Comparative usability evaluation of web systems through ActivityLens

Georgios Fiotakis, Christos Fidas, Nikolaos Avouris

Electrical & Computer Engineering Department, HCI Group,
University of Patras, 26500 Rio Patras, Greece
fiotakis@ece.upatras.gr, fidas@ece.upatras.gr, Avouris@upatras.gr

Abstract

The purpose of this paper is to present the main characteristics of ActivityLens (AL), an environment that facilitates comparative usability evaluation of web sites. AL supports the usability evaluation process with innovative tools designed for this purpose. The comparative usability evaluation is based on a usability assessment model that is also described briefly. Finally, a comparative usability evaluation case study, using AL, of two competitive shipping companies is presented.

Keywords: Comparative usability evaluation, usability evaluation, data analysis tools

1. Introduction

The usability of a software system is an important aspect since it determines to a great percentage whether the provided features will be used by its users or not [Nielsen (1993), Dix et al. (1998)]. Usability refers to whether a system can be used with effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in a particular context of use [ISO 9241-11].

Contrary to traditional desktop applications, usability in web systems is even more a crucial aspect for their success since the user's are able to switch to a competitor's web site easily and fast since it is "only a click away" [Hahn et al. (2006), Johnson et al. (2003)]. Simultaneously, the rapidly growth of the web and the plethora of web services that promise to address similar user requirements makes the selection of the most appropriate web system a difficult task [Brinck et al., 2001].

In this context, comparative usability evaluation of web sites, that address similar user needs, could provide valuable feedback to users in order to choose the best designed web system in terms of task success, completion time, task difficulty and user satisfaction [Brinck at al., 2001]. Although, the term comparative usability evaluation is commonly used in order to describe the comparison of the usability evaluation of two different versions of the same web system [Molich et al., 2004], we use this term

G.Fiotakis, C. Fidas, N. Avouris, Comparative usability evaluation of web systems through ActivityLens, Proc. PCI 2007, May 2007, Patras, Greece

in order to refer to the usability evaluation of different web systems that focus on the same target group of people with similar user requirements. Competitive usability analysis is conducted in order to examine strengths and weaknesses of the evaluated web systems and how a specific system stands against those of competitors and what changes would work up its competitive advantage.

In this context, tools that support the comparative evaluation process could be of significant value. In this paper ActivityLens (AL), a tool designed to support comparative usability analysis of web systems is presented. The architecture of AL is based in ColAT (Collaboration Analysis Tool) a tool for analysis of activities in which more than one actor is involved [Avouris et. al. (2007), Stoica et. al. (2005)].

The paper is organized as follows. The next section presents the process of comparative usability evaluation aiming to describe in detail the lifecycle of such studies and concludes with requirements related with data collection, data integration and data analysis. Following, the architecture and main functionalities of the AL environment are presented aiming to provide usability experts with necessary features which offer clear outcomes about strengths and weaknesses of the evaluated web systems. Finally, we describe a case study concerning the usability evaluation of two web sites that belongs to two competitive shipping companies.

2. Defining the process of comparative usability evaluation of web systems

Comparative usability studies belong to the category of summative evaluation studies in which the same group of test participants are asked to perform identical tasks on two or more systems. These types of studies heavily emphasise on creating an environment that provides the same experience for each test participant ensuring thus the validity of the study is protected.

Although, comparative usability evaluation studies can be performed by using expert, or analytical usability methodologies, the most common used usability methodology is the user testing methodology. User testing is widely recognized in the field of HCI as the most reliable way to examine the usability of a system [Woolrych et. al. (2001)]. The main benefit of the user testing method is that the evaluated system is tested under conditions close to those that exist when it is used “for real". While technical designers and human factors experts may diagnose a large proportion of potential system problems, experience has shown that working with users will reveal new insights that can affect the system design [Maguire et. al. (2001)]. According to user testing methodology, representative users are asked to perform a series of representative tasks with the evaluated system. The aim is to gather information about the users' performance with the system, their facial expressions and comments as they interact with the system, their post-test reactions and the evaluator's observations. In this context, users are encouraged to think aloud so that they can express their

personal opinions about how they perceive various features or controls of the system. This process is well-known as think aloud protocol and has an important role in such studies as it familiarizes usability experts with the users' mental model.

