

Tsvetan Vetskov

**THE INFLUENCE OF GOOGLE'S RANKING ALGORITHM ON
SEARCH ENGINE OPTIMIZATION (SEO)**

Website Design and Development through the Prism of Internet Marketing

THE INFLUENCE OF GOOGLE'S RANKING ALGORITHM ON SEARCH ENGINE OPTIMIZATION (SEO)

Website Design and Development through the Prism of Internet Marketing

TsvetanVetskov

Bachelor's thesis

Spring 2012

Degree Programme in Information Technology

Oulu University of Applied Sciences

PREFACE

The forthcoming Bachelor's thesis was done at Oulu University of Applied Sciences School of Engineering's Raahe Campus.

It is really important for me to give my special gratitude to the mentor and supervisor of my thesis – Mr. Leo Ilkko. He was extremely patient and gave me the necessary guidance and freedom needed to achieve my goal toward this paper. The advice I received from him turned out to be invaluable and I see him as a key contributor for this thesis.

Mrs. Kaija Posio is responsible for the language checking of my thesis and I would like to thank her for having the patience to do this large task. During my years in our school she managed to help me with any language related issues both in the fields of English and Finnish which is the foundation of any knowledge.

Another person that should not go unmentioned is Mr. Jarmo Karppelin as a thesis coordinator and the person who was responsible for reviewing and approving my thesis proposition and assigning me with such a reliable supervisor.

Last but not least, I would like to express my gratitude towards my friends and family who supported me through the whole process of creating this Bachelor's thesis and managed to cope with the fact that I was not always able to give them my full attention.

Raahe, Finland

April, 2012

TIIVISTELMÄ

Oulun seudun ammattikorkeakoulu

Koulutusohjelma: Tietotekniikka

Kirjoittaja: Tsvetan Vetskov

Opinnäytetyön otsikko: Influence of Google's ranking algorithm on Search Engine Optimization (SEO)

Ohjaaja: Leo Ilkko

Termi ja valmistumisvuosi: Kevät, 2012

Sivumäärä: 48

Tämä opinnäytetyö tarjoaa käytännöllisen, ajantasaisen tutkimuksen Internet-sivuston optimointiin pyrittäessä korkeampiin sijoituksiin hakukoneissa. Päähuomio kohdistuu Googlen ranking-algoritmiin, joka on käytännössä sanellut Internet markkinoinnin suuntaviivat muutaman vuoden ajan.

Tämä opinnäytetyö on viimeisimpien trendien mukainen katsaus Internet markkinointiin. Markkinointityössä on tärkeää ymmärtää nykyaikaisen Internet-sivuston suunnittelun ja kehittämisen merkitys ja vaikutukset yrityksen liiketoimintaan ja yhteiskuntaan yleensä ja mitä suuret yritykset ovat valmiita tekemään Googlen hakukoneen ykköspaikan eteen.

Opinnäytetyössä tarkastellaan web-sivustojen ranking-sijoitusten määräytymisen tekniikan perusteita ja myös tehtäväkohtaisia työkaluja sekä kuvataan ranking-algoritmeja ja niiden toimintaa. Lisäksi työssä esitellään semanttiset verkot, mitkä liittyvät oleellisesti nykyaikaiseen hakukoneoptimointiin.

Avainsanat: Hakukoneoptimointi, SEO, Google, Internet markkinointi.

ABSTRACT

Oulu University of Applied Sciences

Degree programme: Information Technology

Author: Tsvetan Vetskov

Title of Bachelor's thesis: The Influence of Google's Ranking Algorithm on Search Engine Optimization (SEO)

Supervisor: Leo Ilkko

Term and year of completion: Spring 2012

Number of pages: 48

This Bachelor's thesis aims to provide a practical up-to-date research and reference paper for a modern website optimization pointed at higher rankings in search engines. The main aspect falls on Google's ranking algorithm that has been dictating Internet marketing for few years now.

A large part of this thesis is rather informational and aligned with the latest trends and changes – something that has not been done in an Internet marketing paper for a long period of time. This is why it is important to understand the impact of modern website design and development for business and society in general and what large companies are ready to do for that sweet number one spot on Google.

Reviewing the fundamentals of website ranking and also more task-specific tools gives a good overall picture on ranking algorithms and how they function. A slight touch to Semantic web was unavoidable due to the evolution of technology and artificial intelligence.

Keywords: Search Engine Optimization, SEO, Google, Internet Marketing.

TABLE OF CONTENTS

1. INTRODUCTION	4
1.1 Aims of the study	5
1.2 The World Wide Web as a constantly changing environment	5
1.3 Structure of the paper	8
2. COMMERCIAL MARKETING AND THE INTERNET	9
2.1 Evolution of online marketing through the years	9
2.2 The concept of Internet based businesses	12
2.3 The Internet as a tool of modeling business strategies	13
3. SEARCH ENGINES AND RANKING ALGORITHMS	15
3.1 Introduction to search engines	15
3.2 Web crawlers	21
3.3 Indexing of content	23
3.4 Searching for results	25
4. GOOGLE'S RANKING ALGORITHM AND ITS INFLUENCE	29
4.1 Pre-Panda period	29
4.2 Panda Update	31
4.3 Semantic web and search	33
4.3.1 Current LSI functionality	34
4.3.2 Knowledge graph and the future	36
5. SEARCH ENGINE OPTIMIZATION AFTER PANDA	39
5.1 Introduction to SEO	39
5.2 On-site search engine optimization	40
5.2.1 Design patterns and user experience	40
5.2.2 Content impression	41
5.2.3 User metrics	42
6. CONCLUSIONS	44
REFERENCES	45

1 INTRODUCTION

The topic for this thesis is something that came to my mind really naturally. I have been working on search engine optimization and Internet marketing as a private contractor for more than 5 years now and it was important for me to choose a topic that will make sense in my future development and that will certainly contribute not only to myself but also to anyone who would like to achieve a certain level of expertise in this field.

The core concepts that are explained in this paper in their majority are known for quite some time but the fast-paced Internet universe changes their nature rather often. There are also a few things that certainly are new to the World Wide Web and have not been dissected in such a document before.

The audience that this Bachelor's thesis is aimed at is quite large. Anyone from an online developer to a PR manager can gain invaluable knowledge and know-how by reading this research paper. I can identify website designers and computer engineers as primary target. Those professionals should always strive to add knowledge and diversify their expertise and search engine optimization is valued by employees due to its critical business impact in modern days.

Aligning IT with business has been a hot topic for a few years and there has been a lot of work put in optimizing companies and their assets in order to provide an increased profit with a minimum setback. Some global improvements as ITILv3 (IT Infrastructure Library) are now a major part of any corporate-grade company that wants to have a continuous improvement and a result-driven policy. Having that in mind, it is fair to say that Search Engine Optimization should be considered as a major task as it drives the targeted traffic to a website which equals more business.

As paper publishing of Internet marketing related materials has been neglected, it is important to note that most of the references will be to online sources,

which can give up-to-date information rather than a source that is a few years old. This is important to have in mind because of the rapid speed and dynamic nature of the Internet as a whole. As part of this environment, search engines and ranking algorithms have had to change extremely fast and as far as web developers and SEO professionals are concerned this has always been an unending chase between them and large search engines like Google, Yahoo! and MSN.

1.1 Aims of the study

This study aims at providing an up-to-date competent knowledge base related to Search Engine Optimization and its relation to Google's ranking algorithm. Being an IT professional in a small or large company can sometimes require working in the field of online marketing and well structured, credible materials are scarce.

Ranking algorithms are the core of any search engine and nowadays there is more and more semantics involved. This strives to make user experience more natural and coherent. This however comes at a price that is the complexity of a website development in terms of optimization and ranking on major search engines.

That is why this Bachelor's thesis addresses those complications and many question marks that go hand by hand with modern website planning and creation.

1.2 The World Wide Web as a constantly changing environment

The World Wide Web (also referred to as the WWW) is the environment that our society gives a huge amount of attention and this is not without its gains. First of all it is important to say that the WWW is not the Internet. It could be described as a communication medium that operates on top of the Internet, doing the same as the Email technology does. Since its discovery in the late 1980's, the

World Wide Web exploded into the single most used piece of virtual technology in the world.

TABLE 1. Internet usage evolution from 1995 till present time
(Internet Growth Statistics)

DATE	NUMBER OF USERS	% WORLD POPULATION	INFORMATION SOURCE
December, 1995	16 millions	0.4 %	IDC
December, 1997	70 millions	1.7 %	IDC
December, 1999	248 millions	4.1 %	Nua Ltd.
December, 2001	361 millions	5.8 %	Internet World Stats
December, 2003	719 millions	11.1 %	Internet World Stats
December, 2005	1,018 millions	15.7 %	Internet World Stats
December, 2007	1,319 millions	20.0 %	Internet World Stats
December, 2008	1,574 millions	23.5 %	Internet World Stats
December, 2009	1,802 millions	26.6 %	Internet World Stats
September, 2010	1,971 millions	28.8 %	Internet World Stats
December, 2011	2,267 millions	32.7 %	Internet World Stats
March, 2012	2,280 millions	32.7 %	Internet World Stats

Since the dawn of the first graphic websites until modern days the speed with which the WWW usage is increasing is incomparable to nothing that humanity

has ever created before. The number of users has increased dramatically in the past 17 years, as it can be seen in Table 1.

All the background data is necessary in order to be able to understand the scale of growth that we are facing. Those figures can help us explain one of the most profound issues that humanity faces when analysing the World Wide Web – wrong assumptions. This technology has developed so fast that scientists and professionals could not cope with the rate and were misled into making wrong conclusions over and over again.

