

IKT 407

Web-mining and data analysis
Faculty of Engineering and Science, Grimstad



AGDER UNIVERSITY
COLLEGE

Title: Temporal web structure mining	Report nr.: Project 4 Field of research: Computer science Number of pages.: 32
Employer: Agder University College	Date: November 26 th 2006
Authors: Sølve Oppheim Bjørn Roalkvam	Supervisor: Morten Goodwin Olsen
Theme: Web usage mining	The web evolution

Resumè:

We have described our 4 hypothesises concerning the evolution of websites and navigation and set forth a solution path to test and discover if the hypothesises are correct. The results have been evaluated and discussed and a conclusion has been reached.

Summary

We have chosen the Temporal web structure mining project, where we go deeper into the growth and change of websites over time. We show in our report that WebPages are growing; more links and picture are being used, more commercial use, webpage size and numbers drastically increasing. And that the Flash technology is replacing the complicated Java applets. This is all connected to how users navigate the through the Internet. By adding and changing the content on WebPages, navigation grows more difficult.

By using certain tools to discover the changes over time and analyzing this data, and extracting data from user tests, we have come to the conclusion that our hypotheses are indeed correct.

Preface

This project is the second grading tool in the course IKT407 Web-mining and data analysis at Agder University College in Grimstad. The project work has been done from august to the end of November 2006.

The members of the project group are Sølve Oppheim and Bjørn Roalkvam
Our supervisor has been Morten Goodwin Olsen.
Grimstad November 2006.

Table of Contents

SUMMARY.....	3
PREFACE.....	4
TABLE OF CONTENTS.....	5
TABLE OF FIGURES	6
TABLE OF TABLES.....	6
TABLE OF GRAPHS.....	6
GLOSSARY AND ABBREVIATIONS.....	7
1 INTRODUCTION.....	8
.1.1 PROBLEM DESCRIPTION - GENERAL.....	8
.1.2 THE ACTUAL PROBLEM.....	8
.1.3 MOTIVATION.....	8
.1.4 REQUIREMENTS.....	9
.1.5 SOLUTION STRATEGY.....	9
.1.6 APPROACH.....	9
2 LITERATURE REVIEW.....	10
.2.1 PROJECT BACKGROUND.....	10
.2.2 WEBSITES HYPOTHESIS.....	10
.2.3 WEBSITES ARE INCREASING.....	10
.2.4 MORE LINKS AND PICTURES THAN BEFORE.....	11
.2.5 THE USE OF WEB CONTENT TECHNOLOGIES.....	11
.2.6 INCREASE IN COMMERCIAL USE.....	13
.2.7 USERS ARE HAVING MORE TROUBLE NAVIGATING SITES.....	13
3 EXPERIMENTAL SETUP.....	15
3.1 GENERAL.....	15
3.2 QUANTITATIVE ANALYSES.....	15
3.2.1.1 Markup Validation Service.....	15
3.2.1.2 Site Link Analyzer.....	18
3.2.1.3 Web Page Analyzer.....	18
3.3 QUALITATIVE ANALYSIS.....	20
4 RESULTS.....	21
4.1 QUANTITATIVE ANALYSIS.....	21
4.1.1 Quantitative analysis for web pages from 1997, 1998 and 1999.....	21
4.1.2 Quantitative analysis for web pages from 2006.....	22
4.2 QUALITATIVE ANALYSIS.....	23
4.2.1 Qualitative analysis for websites from 1997, 1998 and 1999.....	23
4.2.2 Qualitative analysis for websites from 2006.....	23
4.3 QUANTITATIVE VS. QUALITATIVE ANALYSIS.....	23
4.3.1 Number of links within a web page.....	25
4.3.4 Number of errors in a web page.....	28
5 DISCUSSION.....	30
6 FURTHER WORK ON THIS PROJECT.....	31
7 CONCLUSION.....	31
8 REFERENCES.....	32

Table of figures

Figure 1 Flash vs. Java	10
Figure 2 Validate by URL	14
Figure 3 Validate by File Upload	14
Figure 4 Validate by Direct Input	15
Figure 5 Result of validating http://grm.hia.no	15
Figure 6 Snapshot of the Site Link Analyzer tool	16
Figure 7 Snapshot of the Web Page Analyzer	17

Table of tables

Table 3.3 Statements	18
Table 4.1 Old websites found with The Wayback	19
Table 4.2 New updated websites found at the respectively domains	20
Table 4.3 Qualitative analysis (1997, 1998,1999)	21
Table 4.4 Qualitative analysis (2006)	22

Table of graphs

Graph 4.1 Number of links within a web page (1997)	23
Graph 4.2 Number of links within a web page (2006)	23
Graph 4.3 Number of images within a web page (1997)	24
Graph 4.4 Number of images within a web page (2006)	24
Graph 4.5 Size of web pages (1997)	25
Graph 4.6 Size of web pages (2006)	25
Graph 4.7 Number of errors within a web page (1997)	26
Graph 4.8 Number of errors within a web page (2006)	26

Glossary and abbreviations

Links

Like an hyperlink, an reference in a given html document to another html document.

URL

A Uniform Resource Locator (URL) is a web address, and is a standardized address name layout for resources.

Size of web page

Size of web pages given in Bytes.

WayBack Machine

A web service which makes it possible to search for old sites within a domain. For instance find web sites of www.hia.no from back to 1997.

Web page

A web page in a given web site. Usually a html document, for instance <http://www.hia.no/index.html>.

Web site

A web site is the whole site under a domain, for instance all documents under www.hia.no

Web technology

Technologies used in web sites. For instance Flash, Java etc.

1 Introduction

.1.1 Problem description - General

Temporal web structure mining mainly focuses on the structure of the web-pages. For example growth of links, the size of each web-page, web-technologies on a website through time. Based on the topology of these structures, web structure mining can categorize the web pages and do it possible to generate information, such as similarity and relationship between different web pages. It would for example be interesting to see the change of number of links of a web page from 1997 until 2006. It would also be interesting to see how long time a user would use to find a specific web page on, let us say www.hia.no, and then compare the time when using a web page from 1997 and 2006.

As we already know the technology is changing a lot on only a few years. The size of web pages is increasing through time and the change of used technology. This growth of the size of the web pages and change in technology could lead to more complex web pages, which again would do it more complex for low level users.

.1.2 The actual problem

This project focuses on doing a research where the goal is to map the web-content evolution. The main goal is then to compare different web pages from different years and then take a closer look at the changes of all these web pages. It would also be done a research with 5 users who will do some tasks (see chapter 4.2) to see if growth of web pages and changes of technology really would do it more complex or easier for different users. We will then make a conclusion of these tasks and the analysis of the web pages with different analysis tools. The original project description is given below;

The goal of the project is to do an analysis of the change, growth, accessibility and interlinking of web pages over time, by using The Wayback Machine in combination with a web crawler. The case study should include AUC's home page(<http://www.eiao.net/webmining/projects>).

.1.3 Motivation

Our motivation for doing this research is to get a confirmation on our hypotheses. In addition to these hypotheses it would be interesting to see if it is possible to make a conclusion with both practical analysis with data tools and a research with users involved.