An important step in user testing methodologies is the gathering of information related to the accomplishment of specific tasks performed by the users. According to Maguire [Maguire et. al. (2001)], user interactions and comments can be recorded during each test session various ways.

- Automatic system monitoring may be set up whereby the system itself records interaction events of importance.
- An evaluator observes and manually records events during the interaction session including time to complete task, points of apparent user difficulty, user comments, errors made, number of times assistance is required, etc.
- A third method is to record each user session onto videotape for farther analysis.

Commonly the usability experts use a combination of the above three methods in order to achieve a complemented users' observation. The collected information includes quantitative data, such log files etc., that mainly derive from automated system monitoring but also qualitative data derived from questionnaires or audio and video recording. While quantitative data provide to usability experts information about user performance, the qualitative data provide them with valuable feedback about the user preferences [Nielsen et. al. (1994)]. Analysis of these valuable data helps analysts to identify usability flaws of evaluated systems.

In Figure1 a graphical representation of the comparative usability evaluation process that has been described can be seen. The participants use sequentially the competitive web systems in order to carry out specific scenarios. Their interaction during test sessions is videotaped and also logged for later analysis. So a big amount of observational data is produced.

All these structured (log files) and unstructured data (video, audio, hand-taken notes) need to be farther analyzed in order to cut off the critical points of interaction between user and system. Analysis is a tedious and cumbersome process and usually requires much time and many human resources. In this point, it becomes obvious that there is the need for tools that help usability experts to organize and analyze all the above data. Therefore, during the last years, there is a large growth of tools that support the analysis process.

Most of them utilize extensively multimedia data derived from usability studies, synchronizing them with textual files with hand-taken notes. This combination of sources seems to constitute a really strong dataset that is full of invaluable information as it includes peoples' activity and reactions.

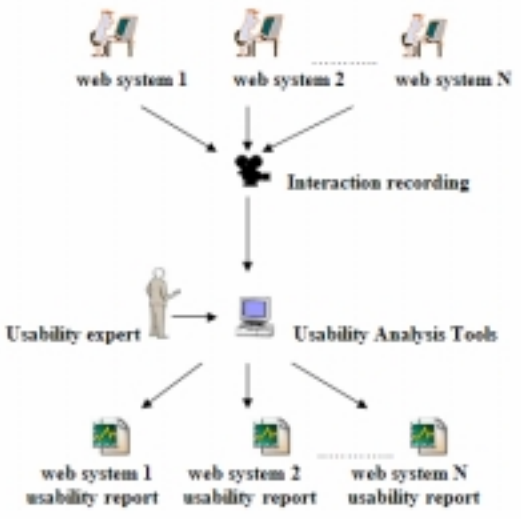

Figure 1. *The competitive usability evaluation process*

The Observer XT [Noldus (2006)], HyperResearch [ResearchWare (2006)], Transana [Transana (2006)], NVivo [QSR (2006)] are some of the most-known packages intended to provide analysis of collected data. Although all these packages provide tools to evaluate the usability based on collected data, they don't provide directly comparison of competitive systems. They just allow discrete analysis of data that come from usability studies.

The challenge is to create a powerful dataset and to provide flexible tools that give to the usability experts the ability to handle any useful information in a way that unveil lucid results about the ascendancy of a system against the competitors. In the next section we describe the AL tool that enables usability experts to compare competitive systems focusing on usability problems emerged from methodical data observation and analysis.

3. The ActivityLens tool

AL is an activity based usability analysis tool that embodies features especially designed for competitive usability analysis. Its main advantage is its ability to integrate multiple heterogeneous qualitative and quantitative data. These data can be video and audio files, log files, images and text files including hand-taken notes of the observers. AL permits integration and synchronization of the heterogeneous collected data.