I will insulate my attention to Internet marketing in detail. At first companies were not eager to assign assets especially to their online-based parts of the business. This changed when Email marketing was discovered in the middle 90's of the past century. What happened then was that most professionals in the field a wrong assumption that Email marketing will be the thing to last.

During a few years, search engines were introduced to the masses and Email marketing was no longer that popular. This was caused by the same increasing number of users mentioned before. It was not possible to remember all places by heart and a more profound means to locate online resources was needed. The search engines gave birth to a search engine optimization as a process. The ranking was crucial and businesses all around the world began assigning larger budgets to online campaigns pointed at higher position in Google or Yahoo!

Once Search Engine Optimization was a legitimate profession everyone interested in Internet marketing knew that the next big thing is a matter of time as the usage of online resources became more and more valuable.

After the arrival of social networks and social media some companies rushed to assumptions and decided that SEO was no longer to be and social media marketing was the future. However, the future turned out to be a combination of both as search engines managed to cope with the pressure and produced

extremely elaborate artificial technologies that can rank websites by applying complex filters and implementing the semantic web more and more. Google is the pioneer in such technologies and this is not a small part of the reason why it is the most popular IT Company in the world.

Those examples and observations prove the dynamic nature of online marketing and why it is important to keep up-to-date with the latest trends in marketing approach, coding patterns and ranking algorithm compliance.

1.3 Structure of the paper

The paper is structured in a way that can be followed easily. Complicated and in-depth information is provided layer by layer which makes the comprehension of the topic rather effortless.

In the first chapter I have placed the introduction part, which gives an overall impression of the topic by providing some major cornerstones and a solid foundation to supplement the following statements and facts.

The second chapter is turning the reader's attention to Internet marketing as a business tool and follows its development through the past years. Going deeper into the third chapter, the paper covers search engines and their ranking algorithms. Having a solid knowledge about those technologies is a valuable skill that cannot be overlooked by any IT professional and it is a sure prerequisite for successful online business.

The fourth chapter explains in detail the functioning of the ranking algorithm of Google both before and after the Panda update.

The fifth chapter dives deeper into the world of the modern Search engine optimization by layering all major approaches to build a successful and well ranked website – from an onsite optimization and design patterns to a content structuring and important metrics.

2 COMMERCIAL MARKETING AND THE INTERNET

The phenomenon called Internet swept the entire world like a storm and nowadays it is a vital part of our day-to-day operations. That is why it is only natural to observe a constantly increasing interest of small and large companies to the development of online branches of their business. The past year proved that online businesses can be self-sustainable and extremely profitable at the same time. Good examples of that are companies like Ebay and Amazon that rely solely on Internet customers and are still the leading resellers of our time.

If we look at the Internet marketing as a part of a given business strategy it is easy to understand why it is preferred more than other conventional marketing approaches. It is cheaper to set up compared to offline methods and it can target vast masses of potential customers that are laser targeted to the specific niche that the business is placed in.

2.1 Evolution of online marketing through the years

As a relatively new technology the Internet was not expected to provide very significant financial results in terms of marketing budgets when it started spreading. At first users were not so many and companies were hesitant on spending money for advertisement and marketing on the Internet.

With the increase of users the mind-set of bigger companies started changing and a turning point can be identified in the year 1995. The budgets for online marketing purposes in the U.S were equal to \$0 in 1994 and a total of \$301 million was spent in the United States of America in the year 1996. This by itself is a monstrous increase that turned attention all around the world to the untapped potential of online advertisement. Internet surfers continued to grow and companies from all niches started looking for a way to drive those prospects to their business. The increase continued with unparalleled tempo

and in 1997 the online industry was worth nearly \$1billion. (Free Encyclopaedia of Ecommerce, 2011)

It is important to point out that those budgets were spent on traditional advertisement campaigns that cannot be classified as pure online marketing but rather as traditional advertisements placed on the World's new leading media medium.

The first company to launch an online optimized marketing campaign was Bristol-Myers Squibb Co. – a US based drug company. They developed and launched a strategy that was meant to build the brand awareness by promoting their product Excedrin during a Tax season. The idea was as brilliant as it was simple because they placed advertisements on major financial websites and promoted their product as a “Tax headache pill”. The results were fast and the company reported an increase of 30 000 customers to their online list in just a week. After their success, many large corporations like Microsoft and IBM began to implement what was already done and started gaining huge benefits from that. (Hathorn, R., The History of Internet Marketing)

In the year 2000 the dot.com bubble busted and many companies were forced to significantly tighten their budgets for online marketing compared to the previous 4 years which were characterized by vast amounts of money being spent without much thought on key indicators like ROI (Return over Investment). The original banner and pop-ups were no longer appealing and the 21st century provided a challenge to marketing professionals as the average Internet user became more and more demanding.

In the following years newsletters and Email marketing were used all over the world to approach and involve new customers to any online business. In their shadow search engines slowly started gathering popularity as the number of websites was increasing uncontrollably. People began searching information more often than they did in the past and the pure product search was falling behind.

This maturing of the average user was the driving force behind the evolution of search engines and Internet marketing as a whole. Companies understood that people are interested in relevant information rather than in flashy banners or sales pitches and that search engines were capable of providing that information. That is when Search Engine Optimization was born – websites trying to rank higher for related keywords in order to get more targeted potential customers.

In the following years search engines became a dominant way of finding and accessing online media and then logically turned into the main investment target for businesses that wanted to develop an online branch. The competition was getting tougher and tougher and ranking algorithms were getting more and more complex and sophisticated. The situation produced a certain set of techniques that could provide fast ranking results by using unorthodox and unethical techniques for optimization – the so called “Black Hat Optimization”. This issue was noticed by major search engines and Google was the first to implement prevention algorithms that looked for websites optimized with Black Hat techniques and removed them from their search results. This process is well known to IT professionals and is called “Sandboxing”.

The past few years have been more dynamic than ever – more and more creative ways of marketing have been implemented in the virtual space we know as Internet. Nowadays there are huge companies that offer services in the field of search engine optimization, online presence and brand management and social media marketing. The market is huge and there are more means to deliver a message to potential customers than ever.

This brief timeline review of Internet marketing was necessary for the paper’s core idea in order to give the reader a more scalable view of the development and tendencies that precede any major outbreak or innovation in the world of online marketing and search engine optimization.

2.2 The concept of Internet based businesses

Internet-based businesses are a totally different branch of online marketing as they are defined by companies solely counting on Internet customers and sales as their main income source.

In its nature this type of approach is different from offline businesses in terms of strategic planning, IT alignment and resource allocation. Some great retailers like Ebay and Amazon have their companies set up entirely online and this does not stop them from being the leaders in retail all around the world.

The biggest advantage of online based companies is the significantly lower start-up fee. Basically you can always register a domain name, have a website set-up and license a company. After that you are ready to go and the initial work can be done from the comfort of your own home. Those lucrative factors are often the driving force behind small and medium start-up businesses, which are smart on their initial investment and aware of the unlimited potential of online customers.

Private companies and businesses have the habit of using multiple marketing mediums to communicate information directly or indirectly to their potential customers and target audiences. We already discussed how and why the Internet and the World Wide Web became such a dynamic and constantly evolving information channel that takes both verbal and nonverbal communication to the next level.

Having this said we can now look at the major tool of business in their fight for customers – a website. The website is one of the most versatile and cost-effective ways of providing information to a certain target group of people. Businesses are nowadays well aware of the ways in which they can benefit from this fact. Utilizing their websites in different ways and for different purposes provides all-in-one marketing solutions that used to consume vast amounts of resources before the Internet era.

A well planned start-up online based business can never afford to have the website built without an initial planning. Sometimes this is just done as a tick in the checklist and neglecting it often costs dearly to the company. A lot of thought and planning are needed for a website to be placed on the correct layer of prospects while keeping informational and aligned with the idea of the company. It is always vital to consider the website as a major part of any online business strategy and this is a usual misconception when companies regard the role of a website in terms of marketing. (Pakroo, P., & Caputo, C., 2008. The small business start-up kit)

The main cornerstone of the concept of an online based business is the website and the information it provides. It is the medium that allows a company to create a picture of the business for the target audience and generate a revenue. However, this cannot be done without the initial planning that will enable the generation of measurable results and traceable goals. The website as well as the whole IT department should be aligned in order to help the business in achieving its goals, generating its revenue and customer base and what lies in the foundations of this concept.

2.3 The Internet as a tool of modelling business strategies

In the past years the Internet has turned into a whole new marketing field that allows businesses to interact with new and old customers. Purely methodically the tendencies have switched and nowadays marketing experts are no longer trying to make people sit in front of their TV or read a newspaper. Instead of that people are nurtured and provoked to search information themselves. A skilled marketing expert wants to have those people searching and reaching conclusions by themselves. At the same time those users are actually being “funnelled” from one channel of relevant data to another until they reach the destined webpage or online property. This way the whole process of buying

online is centred on the customer and the amount of time spent on a certain matter is purely his choice.

The business strategies of major companies all around the world are getting biased towards the Internet more often than earlier because the World Wide Web offers a new market that allows you to reach the whole world as an audience – something extremely difficult in offline marketing. It is fair to say that this new marketing environment is more limited by the boundaries of imagination than by the availability of resources. (Sterne, J. 1998. World Wide Web Marketing: Integrating the Web into Your Market-ing Strategy.)

Business strategies worldwide have been modelled by the Internet due to the increasing transparency in all operations, unrivalled speed, ROI (return over investment) and instant global reach. Information can be harvested faster than ever before thanks to the Internet and interaction between business and customer is becoming more intuitive with every passing day. The functionality that the technology provides allows all parties to communicate on both vertical and horizontal level thus aligning the business goals with measurable results.