.1.4 Requirements

To solve this project it would be essential with some tools to do the analysis. The list of required tools in our project is listed below.

Relaxed (used mostly for accessibility)

HTML – validator (mostly used for analysis of web pages)

- Web Page Analyzer™
- Site Link Analyzer
- Markup Validation Service

In addition to these tools it would be necessary to use different web browsers (to browse web pages), notepad (to analyze web pages manually), Wayback Machine (a online service tool used to search for a number of saved web page of for example www.hia.no) and a web crawler (search engine, for example www.google.com).

.1.5 Solution strategy

The strategy for solving this project will be a combination of two main approaches which is 1. Doing statistical analysis with either a finished developed analysis tools or develop a tool by our self, or 2. Doing a research where we catch the user experience through time.

By combining these two approaches we will be able to do a conclusion based on both analysis tools and user experience through time. We think it can be very interesting to see if this really can be possible.

In both approaches we will use the same web pages to make statistics of. When doing the first approach we will only use one or more analyze tools to make useful statistics which can be used to compare with the user experienced analyze. When doing the second approach we will use a few question (tasks) which the users are supposed to solve. Based on these tasks we will make statistics which will be compared with the statistics from the first approach (as mentioned above).

.1.6 Approach

The project consists of several tasks. The first step is to define our hypothesis questions through a problem description, then on to a literature review to gather the most up to date data on the subjects. The third step is to use the chosen tools to gain knowledge of our hypotheses through analysis and evaluation of results discovered through research and from user surveys. A discussion is the forth step, that will end in a conclusion.

2 Literature review

.2.1 Project background

Web structure mining is the process of using the graph theory to analyse the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds: The first kind of web structure mining is extract patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyse and describe the HTML (Hyper Text Markup Language) or XML (eXtensible Markup Language) tags within the web page [8].

Over the last years the web has grown enormously in size, importance and finally but not last, in complexity. As the websites become more and more complex and is used by more and more people a proper understanding of how a site functions becomes important in order to organize the content in the most efficient way.

This project will focus on how a website is navigated, employing “web usage mining” techniques to find user traversal patterns. Based on these patterns the project will look at how to analyze them to see what information they can reveal and finally how this new information can be used to improve navigability.

.2.2 Websites Hypothesis

We have put forth certain hypothesis which we explain in the following subchapter. Our first research question concerns websites increasing over time in size. We will use the Wayback Machine and similar applications to determine this. Our second hypothesis being that more pictures are used compared to past uses is directly connected to our first hypothesis.

The use of java and flash - technologies on websites is changing, java applets decreasing while more and more go for the flash technology. Our forth hypothesis is the increase commercial use on websites. All four hypotheses are connected in a way that we think websites demand more resources to be viewed. Websites seem to grow continuously as more powerful computers and faster internet connections are available, keeping a certain balance around the time to access a site and time to find a site.

.2.3 Websites are increasing

First of all we should define a website which is nicely described by Answers.com [1]: A Web page is a text document embedded with HTML tags that define how the text is rendered on screen. Web pages can be created with any text editor or word processor. They are also created in HTML authoring programs that provide a graphical interface for designing the layout. Authoring programs generate the HTML tags behind the scenes, but the tags can be edited if required. Many applications export documents directly to HTML, thus basic Web pages can be created in numerous ways without

HTML coding. The ease of page creation helped fuel the Web's growth. A collection of Web pages makes up a Web site.

Websites continuously increase in size, where faster connection and powerful computers make it possible for people to access and view this new internet content and it follows: "The Internet is an efficient and cost-effective resource for disseminating public information. However, as the amount of Internet content increases, it's increasingly difficult to find information. Reducing the complexity of locating information is a high priority research area." [2], as the goal of any good website logically is to have a short access time and being easy to find and understand.

Although it is said that only death and taxes are certain, it seems most likely that the Internet and the World Wide Web will only continue to increase in size and complexity. As more and more individuals, universities, and businesses gain access, the amount and quality of information will continue to expand. At the same time, the amount of disinformation and propaganda will also expand [7].

.2.4 More Links and Pictures than before

The use of pictures and links, more hits on their website, and more pictures making it more pleasant for visitors. Better understanding. But pictures again use more space and takes longer to load than normal contents. A certain balance seems appropriate.

In the 1990s Internet speeds increased, and Internet browsers capable of viewing images were released, the first being Mosaic. Websites began to use the GIF format to distribute small graphics such as banners, advertisements and navigation buttons on web pages. Web graphics are useful in providing a truly graphical user interface to websites rather than plain text. Modern browsers now support the use of jpeg, png and increasingly svg images in addition to gifs on web pages[9].

Numerous websites have been created to host communities for web graphics artists. A growing community consists of people who use photoshop or paint shop pro to create forum signatures and other digital artwork [9]

.2.5 The use of web content technologies

It was predicted that java would be the leading technology used in websites but Java never gained as much acceptance as Sun had hoped as a platform for client-side applets for a variety of reasons, including lack of integration with other content (applets were confined to small boxes within the rendered page) and poor performance (particularly start up delays) of Java VMs on PC hardware of that time [1].” It is also said “In five years, Java EE will be the CORBA of the 21st Century. People will look at it and say, 'It had its time but nobody uses it any more because it was too complicated” [3] Further it is said: "Java EE's days have been numbered for a while now," said Jason Bloomberg, senior analyst with ZapThink LLC, who also sees the main culprit being the increased complexity that comes with each new version. "Clearly, every time a new

version comes out or module gets added, it only adds to the complexity. Eventually, it'll simply collapse under its own weight. It's not like there will be a future version of Java EE that's more lightweight than its predecessor."

Complexity aside, Bloomberg, who specializes with SOA and Web services, sees the Java platform as fatally flawed when it comes to moving into the service-oriented enterprise era.[3]. The Flash technology on the other hand is growing "Macromedia Flash is the standard for creating rich interactive content. Every major website uses it."

April 2006: Java is losing the battle of the browser plug-in technologies with about 57% penetration. The Flash Player is ubiquitous with nearly 96% of users being able to view Flash content [4].

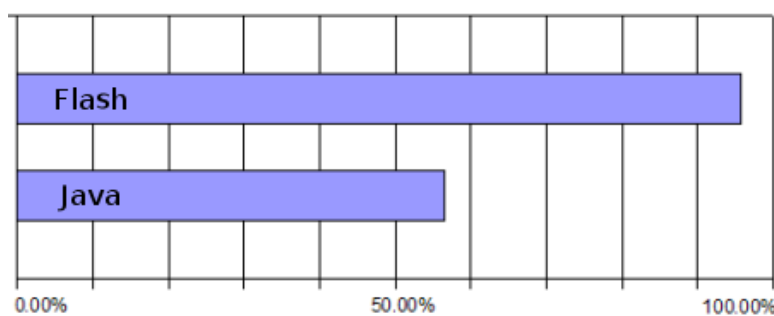


Figure 1 Flash vs. Java

A very interesting fact is that the 7-years old MS VM 1.1.4 still accounts for more than the half of Java penetration; with new versions of Windows shipping without that JVM, it is easy to predict that the whole Java share will continue to shrink. What is even worse is that developers who target Java on the browser side, will still be forced to use MS JVM, at least for a few more months [4].

