2006). There appears to be an intrinsic trust value associated with the rate of a listing. Another point is that search engine user confidence has increased over the last four years, as has the use of longer keyword searches. In this sense, 82% of search engine users report that they re-launch an unsuccessful query using the same search engine, but using more keywords. The same figure was just 68% in 2002. The implication for marketers is the need to better target these longer keyword phrases to be found by searchers. The last key finding is that search engine results continue to impart brand equity on those

companies that appear at the top of search results. 36% of search engine users report that they believe that companies whose websites appear at the top of the search results are leaders in their field. This belief has increased slightly over the last four years (33% in 2002). Brand marketers and search engine marketers should be heartened that one-third of users still believe that top search results equals top of field. So for this segment of the user population, top search rankings impart brand equity and create the perception of industry leadership.

- Jansen and Resnick (Jansen and Resnick December 2006) conducted an investigation into the effect of sponsored links on ecommerce information seeking on the Web. Because searching is a very task-oriented behaviour, it is essential to understand how sponsored listings fit into the tasks that searchers typically execute when using Web search engines. As a matter of fact, understanding the behaviour of online shoppers is a priority issue for competing in the virtual market place. The results of the study indicate that there is a strong preference for nonsponsored links, with searchers viewing these results first more than 82% of the time. This conclusion seems to collide with the fact that Paid Search is the prevalent business model for searching on the Web and businesses see them as the future of Web marketing. The key to whether paid search is a viable business model comes down to perceived relevance but certainly, for the near future, it appears to be the predominant revenue source for the Web search engines, although some commentators have questioned sponsored links as a long-term business model. However, it is important not to forget that sponsored links are primarily transactional. There are demographic factors also influencing ecommerce searching. For example, the percentage of searchers making a purchase online increased as a function of time spent online. That means, the longer the amount of time spent on the Web in a given episode, the greater the chance of making an online purchase. From the data extracted, the main results are the following:

- When using a Web search engine for ecommerce searching, searchers will examine organic results before sponsored results.
- When using a Web search engine for ecommerce searching, searchers will examine organic links as more relevant than sponsored links.
- It appears that the Summary (brief description under the link) can have a positive impact on judging a link as relevant, but the Title is the decisive factor used by searchers when determining a link as not relevant.
- If the ecommerce query is brand specific, the searcher will be more likely to view a sponsored link.
- The lower the rank (i.e., higher on the page) of an organic link, the more likely a searcher will view it and evaluate it as relevant.
- If sponsored links are to be a long-term business model for Web searching, the lack of trust and bias against these paid links must be overcome.

- Ravi Sen carried out a study about the Optimal Search Engine Marketing Strategy (Sen 2005). In this study he analyzes the fact that although people tend not to trust paid placements, 82% of the amount spent on SEM activities go to paid placement campaigns, Much less is invested in SEO. The question come then: if SEO can generate more traffic for the same keywords, why are companies spending more on paid placement? One reason is that it costs less to purchase paid placements for campaigns for thousands of keywords than to implement SEO programs for even a few hundred keywords. The main hurdle to implementing an effective SEO program is the fact that each search engine has its own requirements. In addition, SEO does not consistently result in high rankings and therefore leads to unpredictable traffic. This is because search engines tend to vary their ranking algorithms on “natural search listings”, and in response, SEO specialists have to “guess” and adapt to ever-changing strategies and tactics to either maintain their positions or

improve their rankings. In this investigation, Sen developed an analytic model that makes it possible to compare search engine marketing strategies in terms of their impact on the profitability of on-line sellers. Some of the results indicate that on-line sellers would be better off not investing in SEM strategies in the case of electronic markets characterized by buyers with high search intensity or by products that have relatively few relevant Web sites (e.g., vintage car sellers in and around certain area). However, if the buyer's intensity is low, the product has a large number of relevant sites (e.g., computers), or the products is of low value (e.g., books), the probability to be listed among the high-ranked results on the SERP and therefore become part of a buyer's consideration set would be relatively low. In such a scenario, on-line sellers would want to invest in SEM whenever the cost is not high.