As companies mature in online marketing their strategies become more sophisticated and information channels are formed for different groups of potential or current customers. Personalized messages are laser targeted to the needs of the destined audience and this increases efficiency exponentially. A strong side, which is added to any strategy indulged with the Internet is the diversity of means in which the company can interact, provide and collect information from the users any time of the day, any day of the year.

The World Wide Web offers more strategic value to businesses than any other medium and this is no small part thanks to the limitless range of methods and approaches that can manipulate information which on its terms is without doubt the driving force behind modern civilization.

3 SEARCH ENGINES AND RANKING ALGORITHMS

In the previous chapters we looked into the commercial side of online marketing and its impact on modern business strategies that companies implement. In this part analytical information is needed to make the reader to comprehend the following theoretical and practical implementations that lead to better results in ranking a website on search engines and driving more targeted visitors.

This chapter is dedicated to search engines and the driving force behind them – the ranking algorithms. Those technologies are incorporated in any major network on the World Wide Web and most of us are using them on day-to-day basis. Market leaders like Google, Yahoo! and Microsoft are well known for dedicating vast amounts of resources to develop their respective search engines. I will try to present a clear picture on their reasons for doing so and also, what exactly does this technology represent.

3.1 Introduction to search engines

By definition a search engine is designed to search information on the World Wide Web and the search results are generally presented in a list format. The viewed information may consist of web pages, photos, graphs or any other type of files or data. Search engines keep their data real-time by running complex ranking algorithms that go around websites and collect necessary data nonstop. (Web search engine, Wikipedia, 2012)

During the early years of the World Wide Web there were not so many websites and the need for a structured way of finding information was not so obvious. People remembered the name of the website they needed and just entered manually knowing that the necessary information should be there. With time the increasing number of users and websites naturally gave birth to “web directories” where websites could be listed by category. This was useful and

people were happy to visit those places and enter websites from there. As the Internet user evolved, the vast amount of webpages was impossible to remember or list in a single place. At that moment, means of searching through multiple sources of data simultaneously and displaying the interpreted results became necessary.

The pioneer in web searching was called Archie (from “archive” without the v) and was created by Alan Emtage, Bill Heelan and J. Peter Deutsch in 1990. The three students from McGill University in Montreal invented a simple yet brilliant software. The program worked by downloading the directory listings of all files located on public thus creating a searchable database of filenames. Indexing was not performed by Archie as the data at that time was limited and the manual search on the filenames was doable without taking too long. (“Internet History - Search Engines”, Universiteit Leiden, Netherlands, September 2001, web: LeidenU-Archie)

Archie can be classified as primitive compared to today’s standards as there was no content searching, semantic functionality or indexing incorporated but still we have to acknowledge the technological innovation that put a start to the future of web search. In the following years the web content evolved and so did search engines. Many were created just to see them disappear, crushed by the leaders on the market. This is the way natural selection works even in the world of technology and the way for the best to keep getting better and better.

The 90’s of the past century were marked by exponential evolution of the World Wide Web and search engines were making no exception. In the table below we can see a timeline with some cornerstones in the evolution of search engines. Figure 1 contains the launch dates of the most significant names and their current status.

Timeline					
Year	Engine	Current status	Year	Engine	Current status
1993	W3Catalog	Inactive	2005	AOL Search	Active
	Aliweb	Inactive		Ask.com	Active
1994	WebCrawler	Active, Aggregator	2006	wikiseek	Inactive
	Go.com	Active, Yahoo Search		Quaero	Active
	Lycos	Active		Ask.com	Active
1995	AltaVista	Inactive – redirected to Yahoo!	2007	Live Search	Active as Bing – new MSN
	SAPO	Active		ChaCha	Active
	Yahoo!	Active, Launched as a directory		wikiseek	Inactive
1996	Dogpile	Active, Aggregator	2008	Blackle.com	Active
	Inktomi	Acquired by Yahoo!		Powerset	Inactive (redirects to Bing)
	HotBot	Active (lycos.com)		LeapFish	Inactive
	Ask Jeeves	Active (ask.com, Jeeves went		Forestle	Inactive (redirects to Ecosia)
1997	Northern Light	Inactive	2009	VADLO	Active
	Yandex	Active		DuckDuckGo	Active, Aggregator
1998	Google	Active	2009	Bing	Active - rebranded Live Search
	MSN Search	Active as Bing		Mugurdy	Inactive – lack of funding
1999	AlltheWeb	Inactive - redirected to Yahoo!	2010	Goby	Active
	GenieKnows	Active, rebranded Yellowee.com		Black Google	Active
	Naver	Active		Yandex	Active, Global (English) search
2000	Baidu	Active	2011	Yummly	Active
	Exalead	Acquired by Dassault Systèmes		Interred	Active
2002	Inktomi	Acquired by Yahoo!	2012	Yandex	Active, Turkey search
2003	Info.com	Active		Volunia	Active , only Power User
2004	Yahoo! Search	Active, Launched web search			

FIGURE 1. Timeline of search engine launch dates and current status

The data viewed in the table below can lead us to a few conclusions, the most important of which is the diversity of strategies that market leaders are using to keep up to date with latest tendencies. Google is the leader without a doubt and we can see that they have not acquired nor merged with no single other search engine through the years. I find logic in that as they are investing large amounts of money on their own development and this brings confidence in their own capabilities.

Yahoo! bought many innovative search engines during the years, merging with their companies and more importantly with their technology. This can be a consequence of the fact that Yahoo! as a company are not so dedicated in developing its own search engine but rather its other products lines that are more or less corporate oriented. Those above are perfect examples of two very different approaches to online based businesses that should be in any related textbook for years to come.

Conventional search engines function by collecting and storing a certain type of information about many web sites. This information is retrieved from the code of the web page itself and the mechanism that carries out this task is called a Web crawler – an automated type of web browser that follows links both internally in the website and externally in the World Wide Web.

After the gathering of the data is completed for a certain page a process of analysis is initiated during which the engine determines how to index the content. Words can be extracted from titles, headings or meta tags or the body text of a page itself and the indexed data is stored into a database that will allow a fast and easy search at a later time by using queries or predefined search patterns.

The search engine cycle is finalized when a user enters a search query or a keyword and the engine generates the results based on the indexed content related to the search that was requested. Most search engines support Boolean operations that can help the user in defining as precise search query as possible.

An example in the picture below indicates how to use Boolean symbols to narrow down searches. As I use Google for this demonstration it is important to mention that the logical representation of a NOT operator is represented by a minus sign (-) before the word we want to exclude from the search.

Google potato pie

Search About 6,110,000 results (0.14 seconds)

Everything [Always With Butter: Potato Pie](#)
[alwayswithbutter.blogspot.com/2011/04/potato-pie.html](#) - Cached
 6 Apr 2011 – **Potato Pie**. Puff Pastry. A wonder all on its own. Ever since I started this blog and began making making things other than desserts, puff pastry ...

Images

Maps

Videos

News


Shopping

Recipes

More

Sweet Potato Pie I Recipe - Allrecipes.com
[allrecipes.com/recipe/sweet-potato-pie-ii](#) - Cached
 ★★★★★ 1314 reviews - 2 hrs 20 mins - 389 cal
 For this lovely **pie**, sweet **potatoes** are boiled, peeled and mashed together with butter, sugar, milk and eggs, then seasoned with nutmeg, cinnamon and vanilla.

Images for potato pie - Report images



Ingredients

	Yes	No
sweet potatoes	<input type="checkbox"/>	<input type="checkbox"/>
lard	<input type="checkbox"/>	<input type="checkbox"/>
bourbon	<input type="checkbox"/>	<input type="checkbox"/>
pie crust	<input type="checkbox"/>	<input type="checkbox"/>
steak	<input type="checkbox"/>	<input type="checkbox"/>
vanilla extract	<input type="checkbox"/>	<input type="checkbox"/>
cinnamon	<input type="checkbox"/>	<input type="checkbox"/>
nutmeg	<input type="checkbox"/>	<input type="checkbox"/>

Cooks.com - Recipes - Cheese Potato Pie
[www.cooks.com](#) Recipes - Cached
 Grate cheeses in food processor. Place ... Pour into cheese and **potato** mixture and blend thoroughly. Pour into a glass **pie** pan. Bake at 350 ... minutes and ...

FIGURE 2. Searching for “potato pie” on Google

In Figure 2 normal “potato pie” results are giving us vast options but we are not interested in “sweet potato pie” or “cheese potato pie” because we really want to try the famous potato pie with meat. By altering the search query we get more precise results and save time on browsing through the results’ page.