Flash, on the other hand is really up-to-date, with its latest version 8 reaching nearly 75% of all installations.

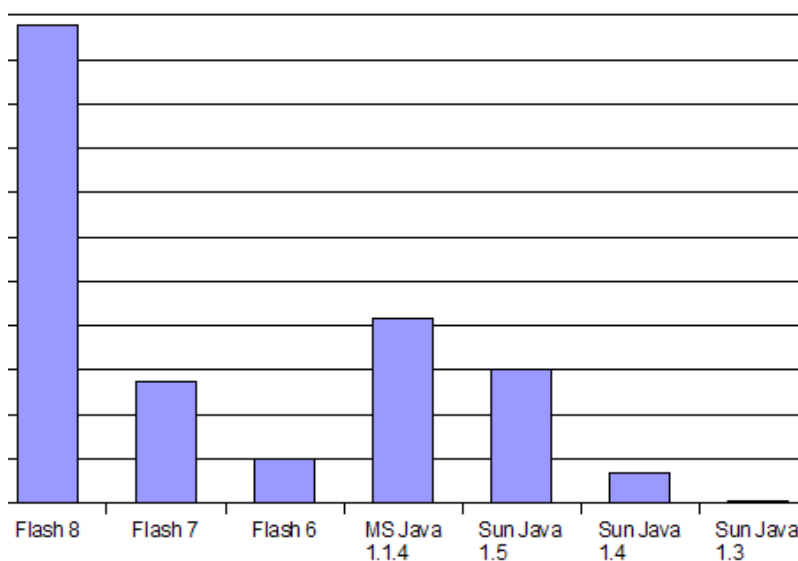


Figure 1 Flash vs. Java

As we can see; Flash is an easier technology to use and more adaptable. Flash has become a universal standard for the delivery of animations on the Internet [5]. Also most pc's today are ready for the flash technology by having the necessary plug-in installed "Recent statistics show that more than 99% of Internet users, so virtually everyone, have the Flash 4 plug-in installed" [5].

.2.6 Increase in commercial use

As it has been more clearly acknowledged that people are affected by commercials on the internet, more resources have been set to take advantage of this. Especially today commercials are often created by the Flash technology. There is hardly a site to be seen without Google ads. Commercials toward every type of product and service can be found on the Internet. Many large websites are run on the money they get for publishing ads and commercials on their sites. "By 1996 it became obvious to most publicly traded companies that a public Web presence was no longer optional. Though at first people saw mainly the possibilities of free publishing and instant worldwide information, increasing familiarity with two-way communication over the "Web" led to the possibility of direct Web-based commerce (e-commerce) and instantaneous group communications worldwide" [6]. Commercial use starts its ever-growing climb to conquer as many Internet users possible.

.2.7 Users are having more trouble navigating sites

Speed issues: Frustration over congestion issues in the Internet infrastructure and the high latency that results in slow browsing has led to an alternative name for the World Wide Web: the *World Wide Wait*. Speeding up the Internet is an ongoing discussion over the use of peering and QoS technologies. Other solutions to reduce the World Wide Wait can be found on W3C.

Science, in particular the social sciences, has been slow to adopt this method of imparting information. Mass media reporting of the Internet as a dark and dangerous place, replete with computer hackers and child molesters, excesses of pornography, and rampant urban legends of security leaks and computer viruses have all played a role in retarding both the growth of the Web and its use by consumers. However, as these issues fade with time and familiarity, some researchers are finding that communicating information without the intervening noise of the press is proving too attractive to dismiss out of hand. Faculty are finding that the Internet allows them to make course-specific material available to everyone in their classes, without having to find a publisher, edit for the widest possible audience and then wait two years for the book to come out [8]

3 Experimental setup

3.1 General

In this section we will describe how to do the analysis. Several validate tools will be tried, to see if we get the similar result. This part is the so called quantitative analyses. The manually analyses will be described shortly, with an explanation on how we are planning to do it. The last part will describe the qualitative analyses, where the user experience is in focus.

To find actual web pages that can be analysed, we are using a so called Wayback Machine [10] to find web pages saved in different years and dates. In addition to this Wayback Machine we need some analyses tools. After testing some tools we have decided to use the following tools for these analyses:

- Markup Validation Service [12]
- Site Link Analyzer [13]
- Web Page Analyzer 0.961 [14]

3.2 Quantitative analyses

To get relevant information about web pages, we used several tools mentioned above. This part of this chapter will be used to present these tools closer. The actual tools for the manual part of the quantitative analysis, which mostly would be notepad, will be used to take a look at the source code for the different websites to see if there are implemented technologies such as java, CSS and flash. One of the analyze tools (Web Page Analyzer) would however determine if there is java or CSS implemented.

3.2.1.1 Markup Validation Service

The Markup Validator is a free service by W3C that helps check the validity of Web documents.

Most Web documents are written using markup languages, such as HTML or XHTML. These languages are defined by technical specifications, which usually include a machine-readable formal grammar (and vocabulary). The act of checking a document against these constraints is called validation, and this is what the Markup Validator does. Validating Web documents is an important step which can dramatically help improving and ensuring their quality and it can save a lot of time and money (read more on why validating matters). Validation is, however, neither a full quality check, nor is it strictly equivalent to checking for conformance to the specification.

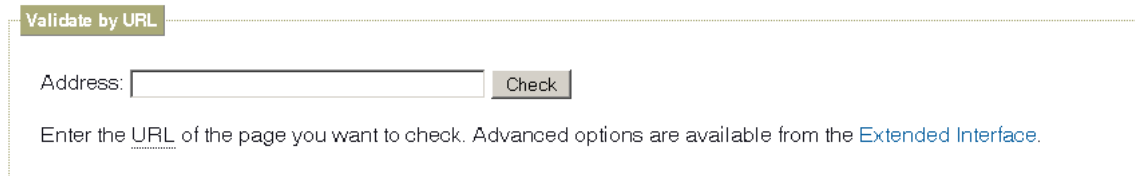
This validator can process documents written in most markup languages. Supported document types include the HTML (through HTML 4.01) and XHTML (1.0 and 1.1) family, MathML, SMIL and SVG (1.0 and 1.1, including the mobile profiles). The Markup Validator can also validate Web documents written with an SGML or XML DTD, provided they use a proper document type declaration [12].

