DEPTH: A Methodology for Usability Evaluation of Web Sites Based on Design Patterns and Heuristics Criteria

YANNIS PSAROMILIGKOS
Technological Education Institute of Piraeus
General Department of Mathematics
Computer Science Laboratory
250 Thivon & P. Ralli, 122 44 Piraeus, Greece
Tel. ++30 210 538 1193, Fax. ++30 210 538 1351
E-mail: gsa@teipir.gr

SYMEON RETALIS and KONSTANTINOS SIASSIAKOS
University of Piraeus
Department of Technology Education and Digital Systems
Karaoli & Dimitriou 80, 185 34 Piraeus, Greece
Tel: ++30 210 414 2746, Fax: ++30 210 414 2753
E-mail: retal@unipi.gr and siassiakos_k@ypan.gr

This paper presents DEPTH (usability evaluation based on DEsign PaTterns & Heuristics criteria) a methodology for performing scenario-based heuristic usability evaluation of Web sites. It is comprised of two operational phases, the preparatory and the execution phase, where specific steps are performed for evaluating the usability along three axis: usability comparison tests among similar Web sites, expert/heuristic evaluation, scenario-based user inspection/enquiry sessions. DEPTH’s innovation lies on the use of usability design patterns both as a mean for identifying best practices of usability design as well as for specifying the scenarios and tasks for user inspection. It is also a method for capturing, adapting and refining usability resources of a Web site. DEPTH is also supported by a CASE tool whose functionality is described. Finally, we present a small example demonstrating the basic steps of DEPTH.

Categories and Subject Descriptors: Additional Key Words and Phrases: Usability evaluation, evaluation methodology, DEPTH, scenario-based heuristic evaluation, usability design patterns, user inspection

1. INTRODUCTION

Nowadays, usability is one of the quality factors defined in the ISO 9126 standard [ISO/IEC 9126, 1991] that plays a key role in software development. Especially at the rapidly growing arena of Web applications there is a growing need for systematic approaches to their development and delivered quality [S. Murugesan and Y. Deshpande, 2001]. The range of Web applications as for example the e-commerce market, the e-learning market, the virtual museums, etc. has exploded over the last three years. It is also projected to continue its rapid expansion. Although in the early years of web-sites, developers mainly aimed in designing them the quickest possible in order to achieve an early presence in the cyberspace, nowadays usability matters. Websites’s user interfaces are now redesigned, taking into account the user needs. Web-sites have also be transformed into dynamic systems the functionality of which has became more complex.

Usability evaluations of dynamic web sites require a lot of effort, since they involve usability experts, raising the cost too high for small and medium sized organisations. A
solution to this problem would be to force developers learn how to design web sites following usability guidelines and patterns.

Formal usability of web sites have attracted the attention of many research and development groups in the last few years. In the literature one can find three main dimensions particularly applicable to e-web sites usability evaluation:

- Usability comparison tests among similar e-commerce sites.
- Expert/heuristic evaluation.
- Scenario-based user inspection/enquiry sessions.

Scenarios provide a versatile and reproducible means of evaluating a system. A scenario is an instantiation of one or more representative work tasks and transitions linking those tasks [M. Rosson et al., 2001]. The granularity of the scenario is not fixed; a scenario can be highly scripted or loosely defined. One of the main difficulties is how to create such scenarios.

In this paper we present the DEPTH approach, an innovative approach for usability evaluation of Web sites based on design patterns and heuristics criteria. This approach tackles all three of the aforementioned dimensions of usability evaluation. Moreover, DEPTH prescribes how to perform the steps of the evaluation giving emphasis on how to compare the e-commerce sites as well as how to easily create scenarios for user inspection. The main aid in this prescription is the usage of design patterns. Design patterns describe a problem and a solution for this problem in a particular context, together with the rationale for using that solution and the consequences (pros and cons) of using it [E. Gamma et al., 1994]. "The pattern is, in short, at the same time a thing, which happens in the world, and the rule which tells us how to create that thing, and when we must create it. It is both a process and a thing; both a description of a thing which is alive, and a description of the process which will generate that thing" [C. Alexander et al., 1977]. The structure of the paper is the following: Section 2 gives an overview of the DEPTH approach while section 3 illustrates an example of its application. Section 4 presents a prototype tool supporting the methodology. The paper concludes with an evaluation report of applying DEPTH at the evaluation of a learning management system as well as our future plans.