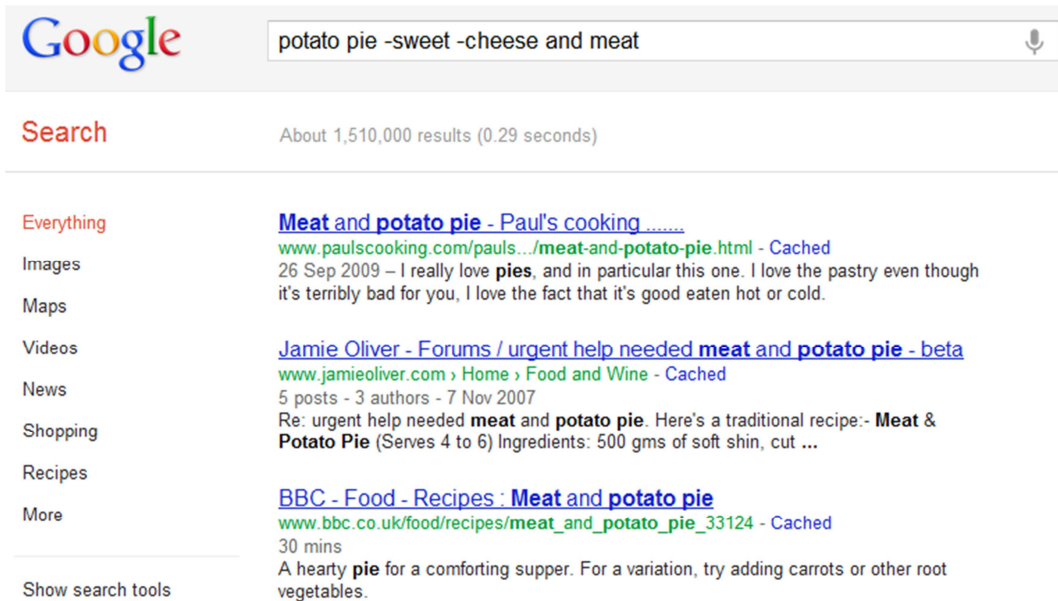


FIGURE 3. Searching for “potato pie” and BOOLEAN operator

This demonstration in Figure 3 showed how flexible and functional modern search engines are. But of course the usefulness of a search engine is defined by the relevance of the results that it gives back to search queries and that is why Google is the market leader with some pretty impressive results that can be seen in Table 2 below.

TABLE 2. Search engines market share (Net Marketshare - World)

Search engine ⇅	Market share in May 2011 ⇅	Market share in December 2010 ^[9] ⇅
Google	82.80%	84.65%
Yahoo!	6.42%	6.69%
Baidu	4.89%	3.39%
Bing	3.91%	3.29%
Yandex	1.7%	1.3%
Ask	0.52%	0.56%
AOL	0.3%	0.42%

As an absolute record holder in terms of a market share we have Google’s with its 86.3% in April 2010 (Net Market share - Google).

After those impressive numbers the next subchapters will try to explain the technical side of the way the search engines work by dissecting their 3 major features – Web crawler, Indexing and Searching.

3.2 Web crawlers

The technology that is responsible for travelling across the World Wide Web and jumping from page to page is called a web crawler. Every adequate engine is using one as the basic functionality and the idea behind web searches is not possible without it.

In its nature the web crawler is a software program that browses the Internet space in an automated manner, following a strictly defined methodical logic. Often referred to as bots, web spiders, web robots or even automatic indexers, those pieces of software are responsible for providing up-to-date information about websites all around the globe. Creating a copy of the contents is the main task that a web crawler needs to address. This copy will later be indexed by the search engine and provided in search results when related query is requested. Additional tasks that those bots usually handle include validating HTML or checking for invalid links.

The amount of websites currently residing on the World Wide Web is a factor that cannot be overlooked when we think about web crawlers and their resource intensive tasks. Jenny Edwards manages to define the issue in one simple sentence: "Given that the bandwidth for conducting crawls is neither infinite nor free, it is becoming essential to crawl the Web in not only a scalable, but efficient way, if some reasonable measure of quality or freshness is to be maintained." (Edwards, J., McCurley, K. S., and Tomlin, J. A., 2001)

It is vital for the efficiency of a web crawler to choose carefully which page to visit next and this strategic behaviour is defined by a combination of policies. The list of pages to be visited is elected by the selection policy, followed by the

re-visit policy that chooses when to check for changes on those pages. Overloading a web site could be an issue and the politeness policy makes sure to avoid such occurrence. The last major policy is the parallelization one and its purpose is to coordinate distributed Web crawlers. (Castillo, C., 2004. *Effective Web Crawling*)

It is fair to say that a web crawler should not only work commanded by a well-organized strategy but also have a highly optimized architecture that will enable efficient and resource caring process.

Building a slow crawler is fairly easy and downloading a few pages per second for a short period of time can be achieved without serious knowledge or resources. When we are talking about building a high-performance system that can process hundreds of millions of pages over few weeks we are presented with a number of challenges and difficulties in terms of system design, input and output interfaces, network efficiency, high availability and manageability. (Shkapenyuk, V. and Suel, T., 2002, 357-368)

Due to understandable reasons most major web crawlers have their algorithms kept as a corporate secret and it has always been a challenge for Search Engine Optimization professionals to find the key components to rank a website higher in the results. Google is the main goal for any online marketing expert as they hold the largest market share and unveiling their algorithm would certainly take search engine spamming to the next level.

Exactly those concerns are the reason why the algorithm behind Googlebot (the web crawler of Google Search Engine) has not been unveiled in the past years and the only existing documented reference is regarding the first versions that are now obsolete. (Brin, S. and Page, L., 1998., 30(1-7):107–117)

However, the corporate confidentiality addresses only specific modifications and functions and it is not hard to understand the basic workflow of a web crawler by looking into Figure 4 – a high-level architecture of a standard web crawler. It

can be described as a self-sustaining content scraper that has a logical scheduling pattern and storage capabilities.

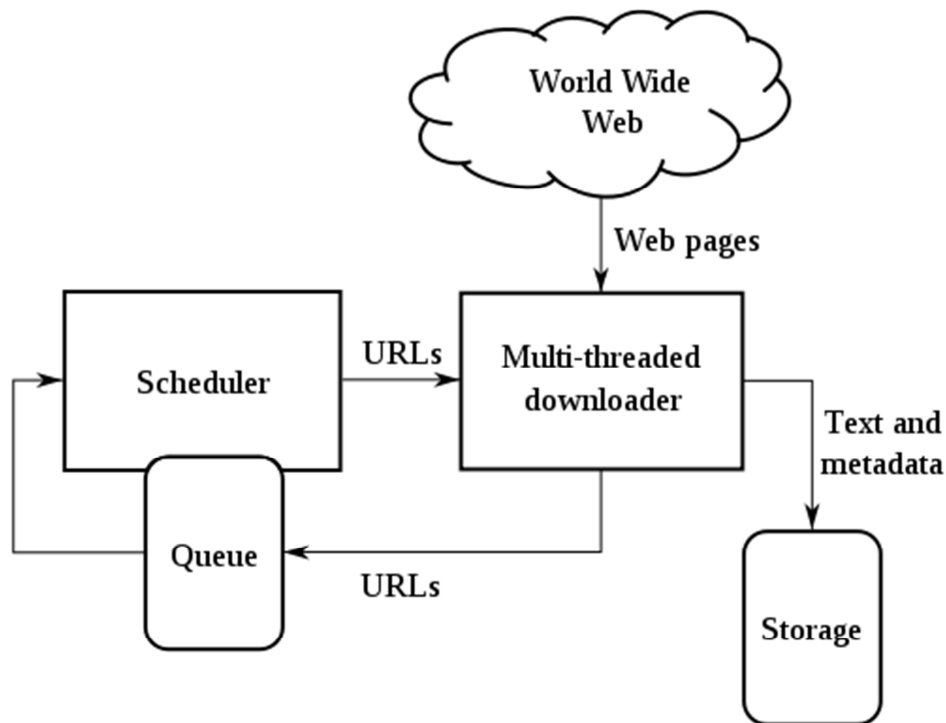


FIGURE 4. High-level architecture of a standard Web crawler

3.3 Indexing of content

After the web crawler of the search engine has stored the necessary information from a given amount of web pages, a process called indexing takes place. It collects, interprets and stores data in a desired format that will enable a fast and accurate retrieval once it is required. This process incorporates multiple approaches and methodologies from linguistics, mathematics, physics, informatics and even cognitive psychology and it is also referred to as web indexing.

The concept behind storing and indexing web page data is to allow an optimized speed and precision when searching and finding relevant information defined by a certain search query. Without the indexing in place, a search

engine would need to check every single document that has been stored by the web crawler each and every time a search is submitted which is resource-intensive and daunting task. Of course, there is a need of additional virtual storage for the index but this expense is compensated by the processor time saved while retrieving the indexed information.

Major cornerstones to be considered in engineering a search engine's architecture are the merge factors. They define how exactly data enters the index or more precisely – how words or object features are added to the index during asynchronous work of the indexers. Checking if the content is being updated or a new one is added is crucial to the merging process and those factors directly correlate to the data collection policy referred to earlier while reviewing the Web crawler. (Brown, E.W, 95-81, 1995)

The broad topic of search engine indexing has many aspects both theoretical and purely mathematical. However for the sake of this thesis paper I would like to pay attention to the so called “Meta tag indexing”. This type of information indexing in search engines relates to the meta tags of a webpage. Some of those are showed on the Figure 5 and used in almost every webpage that exists nowadays.

```
<title>BBC - Food - Recipes : Meat and potato pie</title>  
<meta name="description" content="A comforting homemade beef and potato pie, perfect for a family meal." />  
<meta name="keywords" content="bbc, food, recipes, Meat and potato pie" />
```

FIGURE 5. Title, meta description and meta keywords tag in HTML format

Defining the meta title, keywords and description used to be the most important task for any Search Engine Optimization expert in the past as the architecture of Google's search engine indexer placed significant weight on this information when ranking websites in relevance with the requested search query. Due to the evolution of computer technology and the unethical practices that were incorporated by many websites this is no longer the case and the meta tags lost their major part in ranking a site high in the search results.

Earlier search engines were able to index only the keywords in the meta tags as the full document could not be parsed due to the lack of support from the technology at that time. Initially it is fair to say that the initial design of the HTML mark-up language included meta tags just so that the page could be properly and effortlessly indexed without the need of data analysis. (Berners-Lee, T., RFC 1866, 1995)

3.4 Searching for results

A query in a web search is a string of data that a user enters into a search engine to fulfil a certain information demand. Most often those queries are a plain text or some type of hypertext with optional search parameters like Boolean operators for example.