The low level collected data derived from observation of users constitute an AL project. Projects are farther organized into studies. This way, a study is a collection of projects representing the interaction of several users with a certain web system. For

example a study can be the usability evaluation of search engines that contains 10 projects, concerning the observations of 10 users that interacted with these machines. The projects in a study can be farther organized into profiles which represent the competitive web systems which are under comparative usability evaluation. In the next sections we describe some features of the proposed tool that support the comparative evaluation of web sites.

3.1 Integration and synchronization of heterogeneous data

In Figure 2 the main window of AL can be seen. The main window of AL consists of three (3) main areas which are: area A which contains textual information of observed activity mainly derived from log files, area B which consists of the multimedia files which represent user-system interactions and finally area C which offers various filters through which a subset of the activity can be presented, related to specific actors, or specific types of events.

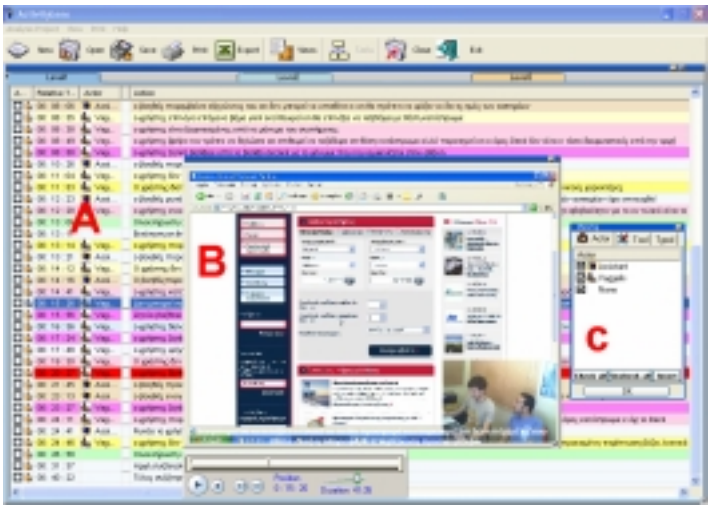

Figure 2. *The main window of ActivityLens*

Heterogeneous interaction data like server log files and video or audio files are fully interrelated through their time stamps and they are also precisely synchronized due to appropriate features of AL. Synchronization allows usability analyst to play back the observed activity by driving a video stream from the log file or to play back the activity through the help of video window controls.

Furthermore, AL provides the ability to describe in detail the observed activity, defining its most important phases. This is provided by a hierarchical structure that allows the aggregation of the events up to three levels of abstraction, taking into account the chronological sequence of the events. This is of high importance since it

permits researcher to distinguish the interesting observed tasks and consequently to focus on a specific type of actions or a repeated behavior that occurs during the task.

3.2 Definition of a usability assessment model

AL has the ability to incorporate various classification schemes that represent specific usability decomposition models. These models are hierarchically structured and decompose usability in a list of low level attributes. The process of defining usability attributes in these models, runs iteratively until the attributes can be observable and measurable in user - system interaction.

In AL tool, the collected data can be annotated order to highlight critical points of interaction between user and system. The annotation of the observed events in AL is based on the elements of the usability assessment model. So, AL provides the usability experts with the ability to correlate the observed behavior of users with specific usability metrics. In AL tool, each low usability attribute is called "Typology". Typologies categorize observable events in a way that make easy their handling and detection. For example a low usability attribute can be "Call for assistance". The usability expert, through AL, defines a new typology that is called "Call for assistance" and correlates this typology with events in which is obvious that user asks assistant to help him.

Based on the usability assessment model it is possible to focus on specific observed behaviors, thus reducing the huge amount of collected data by defining criteria. This ability is of high importance because it helps the usability expert to focus on the interesting sequences of events and make them emerge from the "noise". Filtering of events is possible according to subjects, tools and typologies or by forming any combination between them.

3.3 Supporting comparative usability analysis

The dynamic data model of AL allows the comparison between particular profiles. This comparison is accomplished through an analysis based on the usability assessment model. AL produces tables and graphs representing the average values of occurrences of low usability metrics for each one of the profiles. The window that describes the occurrences of a specific annotated behavior or usability flaw is shown in Figure 3 and it allows the researcher to zoom in the moment that they happened.