2. THE DEPTH APPROACH

As in other evaluation methods (e.g. SUE method and its ancestor called MILE [F. Garzotto, M. Matera, P. Paolini, 1998]) we require two operational phases of evaluation: the preparatory phase and the execution phase. The preparatory phase aims to define the conceptual framework of the evaluation through an iterative process consisting of the following four distinct steps:

1. Select all design patterns related to domain under evaluation and decide for the patterns that will be used in the evaluation.
2. For each design pattern generate a list of usage scenarios according to the underlying functionality of the pattern.
3. For each scenario generate appropriate tasks for user inspection.
4. Apply appropriate usability criteria (heuristics) to scenarios/tasks and generate a toolkit (a set of questionnaires) for quantitative and qualitative evaluation.

The deliverable of the preparatory phase is a set of questionnaires (toolkit) each one for a design pattern, which they will be used at the execution phase. The evaluator may follow
an iterative process among the steps of the phase in order to refine and best fit the underlying usability criteria both in the scenarios/tasks and the toolkit.

The execution phase utilizes the toolkit derived from the preparatory phase in order to perform the actual evaluation. The mathematical approach that will be used for calculations should be defined along with the necessary adjustments. As in [L. Ohnita, 1999] we propose the Logic Scoring of Preference (LSP) model and continuous preference logic as mathematical background especially in cases where the evaluation process includes many design patterns with large amount of elementary attributes and complex relationships among them. So the phase consists of the following steps:

1. Define the mathematical approach of the evaluation and perform the necessary adjustments.
2. Execute the inspection and gather the results.
3. Perform the calculations according to the underlying mathematical approach.
4. Analysis of data gathered and presentation of results.

The ultimate aim of both phases is to support the measurement of the usability of a Web site by examining three dimensions: i) usability comparison tests among similar e-commerce sites, ii) expert/heuristic evaluation and iii) scenario-based user inspection/enquiry sessions.

2.1 The preparatory phase

Most of the times, the preparatory phase is entirely dependent on the experience of the usability expert/engineer. In order to systematize the whole process and provide an effective way for identifying a) the comparative matrix, and b) the main scenarios for user inspection, we propose the utilization of design patterns. According to [C. Alexander et. al, 1977]: “each pattern describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over”. A “problem” is normally a specific functional or non-functional requirement from the system. Because we are interested in solutions that improve the usability of the system in use, we focus on customer-oriented problems and not developer-oriented. Aid in the “pattern mining” approach is the work on hypermedia patterns [F. Garzotto, M. Matera, P. Paolini, 1998], [M. Rosson et. al, 2001], as well as the HCI patterns [Hypermedia Design Patterns Repository, 2002], [J. Borchers, 2001]. Before we proceed to the next step, each design pattern could be categorized as in [A. Stefani, M. Xenos, 2001] in high, middle, and low category. High level comprises of those characteristics of Web sites under evaluation that are most important. Middle level consists of those characteristics that are related to the services provided, but are not as important as those of the high level. Finally, low level includes the least important characteristics.

After deciding for the patterns that will be used in the evaluation, the next step is to generate a list of usage scenarios for each pattern. The list does not include complete scenarios but small descriptions expressing the target-points of each scenario. The scenarios are complemented at the next step where appropriate tasks are added for user inspection. The target-points of scenarios can be easily derived according to the required functionality proposed by the underlying design pattern. We could characterize such a list as “good practices” on how to implement the underlying functionality, which then will be used a) for comparing a specific Web site to an “ideal” one, and b) for evaluation during the inspection process.

However, the existence of required functionality is not enough to make the Web site usable since how each design pattern is implemented will make a difference in usability.
Thus, appropriate tasks should be easily generated to measure the usability of the site. These tasks will guide the user inspection. For each task (and sometimes for a group of tasks) we identify the main heuristic evaluation criteria that should be measured.

The final step of the preparatory phase is the generation of the toolkit, basically a set of questionnaires (one questionnaire for each design pattern) that will provide data for quantitative and qualitative evaluation. Such a questionnaire should consist of two sections: a specific and a general section. The specific section is comprised of questions that measure the existence as well as the user satisfaction, ease of use, and/or usefulness of all the functionality points identified by the list of tasks included in the previous step. The questions are grouped according to the task they belong while an appropriate (small) scenario is preceding in order to guide the user to answer the questions. The general section consists of questions that measure general aspects of each design pattern according to Nielsen’s heuristic evaluation criteria.

2.2 The execution phase

At the execution phase we perform the user inspection of Web sites. The underlying scenarios incorporated into the toolkits developed at the preparatory phase guide user inspection. With the execution phase we compare the completeness of the functionality of a Web site under examination against an “ideal” site, which contains the full range of functionality of a site as described by a design pattern. Moreover, we measure the usability of the functionality offered. Thus, we do not only care about what the site offers but how well it does offer it.