According to Christopher Manning (Introduction to Information Retrieval, 2007, Chapter 19) we can define four categories that cover the context of almost all search queries. The first type is informational queries that cover a broad topic that can lead to thousands of relevant results. Figure 6 is a good example of a simple search for a city or an everyday appliance.



FIGURE 6. Demonstrating informational queries

The second type of search queries is named navigational queries due to the nature of the results that they produce. Those queries are usually used when

the user would like to navigate to a single website or a webpage. A suitable example here would be looking for a train ticket website as in Figure 7.

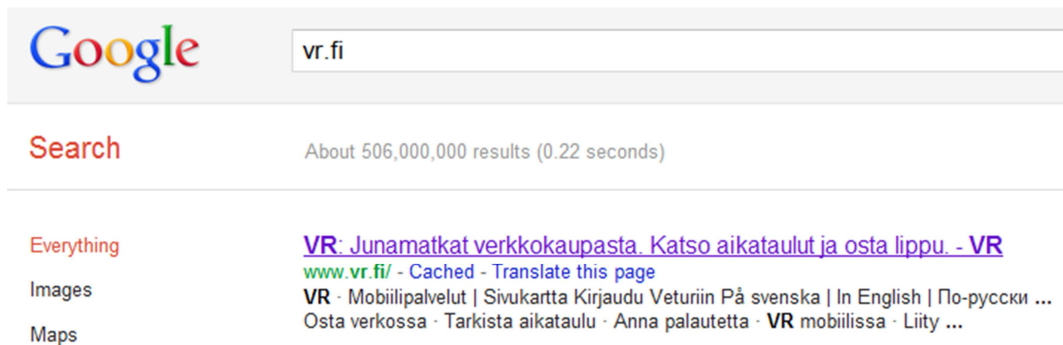


FIGURE 7. Demonstrating navigational queries

The last two types of queries are the ones that are most interesting and valuable for search engine optimization specialists. First there is a transactional query that is aimed at commercial actions like purchasing a new notebook or online music. The results generated by such queries are being targeted by online businesses in order to get potential customers visit their website as seen on Figure 8.

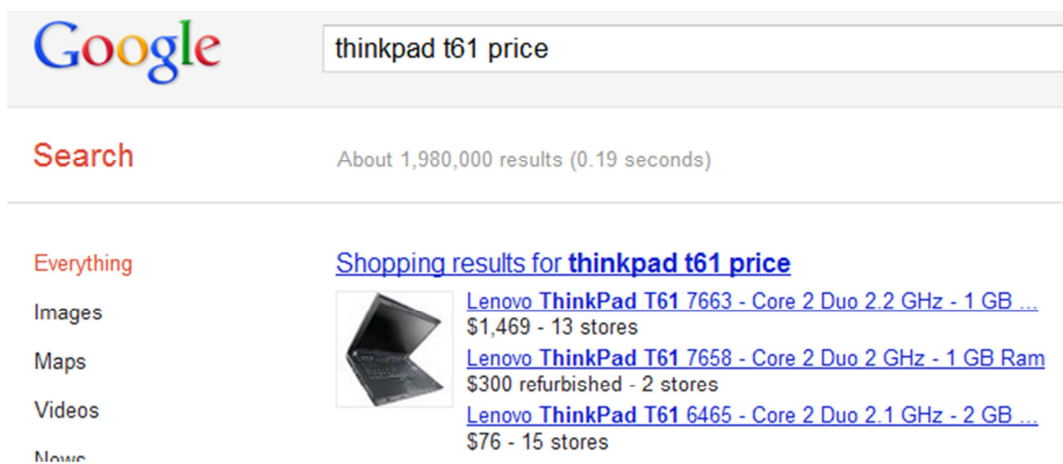


FIGURE 8. Demonstrating transactional queries

Connectivity queries are the last type that closes the group. Those queries are used to gather information on the connectivity and interaction between websites. Skilled webmasters should always use searches of this type to analyse competition and find potential weaknesses in the linking strategy of their own website. Connectivity queries can show how many pages are indexed from a given domain name or how many external links are leading to a given webpage – both extremely important factors in a successful online marketing campaign.

Using the command “link:” followed by a certain domain name, Google’s search engine query will provide all external links that are recognized by the web crawler and pointing to the requested domain as seen on Figure 9.

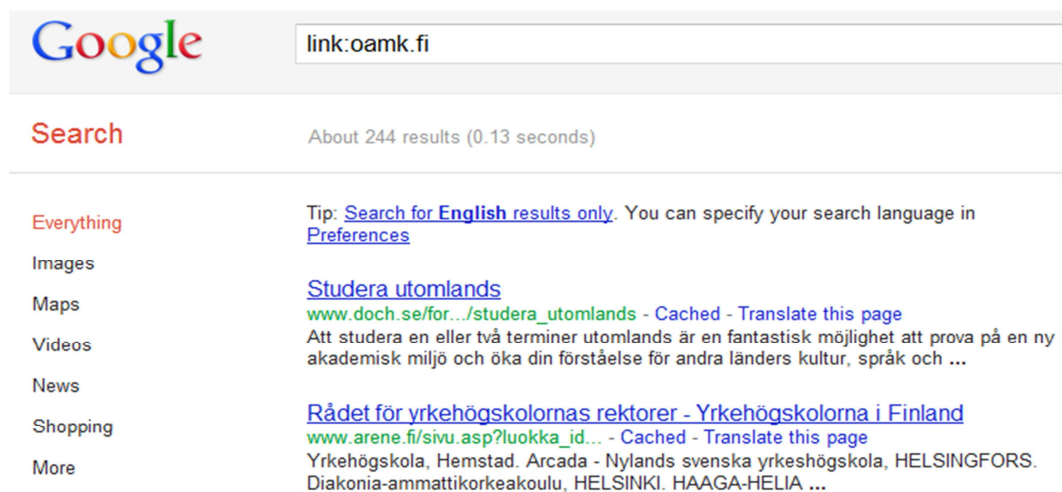


FIGURE 9. Search results for connectivity query with “link:” command

Another useful command in Google’s search engine demonstrated in Figure 10 is the “site:” prefix that should be followed by a domain name. This query will display all indexed pages of the given domain name which can enable optimizers and webmasters to locate not indexed pages and narrow down their search for weakness in the onsite preparedness of the website.

Google

site:oamk.fi

Search About 175,000 results (0.07 seconds)

Everything

Images

Maps

Videos

News

Shopping

More

Show search tools

Tip: [Search for English results only](#). You can specify your search language in [Preferences](#)

[Try Google Webmaster Tools](#) Google promotion

www.google.com/webmasters/ Do you own **oamk.fi**? Get indexing and ranking data from Google.

[Oamk » Pääsivu](#)
[www.oamk.fi/](#) - Cached - Translate this page
Oulun seudun ammattikorkeakoulu - Oulu University of Applied Sciences: Pääsivu.

[Oamk » Sivukartta](#)
www.oamk.fi/sivukartta/ - Cached - Translate this page
Oulun seudun ammattikorkeakoulu - Oulu University of Applied Sciences: Sivukartta.

FIGURE 10. Search results of a connectivity query with “site:” command

Most modern search engines support Boolean operators implemented in search queries and this can help the user in finding information related to a specific topic faster. Defining a search by using multiple descriptive words regarding the wanted topic can be implemented by using the “AND” operator between keywords. A faceted query is a conjunction of such facets; e.g. a query such as *(electronic OR computerized OR DRE) AND (voting OR elections OR election OR balloting OR electoral)* is likely to find documents about electronic voting even if they omit one of the words "electronic" and "voting", or even both. (Mihajlovic V., Hiemstra D., Blok H.E., Apers P., 2006)

4 GOOGLE'S RANKING ALGORITHM AND ITS INFLUENCE

As an undisputed market leader Google's search engine is naturally receiving the largest amount of attention from online marketers as it offers the largest possibilities for profit due to its huge user base. This is why the ranking algorithm behind the search engine has always been under a constant analysis by IT professionals worldwide in an attempt to find the best way to satisfy it and get ranked higher.

There are two significant periods that should be reviewed as both of them had a global impact on website planning and development – pre-Panda and post-Panda update. I will try to give a good overall picture on both in the following pages.

4.1 Pre-Panda period

Before the Panda update was introduced to Google's ranking algorithm in the beginning of 2011, there was a period of about 10 years when the ranking of websites was more or less done by a certain set of factors. During the years the weight of each major factor changed but in general all SEO professionals were used to do the same basic set of operations that would give positive results.

In Figure 11 we can see the evolution of Google's ranking algorithm defined by the importance level of the major ranking factors in the past.



FIGURE 11. Importance of key ranking factors on Google's ranking algorithm

The chart above is based on the statistical data collected through the years in attempt to find out the magic formula of ranking a website high in the results. Going along the key factors is vital in understanding how a search engine optimization was done in the past because the next chapter will cover the optimization of the new era without referring to any morally out-dated practices.

Domain trust has always been a major cornerstone in optimizing a website. This factor is mostly defined by the amount and quality of links that a website is getting from external domains. If a website is being linked majorly by low quality sources, it will be considered as a low quality website, too. This was a smart way to implement a self-sustaining control module based on the assumption that respected websites link only to other quality resources and vice versa.

The Florida update on Google's ranking algorithm that took place in the late 2003 had a major part in placing a weight on domain authority and trust, giving this metric a dominant role in the successful ranking on Google's search results. (Google's "Florida" Update, 2003)

An anchor text displayed on external links used to be a powerhouse between 2004 and the Panda update in 2011. This text carried out a great value on the targeted keywords by bringing the external links to a given domain together with the perspective of a relevant content. At some point unethical practices corrupted the whole Internet by spamming millions of links with keywords as an anchor text in the whole World Wide Web and this forced the algorithm developers to minimize the weight of this factor.