As we can see this window can describe the usability features of the evaluated web systems in a comparative way. The average values of each one of the low level usability attributes are shown in a table but they are also represented in dyads of bars. In case of two competitive web systems, the left bar represents the critical points that were found in study of interaction with the first system while the right one refer to second one. This window gives the usability expert the opportunity of a direct

comparison between the competitive systems, as she can easily see both their strengths and their weaknesses.

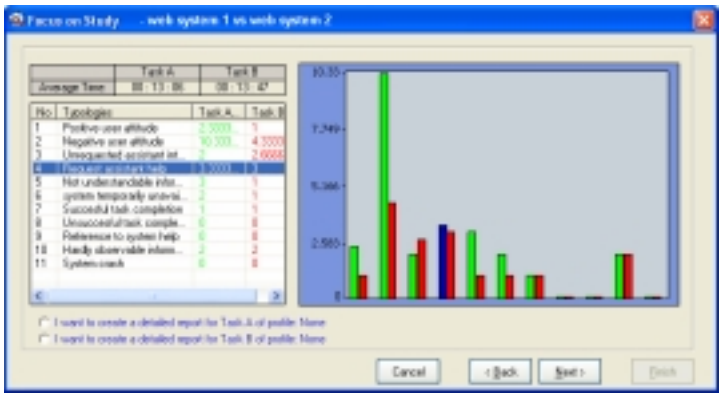

Figure 3. *Comparative usability evaluation features*

Through this representation, the usability expert has the ability to directly access specific types of interaction and to see how they influenced performance of users for all the evaluated web systems. Moreover, detailed comparative reports in form of HTML pages can be produced through AL including thorough description and attached instances of usability problems that users encountered in their interaction with evaluated systems.

4. Case study of comparative usability evaluation

4.1 Description of the case study

A pilot usability study was organized in order to examine how AL facilitates the analysis process in comparative usability studies. In the frame of this study, three usability tests were conducted in a controlled usability lab setting.

In this case study, three undergraduate students of Electrical and Computers Engineering Department of University of Patras, were the participants of this study. The participants had experience of e-commerce services since they used on line web systems for accomplishing similar tasks like the tasks they were asked to perform in this study. They were requested to seriatim use the electronic booking services of the two prevalent Greek shipping companies in order to book tickets to island of Crete. Each of the participants worked in conjunction with the same assistant.

The assistant encouraged the participants to think aloud as they were interacting with the web systems and unobtrusively kept notes in order not to influence user's behaviour. He just interfered the session when the user confronted a deadlock state or when he was asked to help the user about non understandable things of the system.

After the completion of the study, each one of the participants was interviewed about his opinion and experience about the systems design. All these sessions were recorded by a camera for farther analysis.

4.2 Comparative usability analysis

Initially, using AL, we created a new study including two profiles (each profile includes the collected data of every shipping company site evaluation). In order to have a common measure of the usability of the web systems we designated a usability analysis model according to which the collected data (log files, video, hand notes of the observer and interviews) were analyzed using AL. The used analysis model were derived from the usability definition according to ISO 9241-11 and were transformed into low level attributes in order to serve the usability evaluation of the two competitive Web systems.

So, we defined a decomposition usability model that analyses usability in low level metrics so that they can be observable and measurable from the collected data. According to this model, effectiveness is described as an attribute that depends on successful or unsuccessful task completion. Efficiency is described from time on task, occurrences of errors, frequency of assistances and failures of system. Satisfaction is described from positive and negative user attitudes as they were observed during their interaction with the system and as they were described in user interview. In the table1, our proposed usability assessment model is thoroughly described.