The execution phase may have many instances called “evaluation sessions”. At each evaluation session a predefined group of people (user/expert) examine one or more Web sites under a common mathematical approach. An evaluation session may reuse toolkits generated from instances of previous (closed) preparatory phases. This is possible since each evaluation is based on specific design patterns, which correspond to specific toolkits. If another evaluation uses the same design pattern the underlying toolkit could be reused.

When the evaluation problem is rather simple, for example we examine a single design pattern e.g. the “Shopping Card”, we may use a simple mathematical approach to calculate and analyze the data gathered. However, in complex Web sites where we may examine many related design patterns, the preparatory phase may generate a large amount of quality attributes with complex logic relationships among them. In such cases the evaluation process becomes a complex issue and it is difficult the identification of minor differences between similar competitive Web sites [L. Olsina, 1999]. Therefore we propose the Logic Scoring of Preference (LSP) model and continuous preference logic as mathematical background for the calculation of partial and global quality preferences of sites [J. Dujmovic, 1996]. The same approach has been proposed by Web-QEM [L. Olsina, 1999] for the evaluation of academic web sites.

3. AN EXAMPLE

In this section we give a small example to demonstrate the steps of our approach. Suppose we want to evaluate some e-commerce sites with main emphasis on checking the usability of the online purchase of a product. Starting the preparatory phase we should firstly select all design patterns related to domain under evaluation and decide for the patterns that will be used in the evaluation. We can gather a series of design patterns such as:
• Shopping Cart/Basket: Allows users to gather all products first and pay for them all at once and whenever they want.
• Advising: Help the user find a product in the store, assist him according to his wishes.
• Opportunistic Linking: Keep the user interested in the site. Seduce him to navigate in the site even when he has already found what he was looking for.
• Explicit Process: Help the user understand the buying process when it is not atomic.

We identify the “Shopping Cart/Basket” as the most essential capability of an e-commerce site according to our objective and so we will use this pattern in the following evaluation steps. The description of the design pattern “Shopping Cart/Basket” can be found at [M. Welie, 2002].

By examining the underlying pattern we can easily identify a number of target scenarios derived from “good practices” on how to implement the proposed functionality. The total number of the identified target scenarios was 19 as shown in table 1.

Table 1. The list of target scenarios

<table>
<thead>
<tr>
<th>1. Shopping cart basket</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. Appropriate name</td>
</tr>
<tr>
<td>1.2. Ability to add items from anywhere</td>
</tr>
<tr>
<td>1.3. Any type of item can be included</td>
</tr>
<tr>
<td>1.4. Contents viewable at any time</td>
</tr>
<tr>
<td>1.5. Properties “Description”, “Qty”, “Price”, “Availability”, “Category” defined for each line item</td>
</tr>
<tr>
<td>1.6. Additional properties defined describe each line item appropriately</td>
</tr>
<tr>
<td>1.7. Delete Line Item</td>
</tr>
<tr>
<td>1.8. Modify quantity</td>
</tr>
<tr>
<td>1.9. Link to detailed description</td>
</tr>
<tr>
<td>1.10. Total costs calculated according to changes performed</td>
</tr>
<tr>
<td>1.11. Help customers proceed with order</td>
</tr>
<tr>
<td>1.12. Provision of label next to the shopping basket image</td>
</tr>
<tr>
<td>1.13. Links related with shipping and handling costs and their calculation</td>
</tr>
<tr>
<td>1.14. Links for applicable taxes</td>
</tr>
<tr>
<td>1.15. Link for return policy</td>
</tr>
<tr>
<td>1.16. Validation within shopping basket contents</td>
</tr>
<tr>
<td>1.17. Shopping cart saved period</td>
</tr>
<tr>
<td>1.18. Cross selling</td>
</tr>
<tr>
<td>1.19. Upselling</td>
</tr>
</tbody>
</table>

Now, we can complement each scenario with appropriate tasks to guide the user inspection. These tasks are then integrated into questionnaire along with the necessary questions that measure the existence as well as the user satisfaction/easy of use/usefulness of each criterion/characteristic defined in table 1. Table 2 shows an example of how a small scenario is integrated into questionnaire.

Table 2. Scenarios, tasks and questionnaire generation.