The third key factor is an on-page keyword usage. This is probably the most controversial item on the list due to the massive immoral practice of keyword stuffing which is basically putting tens or hundreds of related keywords inside the body of a webpage thus luring older algorithms to rank this page higher for multiple search queries. This practice was common in the early years of this century but succumbed into oblivion after keyword density was introduced. This basically made stuffing impossible as search crawlers became able to locate an excessive usage of a certain word on a page and penalized the website for that. The on-page keyword usage always had a moderate influence on the ranking before the introduction of Panda update in 2011.

A link juice given by the PageRank was a concept that dictated ranking for a short period of time. The big flaw of this factor that led to its decommissioning is the fact that with the evolution of the World Wide Web the users got more access to websites that are not their own and acquiring a link from a trusted website was no longer a hard task. This is why, if we follow the timeline of development of Web 2.0 and at the same time check the influence of the link juice on ranking we can notice they go in an opposite direction.

4.2 Panda Update

In February 2011 Google announced a major update for their ranking algorithm called Panda. This event turned around the world of Search Engine Optimization upside down as the concept behind the new algorithm was

fundamentally different from everything that was present in the past 10 years. The reasons that provoked this major release have been known for years as Google's strategy is strictly aligned with user experience and satisfaction.

Navneet Panda, a Google engineer, was the person who provided the computing capability for the new algorithm to work. He managed to find a way to utilize machine learning algorithms by scaling them up thus enabling artificial intelligence to work at the necessary rate.

Once the capability was present Google engaged their quality raters and asked them to identify a certain amount of websites that are likeable and enjoyable for the user. This process was all about user experience and how the website allowed people to interact with it.

The same process was repeated for websites that are not likeable. A list was gathered and same aspects were covered when the reviewing took place. After both lists were ready Google only had to look at the differences. Monitoring hundreds of metrics that separated those two groups of websites created tendencies. Using the scaled computing power, Google was able to apply that artificial logic to a global scale affecting to both already indexed and fresh web pages.

The Panda algorithm acts as a filter that uses those tendencies to mark websites either as "good" or "bad". Ranking higher the good ones and lower the bad ones was the next logical step in improving the user satisfaction, and this was the turning point in the world of Search Engine Optimization – the beginning of the new era. In its nature the Panda filter is designated to identify what Google believes to be "low quality pages". (Google Panda Update, Browsing Media, 2011)

The algorithm by itself is almost self-sustained as Google has a vast amount of user specific data collected from their pages, browser and other tools they provide. This data is easily channelled and processed by the artificial

intelligence of the ranking algorithm so that it can keep its database of good qualities and bad qualities up-to-date at all times. Thanks to that Google is able to release an official self-produced update of the Panda algorithm almost every 30 days. (Fishkin, R., 2012)

The arrival of Panda algorithm certainly changed things in the field of online marketing and it is fair to say that nowadays search engine optimizers are more like web strategists. It is no longer vital to perform daily tasks in the sake of generating a unique content and fresh link as nowadays ranking is based on a more philosophical concept, which is based upon continuous improvement of the user experience.

4.3 Semantic web and search

Semantic Web (often referred to as Web 3.0) is defined as a collaboration movement that is created and led by the W3C (World Wide Web Consortium) which is following the idea of promoting common formats for data. The semantic content in web pages encouraged by the Semantic Web is transforming the unstructured content on the World Wide Web into a logically connected “web of data”. (“W3C Semantic Web Activity”. World Wide Web Consortium (W3C)., 2011.)

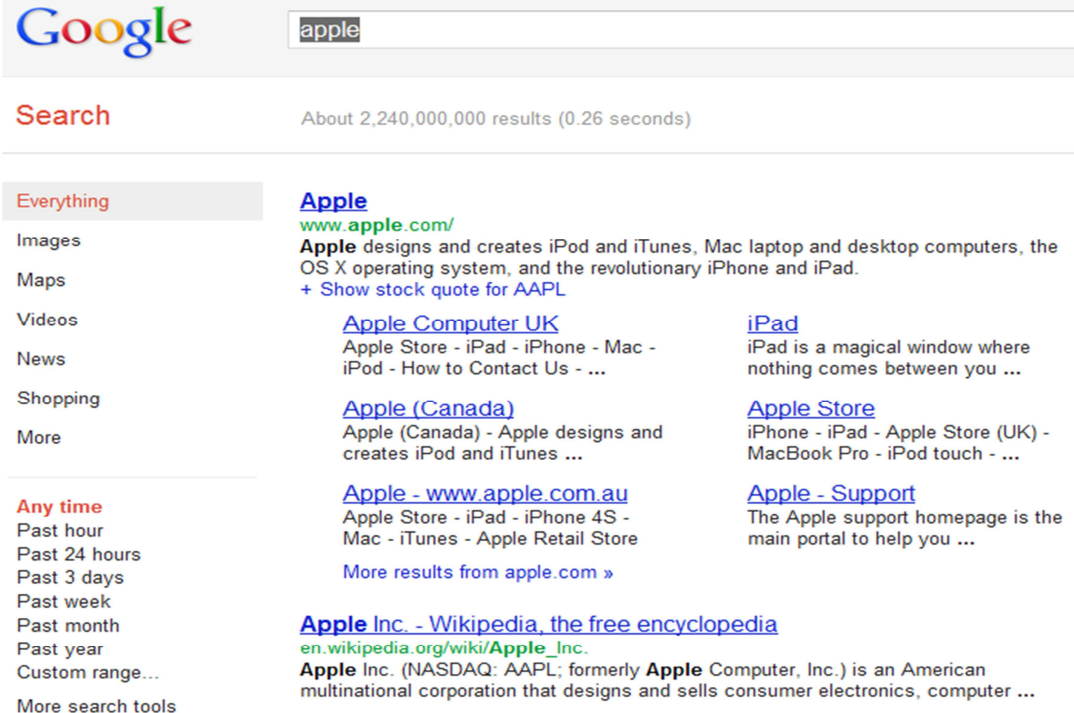
Web 3.0 is all about engaging an innovation technology to interpret, reuse and repurpose information in different situations thus significantly increasing the capabilities of the artificial intelligence used in the data processing.

The flagman of the movement Tim Berners-Lee described his initial vision of the Semantic Web as a virtual web in which computers become capable of analysing all the data on the Web – the content, links, and transactions between people and computers. When a ‘Semantic Web’ like this is present, the day-to-day mechanisms of trade, bureaucracy and our daily lives are to be handled by machines talking to machines. The ‘intelligent agents’ people have touted for ages are going to finally materialize. (Berners-Lee, T., Fischetti, M., 1999.)

4.3.1 Current LSI functionality

Latent Semantic Indexing (LSI) aims at discovering words and phrases that are contextually related to a single document or a group of such. Basically it is the process of discovering the related terms and phrases executed with the help of a mathematical equation that arranges words into matrixes which are analyzed until semantically connected terms are identified.

High-end search engines like Google have an LSI implemented as part of their indexing process and ranking position selection. This way the engine can determine the value of a single page or a whole website related to a searched keyword in a context as close as possible to the one meant by the user that submitted the query. For example searching for “apple” as in Figure 11 will bring the results related to computing and IT industry because statistically the majority of documents on the World Wide Web relate the word “apple” with terms like “computer” and “technology”.



The screenshot shows a Google search interface. At the top, the Google logo is on the left, and a search bar contains the word "apple". Below the search bar, it says "Search" and "About 2,240,000,000 results (0.26 seconds)". On the left side, there are navigation tabs: "Everything" (selected), "Images", "Maps", "Videos", "News", "Shopping", and "More". Below these are time filters: "Any time", "Past hour", "Past 24 hours", "Past 3 days", "Past week", "Past month", "Past year", and "Custom range...". At the bottom left, there is a link for "More search tools". The main search results are for "Apple". The first result is "Apple" with the URL "www.apple.com/". The description says "Apple designs and creates iPod and iTunes, Mac laptop and desktop computers, the OS X operating system, and the revolutionary iPhone and iPad." Below this is a link "+ Show stock quote for AAPL". There are three columns of related links: "Apple Computer UK" (Apple Store - iPad - iPhone - Mac - iPod - How to Contact Us - ...), "iPad" (iPad is a magical window where nothing comes between you ...), "Apple (Canada)" (Apple (Canada) - Apple designs and creates iPod and iTunes ...), "Apple Store" (iPhone - iPad - Apple Store (UK) - MacBook Pro - iPod touch - ...), "Apple - www.apple.com.au" (Apple Store - iPad - iPhone 4S - Mac - iTunes - Apple Retail Store), and "Apple - Support" (The Apple support homepage is the main portal to help you ...). At the bottom, there is a link "Apple Inc. - Wikipedia, the free encyclopedia" with the URL "en.wikipedia.org/wiki/Apple_Inc.". The description for this link says "Apple Inc. (NASDAQ: AAPL; formerly Apple Computer, Inc.) is an American multinational corporation that designs and sells consumer electronics, computer ...".

FIGURE 11. Demonstrating LSI functionality with a search query for “apple”

If the search query is changed in a different direction, the Latent Semantic Search makes sure to correct the main selection criteria and display contextually related information. In the next example the search query is changed to “apple pie” and the provided results are strictly related to it. Even though there are 100 times more results for just “apple”, the engine understands that what is now required is far from the IT niche and displays contextually adequate results as in Figure 12.