Table 1. *Description of Usability assessment model*

<table>
<tr><td rowspan="12">Usability (ISO 9241)</td><td rowspan="2">Effectiveness</td><td colspan="2">Successful Task completion</td></tr>
<tr><td colspan="2">Unsuccessful Task completion</td></tr>
<tr><td rowspan="10">Efficiency</td><td colspan="2">Task time</td></tr>
<tr><td rowspan="2">Errors</td><td>Not understandable information</td></tr>
<tr><td>Hardly observable information</td></tr>
<tr><td rowspan="3">Assistances</td><td>Requests Assistant help</td></tr>
<tr><td>Unrequested assistant intervention</td></tr>
<tr><td>Reference to system help</td></tr>
<tr><td rowspan="2">System failures</td><td>Not responding</td></tr>
<tr><td>System crash</td></tr>
<tr><td rowspan="2">Satisfaction</td><td colspan="2">Positive user attitude</td></tr>
<tr><td colspan="2">Negative user attitude</td></tr>
</table>

Using AL we defined as typologies the low level features of usability as they were described in the above model. We extensively studied the integrated multimedia files from both shipping companies and we annotated the most interesting situations found, correlating the observed user activity with low level features.

The window that describes the occurrences of a specific annotated usability flaw is shown in Figure 4. This certain window describes the usability evaluation results for both shipping companies in a competitive way. The average values of each of the defined low level usability attributes are shown in a table but they are also represented in dyads of bars. The left bar represents the critical points that were found in the study of the interaction with the system of shipping company1 while the right one refer to system of shipping company 2. This window gives to the usability experts the opportunity of a direct comparison between the competitive systems, as the low level usability attributes that measured in system 1 are contrasted with those of the system 2.

We systemically studied the evaluation results in order to acquire a clear opinion about each of the three main usability attributes for both systems. Especially effectiveness for both systems proved to be successful since all users achieved to complete the booking tickets task.

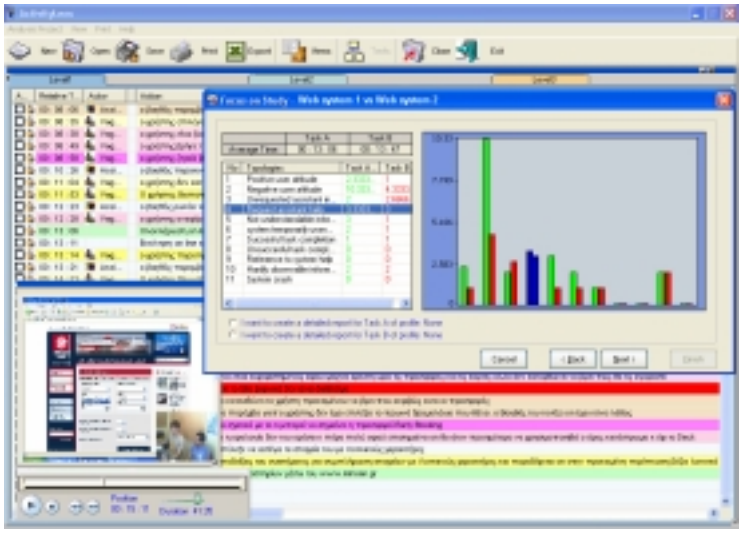

Figure 4. *Competitive representation of low level usability features*

Table 2 focuses on efficiency as it is described by low level metrics in Figure 4. As it can be easily seen the efficiency of system 2 has a clear preponderance against efficiency of system 1, since users of system 1 encountered more errors and needed more help than users of system 2. The main errors in system 1 were caused by strange terminology on its interface.

We derived information about user satisfaction based on interview of users after the task completion. The user attitudes are described in Table 3.

Based on our analysis using the above usability assessment model through AL, we managed to acquire an overall view about the usability of both systems. Although all

the users managed to execute the required tasks, both of the systems need a lot of improvements in order to help their users to be more efficient and satisfied when they use them. In system 1, most of the usability problems were related to the information architecture that lead the users to dead-end and caused the interferences of the assistants, while in system 2 most of the usability problems were related to navigability problems.