There are three options when validate a website (however the first one is most actual for our use);

1. Validate by URL (validate an external web page direct from the W3C server)
2. Validate by File Upload (upload a file to the web server and then validate it)
3. Validate by Direct Input (validate a web page by paste some codes into a text field)

This analyses tool gives us information about;

- Modified
- Server
- Size
- Content-type
- Encoding
- Document type

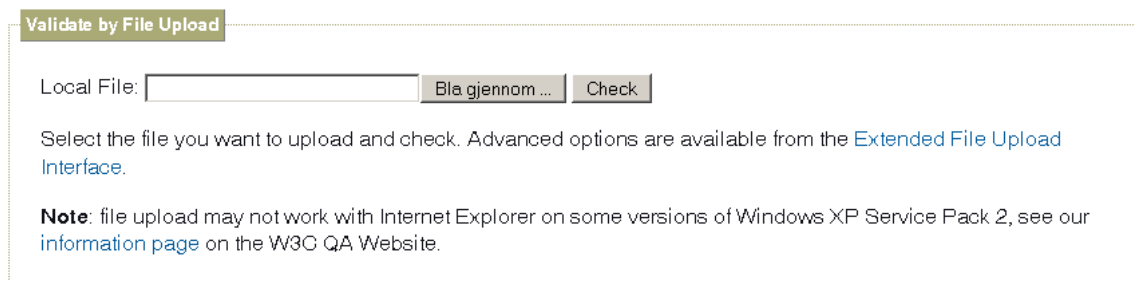


Validate by URL

Address:

Enter the URL of the page you want to check. Advanced options are available from the [Extended Interface](#).

Figure 2 Validate by URL



Validate by File Upload

Local File:

Select the file you want to upload and check. Advanced options are available from the [Extended File Upload Interface](#).

Note: file upload may not work with Internet Explorer on some versions of Windows XP Service Pack 2, see our [information page](#) on the W3C QA Website.

Figure 3 Validate by File Upload

Validate by Direct Input

Input the markup you would like to validate in the text area below.

Only complete documents (along with a [Doctype declaration](#)) will be validated. Advanced options are available from the [Extended Direct Input Interface](#).

Figure 4 Validate by Direct Input

Validating web pages

In this part we will give you an example of validating a web page. Actually a validating does not only give us a validating, but also useful information about a web page. The example below is given from <http://grm.hia.no> .

Result: Failed validation, 47 errors

Address:

Modified: Fri Oct 13 12:17:54 2006

Server: Microsoft-IIS/6.0

Size: 26861

Content-Type: text/html

Encoding: iso-8859-1

Doctype: (no Doctype found)

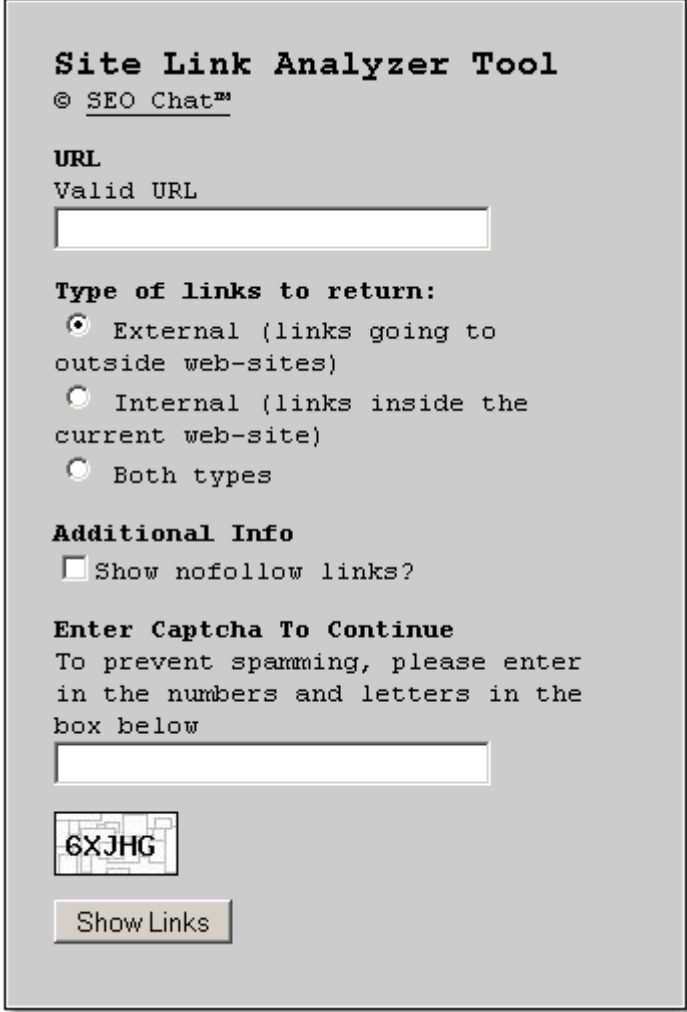
Figure 5 Result of validating “http://grm.hia.no”

As we can see from figure 5 there are 47 validation errors on <http://grm.hia.no>. The validation result also tells us when the server was last modified, what kind of server is used, size on the web page, content-type and encoding. But as we see it does not give us any information about document type. This is probably because of one of the 47 error given in the result. We will come back to this later in this analysis part.

3.2.1.2 Site Link Analyzer

The Site Link Analyzer is developed by SEO chat™, and is analyzing a given web page and then return a table with the links with their associated text. The tool also calculates the total number of links, based on internal and external links.

In this project the service tool will only be used to calculate the total number of links within a web page, to compare a web page from different years.



The screenshot shows a web-based form titled "Site Link Analyzer Tool" with a copyright notice for "SEO Chat™". The form includes a "URL" section with a "Valid URL" label and an empty text input field. Below this is a "Type of links to return:" section with three radio button options: "External (links going to outside web-sites)", "Internal (links inside the current web-site)", and "Both types". The "External" option is selected. There is an "Additional Info" section with a checkbox labeled "Show nofollow links?". A "Enter Captcha To Continue" section follows, with a text instruction and an empty input field. Below the input field is a captcha image showing the characters "6XJHG". At the bottom of the form is a "Show Links" button.

Figure 6 Snapshot of the Site Link Analyzer Tool

3.2.1.3 Web Page Analyzer

The Web Page Analyzer™ is a web based analysis tool which generates useful information about a web page. In our project the analysis tool will be used to find the following information;

- Number of total HTTP request

- Total size of web page
- Total number of images
- Download time

Web Page Analyzer - 0.961 - from Website Optimization

Free Website Performance Tool and Web Page Speed Analysis

Try our free web site speed test to improve website performance. Enter a URL below to calculate page size, composition, and download time. The script calculates the size of individual elements and sums up each type of web page component. Based on these page characteristics the script then offers advice on how to improve page load time. The script incorporates best practices from [HCI research](#) and web site optimization techniques into its recommendations.

Enter URL to diagnose:

Or cut and paste (X)HTML and an optional base href URL to resolve relative URLs:

Enter (X)HTML to diagnose:

Enter optional <base href=> URL
Example: domain.com:

Figure 7 Snapshot of the Web Page Analyzer

3.3 Qualitative analysis

The qualitative analyses are performed by a users experience when browsing a web page given in the quantitative analysis. It is then natural to make tasks where the answers and the average statistics can be indirectly compared to the result we got in the quantitative analysis.

The tasks for this part of the analysis are given in table 3.3 which is made up of claims where the users rank their experience from 1 to 5. 1 being the test users strongly disagree opinion and 5 being that they strongly agree with the statement.

	Statements	Score
1	When looking at this site's amount of Webpage content (pictures, links, data size), it is difficult to navigate this site.	
2	A lot Flash Technology is used in this site.	
3	The commercial use on this site is extensive.	
4	There are too many links on this site.	
5	There are too many pictures on this site.	
6	The site is too large in size, takes a long time to load.	