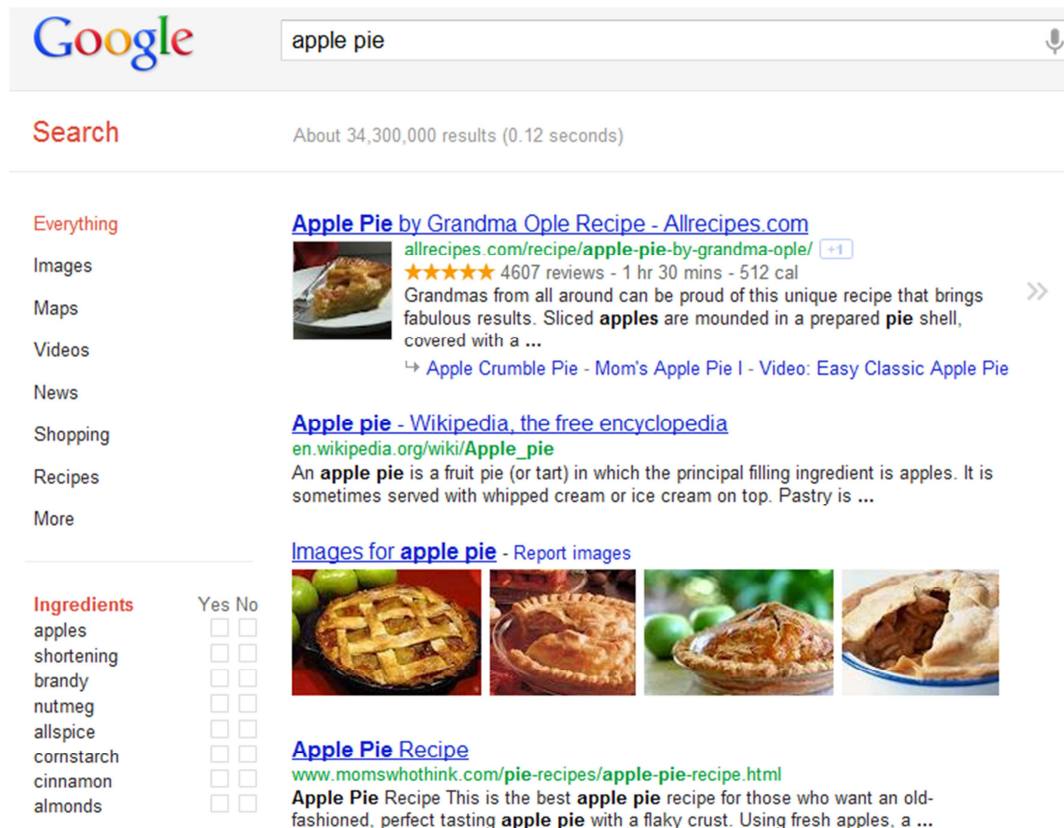


FIGURE 12. Demonstrating LSI with a search query for “apple pie”

Following the situation from the previous figure we can now look at the most efficient way of optimizing one website for LSI ranking – thematic silos. This term governs a specific topology of organizing a website’s internal pages. The

best practice is to create a top level page for the main keyword of the business (in this scenario it is “apple pie”) and then create additional pages under this one revolving around contextually related topics. A silo for our practical example should look like the Figure 13.

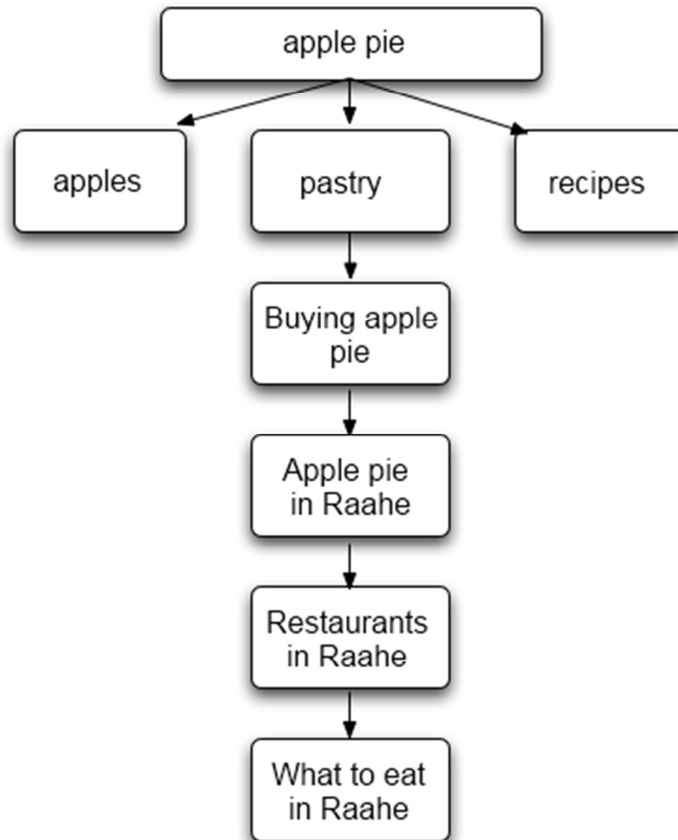


FIGURE 13. Example of theme silo for a website based on apple pies

By following a strict theme silo a website is increasing the chances that the LSI algorithm will give higher ranking for long-tail keywords consisting of semantically related phrases with lower traffic but with a higher accuracy and less competition.

4.3.2 Knowledge graph and the future

Currently in its nature, Google still does not understand what the user entirely means when a search query is submitted. The results given will be extremely accurate and contextually adequate thanks to the advanced ranking algorithm and the LSI functionality but still the search engine itself is unable to interpret completely the meaning of the requested information.

Google Inc. now wants to bring a change to their search results by remaking the way the information is presented. Instead of just words appearing as results, we are slowly starting to see whole data-sets presented in search results by different means that include related information. This correlation between attributes is natural for the human brain but in technology it is an extremely complex process identified by the term Artificial Intelligence.

This daunting task is being worked on as those words are being typed and the Senior Vice President of Google's research and development branch already indicated that they are "building a huge, in-house understanding of what an entity is and a repository of what entities are in the world and what should you know about those entities"

The base layer of this project was acquired when Google purchased a community-created knowledge base with around 12 million contextual entities by the name Freebase in 2010. Followed by a significant investment in resources, the database increased to almost 200 million entries as at the current moment. (Singhal, A, 2012)

The innovative transition from a word-based indexing to the knowledge graph developed by Google is a shift of direction that will increase the complexity of algorithms exponentially and require an immense computing power. Those reasons define the slow pace at which the whole project is being implemented. Currently the knowledge graph is used at the minimum possible capability but we are able to see changes in the way search results interact with users.

A practical example I am going to present is searching for the great painter Vincent van Gogh. As seen on Figure 14 below such search will generate the conventional results related to the search. The innovative idea of knowledge graph is observed at the bottom of the results page where a section called “Artwork searches for Vincent van Gogh” is displaying the most popular paintings of the artist.

[News for van gogh](#)

 [Van Gogh museum unveils new watercolour](#)
[Herald Sun](#) - 2 days ago
A YOUNG Vincent **van Gogh** was so struck by a dead willow leaning "lonely and melancholy" over a pond near The Hague that he knew at ...
Kansas City ...

[Vincent van Gogh | artist | 1853 - 1890 | The National Gallery, London](#)
www.nationalgallery.org.uk/artists/vincent-van-gogh
Explore information about the artist: Vincent **van Gogh**. See list of paintings at the National Gallery, London.

Artwork searches for **Vincent van Gogh**

				
The Starry Night	Self-Port...	Irises	The Potato Eaters	The Raising of Lazarus (after...

FIGURE 14. Demonstrating Google’s knowledge graph functionality

This might not seem as much but it indicates that the search engine now acknowledges van Gogh as a painter and relates him to his work using the knowledge graph. Functionality like that can be seen when searching for popular artists like The Beatles where a list of popular album is displayed again thanks to the new functionality that Google is slowly presenting into its search engine.

5 SEARCH ENGINE OPTIMIZATION AFTER PANDA

As I mentioned earlier nowadays it seems like the term Search Engine Optimization expert should be renamed to a Web Strategist. This feels necessary due to the fundamental changes that Panda update brought to Google's ranking algorithm.

It is fair to say that anything that is done on a website currently can impact significantly the ranking on Google, even the smallest of details. Due to those conditions I would like to emphasize the philosophical aspect of optimizing a website rather than the pure technical work that needs to be done. This topic is rarely covered in any research paper and some major misconceptions appeared in the general public due to the lack of quality summarization.

5.1 Introduction to SEO

Search Engine Optimization can be defined as the logical process aimed at increasing the web visibility and brand recognition for a particular web site or online based resource. The SEO process tries to drive as much natural and targeted traffic to a web site as possible while being aligned with business objectives and resource capabilities.

The largest part of this process goes to the effort put in ranking the particular web site higher in search engine results for desired keywords and this has been around since the dawn of the Internet searches.

The technology evolution affected the overall approach of optimizers worldwide significantly. As mentioned previously, the Panda update to Google's ranking algorithm was a turning point that made people doubt everything that was considered as true until then. It is no longer vital to put countless hours in gaining external links or writing grammatically correct content just to rank for

more relevant keywords. The search engines of 2012 are smarter and more adequate than ever and this calls for equivalent effort from webmasters and SEOs in the battle for higher positions. Time-consuming and repetitive tasks are no longer major factors in SEO but the overall time for a proper optimization has not become shorter because the effort put in research and planning is now increased and fully compensates for them.

As many people consider SEO a 2-part process I would like to make my case in the opposite. The off-site SEO is something that does not directly affect the ranking of a given website but rather increases its visibility which indirectly might lead to a higher traffic. This however should be placed in the publishing niche opposed to the more technical SEO process.