<table>
<thead>
<tr>
<th>Task</th>
<th>1.7.1: Delete Line Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1.7.1.1:</td>
<td>Does the system perform this task?</td>
</tr>
<tr>
<td>Question 1.7.1.2:</td>
<td>Did you find easy to perform this task?</td>
</tr>
<tr>
<td>Question 1.7.1.3:</td>
<td>Are you satisfied with the task implementation?</td>
</tr>
</tbody>
</table>

At the execution phase we should decide upon the mathematical approach we will use at the evaluation. Because of the simplicity of our case study we use simple statistical calculations (e.g. averages). We assume each task and each scenario to be equal weighted, we also define the mark scale of each question and perform the calculations.
4. A PROTOTYPE SYSTEM

We have implemented a prototype Web-based tool for designing and implementing evaluations based on DEPTH methodology (http://softlab.teipir.gr/depth_toolkit). The tool supports both preparatory and execution phase. More specifically, the tool supports the following tasks:

- recording, browsing and searching for design patterns
- assigning scenarios and appropriate tasks to each design pattern
- assigning questions to tasks and generating the questionnaire for the evaluation
- managing evaluation sessions and recording the results
- generating different evaluation toolkits (questionnaires) based on the same pattern (cloning)
- authorized and personalized access for the "Designers", "Session managers" and "Evaluators".

Figure 1 shows how the system manages the various design patterns, which are used in the evaluations.

![Figure 1. Registration of Design patterns.](image)

The tool relates each preparatory phase to one design pattern and generates one questionnaire. The "Designer" of an evaluation should declare the name of a registered (to the system) design pattern when he opens a new preparatory phase. If the specified pattern has been used in other (closed) preparatory phases, it is already associated with corresponding scenarios and tasks. The designer has the option to generate a completely new version of scenarios and tasks or to modify a clone of a previous one (Figure 2).
The “Session Manager” establishes and manages an evaluation session. Each evaluation session is related to:

- a set of Web sites that will be evaluated
- a set of users (evaluators) that will perform the inspections, and
- a set of design patterns that will be used in the evaluation.

Because each design pattern may have many preparatory phases, the session manager should specify the identifiers of these specific preparatory phases in order the system to be able to associate the corresponding questionnaires for the evaluation (Figure 3).
For the present, the system performs simple calculations based on marks and weighting factors defined at question level (Figure 4).

5. CONCLUSIONS

We have applied DEPTH methodology in two separate pilot studies. Firstly, we tried it out for the evaluation of four (4) e-bookstores, namely the www.amazon.com, www.walmart.com, www.ianos.gr, and www.plaisio.gr. Two (2) usability experts took part in this evaluation study. We also tried it out for the evaluation of four (4) e-learning brokerage platforms, such as the Universal, MERLOT, WorldLectureHall, and COREO. Six (6) teams comprised of 3 junior software engineers with interests in Learning technologies participated to the evaluation of these systems.

In the aforementioned pilot studies, we were primarily interested in investigating whether the main idea behind DEPTH, i.e. the use of design patterns, is understandable and adds value to the usability evaluation. We also wanted to examine how easy the users would manage to follow the steps of the DEPTH method.

The results were more than encouraging. Table 3 shows the 8 design patterns used from students.

Table 3. Design patterns used at the evaluation.

<table>
<thead>
<tr>
<th>Design patterns used at the evaluation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Support</td>
</tr>
<tr>
<td>Simple Search</td>
</tr>
<tr>
<td>Advanced Search</td>
</tr>
<tr>
<td>Site Map</td>
</tr>
<tr>
<td>Registration-authentication-access control</td>
</tr>
<tr>
<td>Course creation and customization</td>
</tr>
<tr>
<td>Home Page</td>
</tr>
<tr>
<td>Map Navigator</td>
</tr>
</tbody>
</table>

Some of the comments made are the following:
DEPTH’s Advantages
- Minimises the time and effort at the preparatory phase
- It guides the whole evaluation process
- Novice usability evaluators can measure the usability using this method
- Isolated areas of interest can be evaluated (e.g., checkout process)

DEPTH’s Disadvantages
- Design patterns are not that many. Will there always be a design pattern to validate all areas of interest in a website?
- Increased number of points to be evaluated

Concluding, the DEPTH approach is valuable for examining the completeness of the functionality of the website under evaluation, illustrates the use of scenarios in performing user inspection, and identifies easy-to-measure correlates of more important, but complex, behaviours. Third parties have not extensively applied this approach as yet. Thus we cannot provide any evidence of the efficiency and the effectiveness of DEPTH.

We plan to organise systematic user trials of this approach as well as to start experimenting with applying DEPTH for evaluating the usability of other types of hypermedia systems like Learning Management Systems and adaptive educational hypermedia systems.

REFERENCES