Table 3.3 Statements.

4 Results

In this chapter we will present the result of the qualitative and quantitative analysis. The first part is the quantitative analysis which gives us an overview on the evolution in the web technology. This means that we would be able to compare a web page from 1997 with the same web page in 2006, and then see the difference in for instance size of web page, number of links, images etc. In the second part we are presenting the result of the quantitative analysis, which will give us some information on how the user experience is in proportion to the result we got in the qualitative analysis.

We would also compare both the qualitative and quantitative analysis closer, to see if it is possible to make a conclusion on the evolution.

4.1 Quantitative analysis

This part of the analysis is the automatic part, which will be performed by the three mentioned data analyze tools. We are using ten web pages, which will be used through both the quantitative and qualitative analysis.

It is important to know that these data can vary a bit when it comes to the automatic data analysis, because there could be issues in the data analysis services, which again would give us wrong values. This means that all websites would give us the correct result and information. One of the data tools will for instance not give us the number of errors when the validating is not succeeded. The download time could also vary a bit, because of delays on the Internet and so on.

The quantitative analysis table is divided into two tables, to have a better overview. Thus there is one table for the old web sites and one table for the new websites.

4.1.1 Quantitative analysis for web pages from 1997, 1998 and 1999

Year	URL	Size	Links	Images	Errors	Java	CSS	Flash	DT(s)
1997	http://www.hia.no	5276	29	1	34	-	-	-	1,05
1997	http://grm.hia.no	7604	13	18	-	-	-	-	1,52
1997	http://www.telenor.no	49765	14	15	8	-	-	-	9,92
1997	http://www.kreditkassen.no	5344	5	5	-	-	-	-	1,07
1997	http://www.gulesider.no	8223	48	4	-	-	-	-	1,64
1997	http://odin.dep.no	10454	60	12	-	-	-	-	2,08
1997	http://www.ntnu.no	29106	21	4	-	-	-	-	5,80
1997	http://www.aftenposten.no	97896	18	21	44	X	X	-	20,11
1998	http://www.tv3.no	13431	14	15	-	-	-	-	2,68
1999	http://www.microsoft.no	165877	67	12	-	X	-	-	33,06

Table 4.1 Old websites found with The Wayback Machine at <http://www.archive.org/web/web.php>

4.1.2 Quantitative analysis for web pages from 2006

Year	URL	Size	Links	Images	Errors	Java	CSS	Flash	DT(s)
2006	http://www.hia.no	175695	23	21	-	X	X	-	35,62
2006	http://grm.hia.no	106147	98	52	47	X	X	-	21,58
2006	http://www.telenor.no	71259	8	10	0	X	X	-	14,40
2006	http://www.nordea.no	248682	54	59	95	X	X	-	50,56
2006	http://www.gulesider.no	108090	33	11	40	X	X	-	23,94
2006	http://odin.dep.no	59581	77	11	19	X	X	-	12,47
2006	http://www.ntnu.no	262439	76	15	35	-	X	-	52,50
2006	http://www.aftenposten.no	276006	550	97	25	X	X	X	55,61
2006	http://www.tv3.no	412308	32	120	120	X	X	X	83,17
2006	http://www.microsoft.no	256549	101	17	2	X	X	X	51,33

Table 4.2 New updated websites found at the respectively domains

Year:	Time for saved version of web page
Website:	URL for the given web page
Size:	Size of web page given in Bytes
Links:	Number of internal and external links on the given web page
Errors:	Number of errors (if validation ok)
Encoding:	Type of encoding of the web page
Java:	Is there java technology implemented on the web page?
CSS:	Is there CSS technology implemented on the web page?
DT(s):	Download Time in second for a 56K connection rate
*	The domain has changed, see A-X for details.
-	No value found

4.2 Qualitative analysis

Old versions of the websites vs. today's version of the same sites:

The higher sum, the more trouble navigating, being confused. Lowest 6 and highest 30.

4.2.1 Qualitative analysis for websites from 1997, 1998 and 1999

Page	Year	URL	User 1	User 2	User 3	User 4	User 5	Average
1	1997	http://www.hia.no	6	7	6	7	6	6,4
2	1997	http://grm.hia.no	7	7	8	6	6	6,8
3	1997	http://www.telenor.no	8	9	8	9	9	8,6
4	1997	http://www.kreditkassen.no	6	6	6	8	7	6,6
5	1997	http://www.gulesider.no	7	9	7	7	6	7,2
6	1997	http://odin.dep.no	6	7	7	7	8	7
7	1997	http://www.ntnu.no	9	8	8	8	9	8,4
8	1997	http://www.aftenposten.no	6	6	7	6	8	6,6
9	1998	http://www.tv3.no	6	6	7	6	7	6,4
10	1999	http://www.microsoft.no	8	6	8	8	6	7,2
Total	Average							7,12

Table 4.3 Qualitative analysis (1997, 1998,1999)

4.2.2 Qualitative analysis for websites from 2006

Page	Year	URL	User 1	User 2	User 3	User 4	User 5	Average
1	2006	http://www.hia.no	14	12	15	13	14	13,6
2	2006	http://grm.hia.no	13	13	15	14	14	13,8
3	2006	http://www.telenor.no	22	25	22	21	27	23,4
4	2006	http://www.kreditkassen.no	16	15	14	20	15	16
5	2006	http://www.gulesider.no	12	14	10	12	16	12,8
6	2006	http://odin.dep.no	21	21	23	18	19	20,4
7	2006	http://www.ntnu.no	17	21	15	25	20	19,6
8	2006	http://www.aftenposten.no	27	25	26	26	29	26,6
9	2006	http://www.tv3.no	23	25	20	25	21	22,8
10	2006	http://www.microsoft.no	28	26	27	27	28	27,2
Total	Average							19,62

Table 4.4 Qualitative analysis (2006)

4.3 Quantitative vs. Qualitative analysis

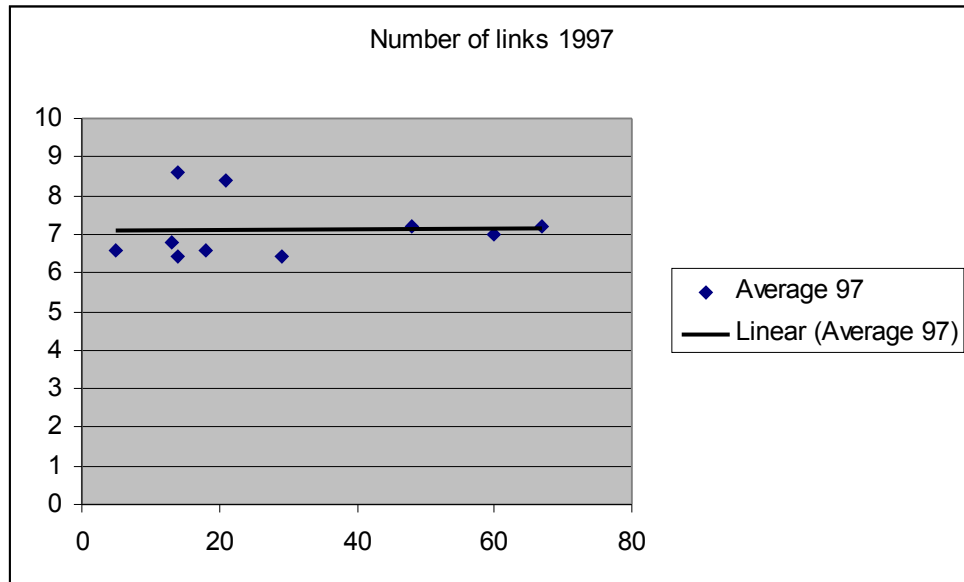
This section will describe the eventually connection between the quantitative analysis in chapter 4.1 and the qualitative analysis in chapter 4.2.