Table 2. *Low level features of efficiency*

	Usability features		System 1	System 2
Efficiency	Task time		13:06	13:47
	Errors	Not understandable information	3	1
		Hardly observable information	2	2
	Assistances	Request Assistant help	3.33	3
		Unrequested assistant intervention	3	2.66
		Reference to system help	0	0
	System failures	Not responding	2	1
		System crash	0	0

Table 3. *Low level features of user satisfaction*

	Usability features	System 1	System 2
Satisfaction	Positive user attitude	2.33	1
	Negative user attitude	10.33	4.33

5. Conclusions

Comparative usability evaluation provides valuable feedback regarding the usability features of systems which promise to fulfil same user requirements. In this paper we presented ActivityLens a tool designed for supporting the process of comparative usability evaluation performed by usability experts. We argue that for supporting this certain task of comparative usability evaluation various user requirements need to be supported like: a) easy integration and synchronization of heterogeneous data derived from the user observations like log files, hand notes, audio and video, b) the ability to allow usability experts to define custom usability assessment models or typologies according to which the web systems will be compared c) the analysis of the interaction data according to the defined usability assessment model and finally d) the presentation of the comparative usability evaluation results.

ActivityLens supports the aforementioned requirements by enabling directly comparative usability analysis of web systems. In order to examine the effectiveness of the tool in performing comparative usability evaluation studies, we conducted a case study concerning two competitive shipping companies. This study showed that using the proposed tool it is possible to perform comparative studies but underlined as

well some weaknesses in analyzing user questionnaires which is a matter of further development and research.

References

Avouris N., Fiotakis G., Kahrimanis G., Margaritis M., Komis V. (2007), *Beyond logging of fingertip actions: analysis of collaborative learning using multiple sources of data*, Journal of Interactive Learning Research JILR, vol. 18(2), pp.231-250, April, 2007

Brinck T., Gergle D., Wood S. (2001), *Usability for the Web: Designing Web Sites that Work*. Morgan Kaufmann, San Francisco.

Dix A., Finlay J., Abowd G., Beale, R. (1998), *Human-computer interaction*, Prentice Hall, Hemel Hempstead, UK

Hahn J., Kauffman R. J. (2006), *A Design Science Approach for Identifying Usability Problems in Web Sites that Support Internet-Based Selling*, (Working Paper).

ISO, 1998. Ergonomic requirements for office work with visual display terminals (VDTs)-Part 11: guidance on usability—Part 11: guidance on usability (ISO 9241-11:1998).

Johnson E.J., Bellman S., Lohse G.L. (2003), *Cognitive Lock-in and the Power Law of Practice*, Journal of Marketing (67:2), pp. 62-75.

Maguire M. (2001), *Methods to support human-centred design*. International Journal Human-Computer Studies, 55, 587- 634.

Molich R., Ede M. R., Kaasgaard K., Karyukin, B. (2004), *Comparative usability evaluation*, Behavior and Information Technology, 23(1), 65-74.

Nielsen J. (1993), *Usability Engineering*, Academic Press, London 1993.

Nielsen J., Levy J. (1994), *Measuring usability: Preference vs. performance*, Communications of the ACM, Vol. 37, no. 4, pp. 66-75.

NVivo 7, QSR International (2006), available at: http://www.qsrinternational.com/, accessed on December 2006

Observer XT. [6.0], Noldus Information Technology (2006), available at: http://www.noldus.com accessed on Decenber 2006

ResearchWare Inc (2006), website available at: http://www.researchware.com/hr/index.html, accessed on September 2006

Stoica A., Fiotakis G., Cabrera J. S., Frutos H. M., Avouris N., Dimitriadis, Y. (2005), *Usability evaluation of handheld devices: A case study for a museum application*. Proc PCI 2005, Volos, Greece.

Transana (2006), available at http://www.transana.org/, accessed on Decemberr 2006

Woolrych A., Cockton G . (2001). *Why and when five test users aren't enough*. In J. Vanderdonckt, A. Blandford, and A. Derycke (Eds.), Proceedings of IHM-HCI 2001 Conference, Vol. 2 (pp. 105-108). Toulouse, France: Cépadèus Éditions.