5.2 On-site search engine optimization

On-site search engine optimization defines the actions taken towards the code and content of a website that result in better ranking on search engine results. This chapter will concentrate on the best practices that should be considered and implemented after the Panda algorithm update has been applied.

It is not uncommon to notice that the sites with better external links, a more unique content and a clearer code are now ranking lower than websites that fall behind in those categories but are evaluated as more appealing by quality rating personnel. Simply, the signals that are designated to be more user-oriented are less obvious in those types of sites and after Panda update this is the big thing.

5.2.1 Design patterns and user experience

In the past the design and user experience were always valued to bring more links and visibility but after the Panda algorithm those factors are certainly bearing primary impact directly on the ranking of a website in Google's search results.

As the human factor is included, it is normal to avoid unpopular practices that can drag your website down in the rankings. Intuitive and appealing designs give great edge over exceedingly long pages with huge vertical scrolls and outdated block-styled frames that have all useful text surrounded by advertisements.

Mentioning coding practices is always important but the way it impacts ranking changed a lot. In the past web pages needed a clear code with hierarchically structured heading tags and clearly defined meta tags that used to dictate ranking in the word-indexing algorithm era. Nowadays having a well-formatted code impacts mostly on the speed of indexing and the return ratio of the web crawler. This means that the crawling speed for a given website is included in the formula that defines how often the crawler will return.

Giving the users a friendly environment is all about the concept of user experience and user-centred design practices to generate cohesive, predictive and the application of desirable designs based on the expectation of users' desired experience. In most cases, User experience design fully encompasses traditional Human-Computer Interaction design and extends it by addressing all aspects of a product or service as perceived by users. ("What is user experience design?", IBM)

Small details should also be analysed. If it is possible to reach to a certain place on the website with one or two clicks less, it should be done because both the user and the rating staff are going to like that.

Carefully planning an internal relation between pages is crucial to the user experience. Having intuitive links between pages is a great way to keep visitors for longer on the website while increasing browse rate-related metrics.

5.2.2 Content impression

Talking about the content of a business oriented website is like talking about the objects in a painting. You can have a great website with a nice design and a fantastic functionality but without content it just will not make sense to anyone. The information provided in a website in a text format has always been considered a key factor in indexing and ranking for specific keywords but after the Panda update things have changed a little.

Still a certain necessity, the modern content is not so much about being grammatically correct or unique and diverse but rather about being eye-grabbing and enjoyable. The contextual relativeness is of course a must but as user satisfaction becomes the main objective more and more websites get ranked lower just because their articles or posts are plain boring.

A great example of how things changed after the Panda update are catalogue-styled online stores. Let us assume that we are making an optimization for an online retailer of plumbing parts. The web catalogue has 5 000 products and we go ahead and outsource the creation of 5 000 paragraphs, each of 200 words, that describe the technical aspect of the products one by one.

Before Panda this would be a vigorous achievement as the large amount of unique content filled with keywords would certainly result in better ranking for a lot of plumbing related keywords. After the change in the ranking algorithm this action most likely will have a negative result because the bottom line is that Google wants a quality content that their users would like to share, bookmark and even personally recommend.

5.2.3 User metrics

Last but not least I would like to mention the metrics that Google are constantly using since the release of Panda algorithm to analyze the performance of a website on a purely statistical level. The information that they collect on a daily basis is most likely the most accurate marketing data ever to exist due to the large range of products they use to gather it. The main contributors in this harvest are Google's search engine itself, Google Chrome and Google Analytics.

The algorithm can take statistical data about average values of certain key metrics in a niche and check how a certain website is compared to them, correcting a search engine ranking correspondingly. Using the specific niche's averages ensures that the evaluation is done based on real tendencies closely related to the nature of the given website.

Panda update turned the engine's attention to several metrics that can identify both user experience and satisfaction. The amount of time spent on a website is the most important metric that directly correlates to quality and user satisfaction. If the website is having users leaving faster than usual this can guide any experienced SEO specialist to the problematic web page or part of the design. Unfortunately as the ranking algorithm evolves, so does its requirements and a shorter stay of users than the niche's average will result in lower rankings.

The second important metric that people tend to neglect is the number of pages per visit. This basically defines how many pages on a given website were visited on average by each user. Having this metric above average signals to the algorithm that not only visitors like to stay on a given website but also they find the structure easy to navigate and the content interesting to read.

All important metrics in SEO nowadays are revolving around the engagement level of users and their overall experience. The websites with beautiful and practical designs are more appealing to the eye and people like to spend more time there, which brings us back to the discussion about designing a user experience.

6 CONCLUSIONS

During the development of this thesis paper I was able to expand my knowledge on this unconventional subject by referring to a large amount of literature as well as analysing a mix of information gained by my previous experience and by absorbing ideas from people that are considered as flagmen in online marketing and SEO.

I am quite aware that this paper is structured more like a literature reading than a technical datasheet and this was done on purpose. SEO is something that many people write about but few understand. Mechanical tasks that were needed in the past are mentioned over and over in countless theses papers and books all over the world. However I was not able to find a good work that dissected the philosophical side of this process and it became my desire to create such a document.

The timing of my thesis is perfectly aligned with the new era of online marketing dictated by the market leader Google and their innovative approach to the artificial intelligence and user experience. Surely, the future seems bright as we are on the verge of seeing a technological triumph in the face of Knowledge Graph and its capabilities and I am more than sure that this thesis paper will be a sufficient fundament for anyone interested in search engine optimization.

I would like to thank you for reading my work and sincerely hope that it will bring a value both to your personal and professional life.

LIST OF REFERENCES

HISTORY OF INTERNET MARKETING, FREE ENCYCLOPAEDIA OF ECOMMERCE, 2011

<http://ecommerce.hostip.info/pages/708/Marketing-Internet-HISTORY-INTERNET-MARKETING.html>, Date of retrieval: 12.02.2012

Hatthorn, R., The History of Internet Marketing

<http://www.evancarmichael.com/Home-Based-Business/5341/The-History-of-Internet-Marketing.html>, Date of retrieval: 18.02.2012

Pakroo, P., & Caputo, C. (2008). The small business start-up kit: 5th ed. Berkeley, CA

Sterne, J. 1998. World Wide Web Marketing: Integrating the Web into Your Marketing Strategy. New York. John Wiley & Sons Inc.

Web search engine, Wikipedia, 2012

http://en.wikipedia.org/wiki/Web_search_engine, Date of retrieval: 05.03.2012

Internet History - Search Engines", Universiteit Leiden, Netherlands, September 2001, web: LeidenU-Archie

<http://www.leidenuniv.nl/letteren/Internethistory/?c=7>, Date of retrieval: 05.03.2012

Net Market share - Google

<http://marketshare.hitslink.com/report.aspx?qprid=5&qpcustom=Google%20-%20Global&qptimeframe=M&qpsp=120&qpnp=25>, Date of retrieval: 11.03.2012

Edwards, J., McCurley, K. S., and Tomlin, J. A. (2001). "An adaptive model for optimizing performance of an incremental web crawler". In Proceedings of the

Tenth Conference on World Wide Web (Hong Kong: Elsevier Science): 106–113

Castillo, C., 2004. *Effective Web Crawling* (Ph.D. thesis). University of Chile.
http://chato.cl/research/crawling_thesis, Date of retrieval: 11.03.2012

Shkapenyuk, V. and Suel, T., 2002. Design and implementation of a high performance distributed web crawler. In Proceedings of the 18th International Conference on Data Engineering (ICDE), pages 357-368, San Jose, California. IEEE CS Press.)
<http://cis.poly.edu/tr/tr-cis-2001-03.pdf>, Date of retrieval: 15.03.2012

Brin, S. and Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117
<http://infolab.stanford.edu/~backrub/google.html>, Date of retrieval: 15.03.2012

Brown, E.W.: Execution Performance Issues in Full-Text Information Retrieval. Computer Science Department, University of Massachusetts at Amherst, Technical Report 95-81, October 1995

Berners-Lee, T., "Hypertext Markup Language - 2.0", RFC 1866, Network Working Group, November 1995
<http://tools.ietf.org/html/rfc1866>, Date of retrieval: 16.03.2012

Manning C.D., Raghavan P. and Schütze H., *Introduction to Information Retrieval*, Cambridge University Press. 2007, ISBN: 0521865719

Mihajlovic V., Hiemstra D., Blok H.E., Apers P. "Exploiting Query Structure and Document Structure to Improve Document Retrieval Effectiveness", 2006
<http://doc.utwente.nl/66353/>, Date of retrieval: 02.04.2012

Google's "Florida" Update, 2003
<http://www.webworkshop.net/florida-update.html>, Date of retrieval: 08.04.2012

Google Panda Update - what, why, who and what next?, Browsing Media, August 1, 2011

<http://www.browsermedia.co.uk/2011/08/01/google-panda-update-what-why-who-and-what-next/>, Date of retrieval: 02.04.2012

Fishkin, R., Seomoz.com, 2012

<http://www.seomoz.org/google-algorithm-change>, Date of retrieval: 10.04.2012

"W3C Semantic Web Activity", World Wide Web Consortium (W3C), 2011

<http://www.w3.org/2001/sw/>, Date of retrieval: 10.04.2012

Berners-Lee, T., Fischetti, M., 1999. Weaving the Web. Chapter 12.

HarperSanFrancisco. ISBN 978-0-06-251587-2.

Singhal, A, 2012, Date of retrieval: 12.04.2012

<http://mashable.com/2012/02/13/google-knowledge-graph-change-search/>

"What is user experience design?", IBM

<http://www-01.ibm.com/software/ucd/designconcepts/whatisUXD.html>, Date of retrieval: 15.04.2012