To see if there is any relation between the two analyses, we will put the results into scatter diagrams. This will be done by each value (except of the download time, since this will be the same diagram as the size) in table 4.1 and table 4.2 in the quantitative HiA, Grimstad
November 2006
Page 23 of 33

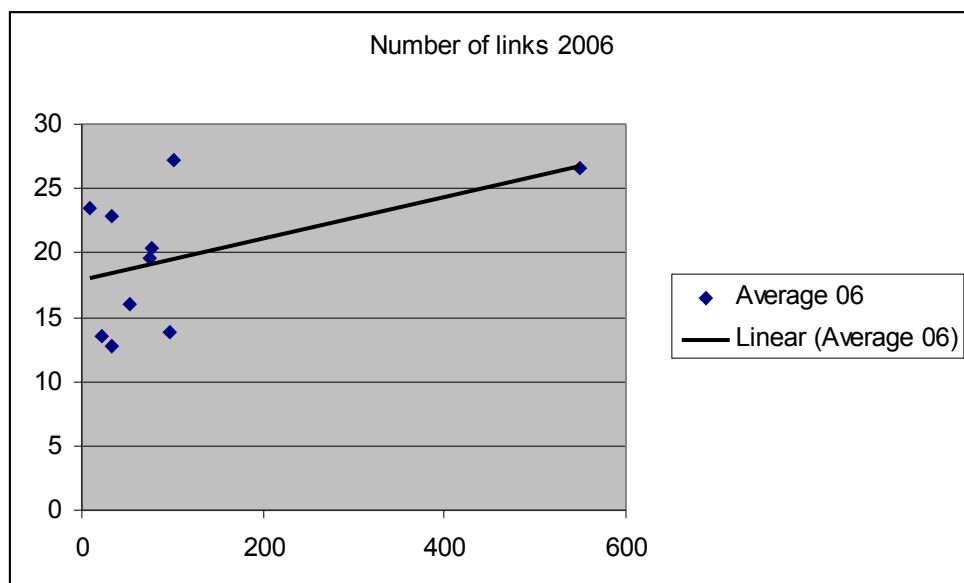
analyze will make up to us if we put them together.

It is important to keep in mind that these analysis is based on only 10 web pages from only two different dates (years). Also, that the user experience survey only is based on the experience of five persons, which is familiar with the use of computer and Internet. Anyway, it would maybe be possible to see an indication, but we won't conclude with anything based on the data we got in this project. To make a conclusion it would be necessary with a lot more data and of course many more users in the survey.

4.3.1 Number of links within a web page



Graph 4.1 Number of links within a web page (1997)

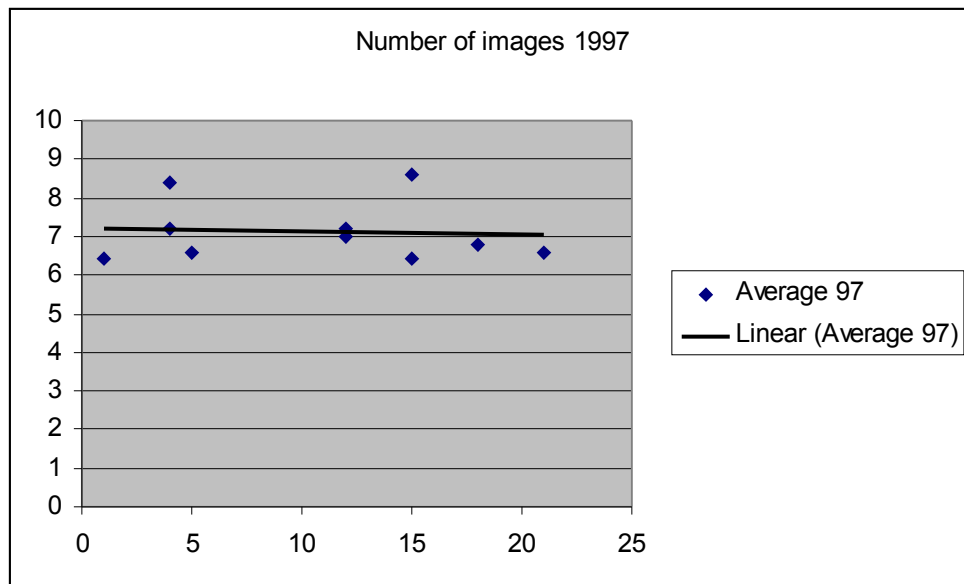


Graph 4.2 Number of links within a web page (2006)

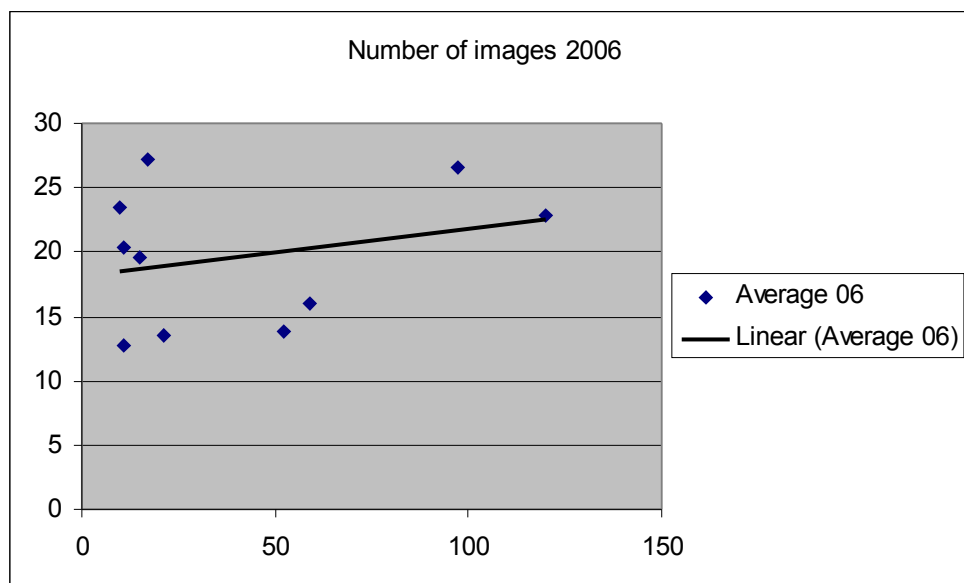
These graphs show the relation between the user experience and the number of links within a web page from 1997 and 2006.

Both graphs gives an increasingly curve. This means that the users are less satisfied with a web page with many links. The graph for web pages from 1997 has a less steep curve than the graph from 2006. This could be because the links on newer web pages are better organized than on old web pages.

4.3.2 Number of images within a web page



Graph 4.3 Number of images within a web page (1997)



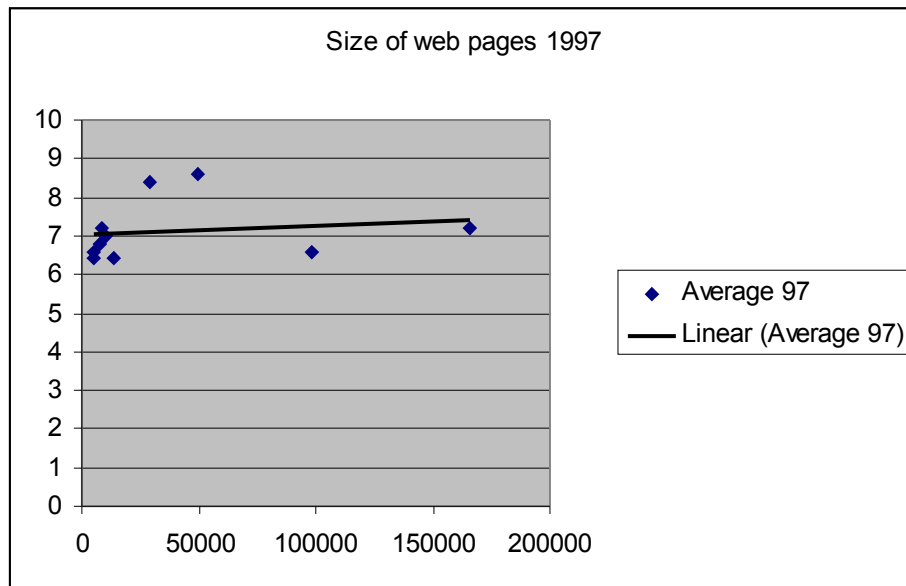
Graph 4.4 Number of images within a web page (2006)

These graphs show the relation between the user experience and the number of images within a web page from 1997 and 2006.

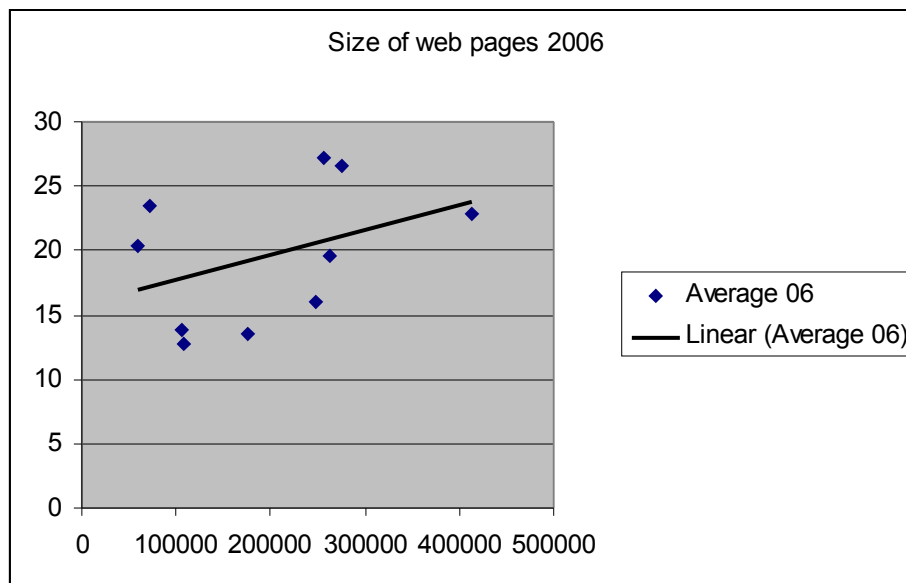
These graphs are the only ones which have two linear curves with different directions. The first graph (1997) indicates that the users are less satisfied when the number of images on a given web page is increasing. The second graph (2006) indicates that the users are more satisfied when the number of images is increasing.

The difference between the websites could be a result of the development of web pages related to technology and overview.

4.3.3 Size of web pages



Graph 4.5 Size of web pages (1997)

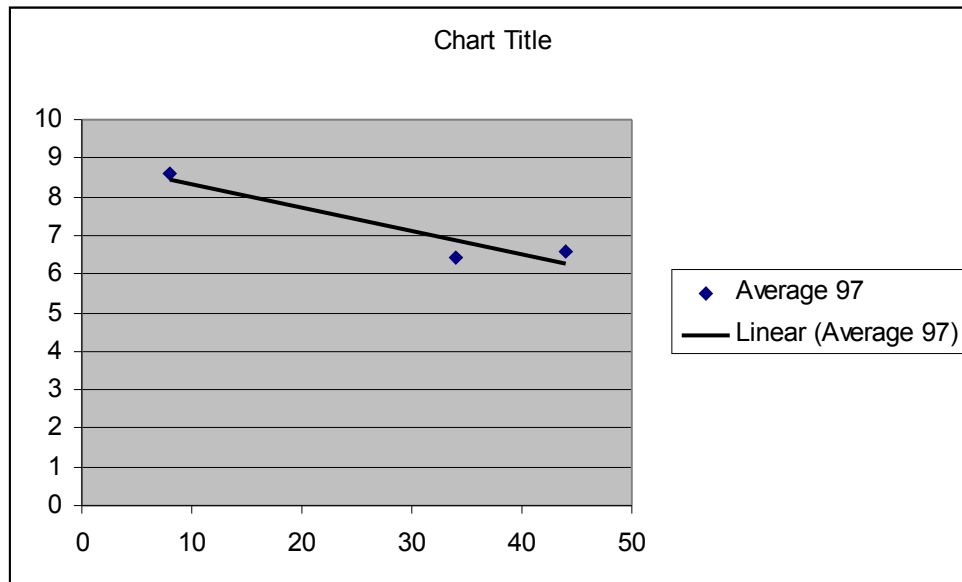


Graph 4.6 Size of web pages (2006)

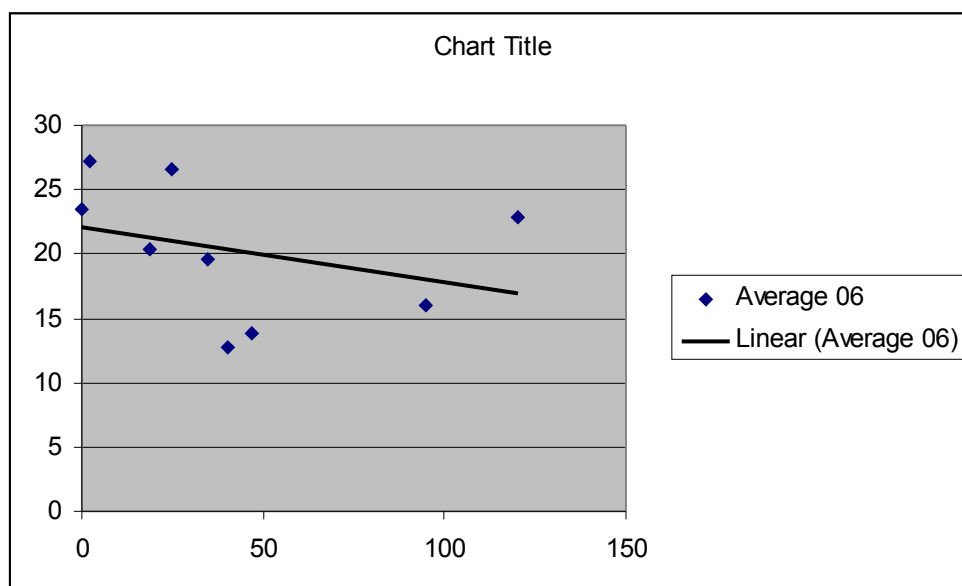
These graphs show the relation between the user experience and size of a web page from 1997 and 2006.

As we can see of these graphs, there is a positive linear curve in both 1997 and 2006. This means that the users are less satisfied with the web page when the size of the web page is increasing.

4.3.4 Number of errors in a web page



Graph 4.7 Number of errors within a web page (1997)



Graph 4.8 Number of errors within a web page (2006)

These graphs show the relation between the user experience and number of errors in the source code on a web page from 1997 and 2006.

Because of an error in the Markup Validation Service, we were only able to get the number of errors on some of the web pages which were analyzed. Especially on web pages from 1997 there were some problems with getting a successful validating.

As we see in both graphs there is an indication that users are more satisfied with web pages which contains several errors in the source code.

5 Discussion

First of all we have to mention that the quantitative data analysis should have been analysis of websites, but there were no available analysis tools which would generate information for the whole website for both dates (1997 and 2006). There were some actual web spiders and web crawlers which could generate and eventually download the websites for us which gave us a useful result for the new updated websites (from 2006). But when we used the wayback machine and tried to collect some information based on a given link from this archive (for instance the address <http://web.archive.org/web/19970502232742/http://www.grm.hia.no/>) was not possible to generate any information about the web pages from level 2, only the first level. We even tried these websites in the analyze tools developed in a similar project in IKT407 2004, but this service contained a lots of bugs, so it was impossible to use. This issue leads us to perform a qualitative analyze based on only the first web page on a given web site.

More time can be used on the html validator. The result from this validator tool gives us information about the errors in a given web page. This could be useful when for instance programming a web page especially for a blind person. In these cases it would be necessary to for instance have an “ALT” value in the source code for a picture. This is because a blind person can’t see the picture, but reads the values in the “ALT” attribute. All web pages analyzed in this project had at least one error (when it was possible to get a value) unless of www.telenor.no (2006) which was the only web page which passed the validator test.

However, we think that the quantitative analyze we did with each web page from a given website (all from the index file), will give us an indication on what the tendencies would be among the users (in the qualitative analysis) when surfing the World Wide Web based on the data we got in the analysis.

Even if the data we gathered by the data tools looks true, there could be some errors. As mentioned earlier in this report, there could be some issues in the data analysis tools which would give wrong values, and then give us a wrong statistics, which last but not least could give us a wrong indication on the relation between the quantitative and qualitative analysis.

6 Further work on this project

The result of this project could be more interesting if there were available analysis tools which work 100% in the case with the Wayback Machine and in cases where the websites and web pages are quite big. The crawler developed in web mining 2004 contained some bugs which could be fixed, like it is described in the report [8]

To continue the work on this project there would be needed better data analysis tools, and it would be necessary with more users and more websites or as in our case web pages to get a better indication on whether the quantitative and qualitative analysis can be comparative.

7 Conclusion

As time goes by, WebPages in websites do indeed increase as demonstrated in our quantitative- and qualitative analysis. The dates picked (1997 and 2006) were specifically chosen to show more clearly how much the Web is “growing”.

There are various reasons why for example pictures and the Flash technology steadily increases related to more attractive user experiences for viewers and easier, cheaper solutions for the companies providing the content. The growth of WebPages in size and numbers are a result of better recourses as faster internet connections and newer technology is available.

Although the number of sites increases, this directly makes finding specific information on the Web continuously more difficult. The navigation experience is further interrupted by endless links and enforced commercial ads. Our quantitative user survey directly asks the test persons what they think about certain statements in our chosen websites, their answers reveal that our hypotheses are correct. Although the Web is growing, and resources are making it easy to process more content, the goal of any site is to make navigations and user ability easy and enjoyable for intended users. As our goal in this paper was to discover if our hypotheses were correct, in the qualitative analysis however we were only able to test specific pages of the sites in the qualitative analyses but they are a good indication of the site.

Together the quantitative- and qualitative analysis has supported our hypothesis, but to make an even better qualitative analyses, further research and better tools are required to analyses whole sites.

8 References

- [1] Encyclopedic in almanacopedia 2006.
<http://answers.com>
- [2] National institute for occupational safety and health - Improved Technology Transfer via the Web 2006
<http://www.cdc.gov/niosh/nas/mining/potentialintermediateoutcome66.htm>
- [3] Analysts see Java EE dying in an SOA world 2006.
http://searchwebservices.techtarget.com/originalContent/0,289142,sid26_gci1198211,00.html
- [4] Java vs. Flash which technology dominates on the client side 2006.
<http://www.realchat.com/blog/java-vs-flash/>
- [5] Flash facts 2006.
<http://www.mix-fx.com/flashfacts.html>
- [6] History of the World Wide Web – Wikipedia 2006.
http://en.wikipedia.org/w/index.php?title=History_of_the_World_Wide_Web&oldid=76754416 >]
- [7] Mining the Web: Techniques for Bridging the gap between content Producers and Consumers 1997.
http://www.firstmonday.org/issues/issue2_10/hirst/index.html
- [8] "Web mining." *Wikipedia* " 2006.
http://en.wikipedia.org/w/index.php?title=Web_mining&oldid=90226471
- [9] "Graphics." *Wikipedia*
<http://en.wikipedia.org/w/index.php?title=Graphics&oldid=89544560>
- [10] WayBackMachine
<http://www.archive.org/web/web.php>
- [12] Markup Validation Service version 0.7.4
<http://validator.w3.org/>
- [13] Site Link Analyzer
<http://www.seoachat.com/seo-tools/site-link-analyzer/>
- [14] Web Page Analyzer
<http://www.websiteoptimization.com/services/analyze/>
- [15] Web mining research, Raymond Kosala & Hendrik Blockeel
<http://delivery.acm.org/10.1145/370000/360406/p1-kosala.pdf?key1=360406&key2=3706683611&coll=portal&dl=ACM&CFID=1111111&CFTOKEN=2222222>

- [16] Web mining technology and Academic Librarianship
http://www.firstmonday.org/issues/issue4_6/chau/index.html
- [17] Survey: Website navigation, Kathleen Kotwica
<http://www.cio.com/research/behavior/edit/survey6.html>
- [18] Web Structure Mining, project 2004