Journal of Visual Languages and Computing 28 (2015) 1–22

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Empirical validation of an automatic usability evaluation method ☆

Rosanna Cassino, Maurizio Tucci, Giuliana Vitiello *, Rita Francese

Department of Management and Information Technology, Università di Salerno, Italy

ARTICLE INFO

Article history:
Received 11 October 2014
Received in revised form
2 December 2014
Accepted 3 December 2014
Available online 13 December 2014

Keywords:
Interactive systems evaluation
Usability engineering
Summative usability evaluation

ABSTRACT

Today, the success of a software application strongly depends on the usability of its interface, so the evaluation of interfaces has become a crucial aspect of software engineering. It is recognized that automatic tools for graphical user interface evaluation may greatly reduce the costs of traditional activities performed during expert evaluation or user testing in order to estimate the success probability of an application. However, automatic methods need to be empirically validated in order to prove their effectiveness with respect to the attributes they are supposed to evaluate.

In this work, we empirically validate a usability evaluation method conceived to assess consistency aspects of a GUI with no need to analyze the back-end. We demonstrate the validity of the approach by means of a comparative experimental study, where four web sites and a stand-alone interactive application are analyzed and the results compared to those of a human-based usability evaluation. The analysis of the results and the statistical correlation between the tool's rating and humans' average ratings show that the proposed methodology can indeed be a useful complement to standard techniques of usability evaluation.

1. Introduction

The most intuitive definition of usability is the property of the system that defines its degree of simplicity of use in terms of learning, storage and efficiency. The ISO 9241 standard, on "Ergonomics of Human System Interaction", defines usability as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."

From 1980 the value of a software system is measured on the basis of its GUIs (graphical user interfaces) and the related power of expression and communication. The interface has to be user-friendly because it is often the only part of the system with which the user interacts [20].

To create a "usable" system, the designer must create a good conceptual model of the application (correct, consistent, and simple) and effectively transmit it to the user through the interface that must "accommodate" user's mental model, namely her expectations about system behavior. Several evaluation methods and tools are available to measure to what extent a GUI is "usable".

Usability inspection methods involve usability experts and different techniques (cognitive walkthrough with task-specific, heuristic evaluation, and pluralistic walkthrough) to evaluate a user interface without involving end users. These approaches, generally, can be used early during the development process by evaluating system prototypes or specifications that cannot be tested on end users [18].

Empirical methods refer to usability testing used in user-centered interaction design to evaluate a product by

☆ This paper has been recommended for acceptance by Shi Kho Chang.

* Corresponding author.
E-mail addresses: rcassino@unisa.it (R. Cassino), mtucci@unisa.it (M. Tucci), gvitiello@unisa.it (G. Vitiello), francese@unisa.it (R. Francese).

http://dx.doi.org/10.1016/j.jvlc.2014.12.002
1045-926X/

testing it on users. This can be seen as an irreplaceable usability practice, since it gives direct input on how real users use the system [4].

Such approaches detect usability deficiencies of the graphic interface by running and inspecting test cases and/or by analyzing the results of questionnaires or interviews. In addition to being very expensive and laborious, they often produce results that can be biased by the acquisition method and considerably depend on the adopted definition of usability, on the type and number of tasks, on the data and on the evaluation standards.

On the contrary, the automatic evaluation techniques are designed to avoid those problems both in terms of cost and in terms of running time: an automatic tool is able to locate in minutes (if not seconds) many critical issues. To get the same results with the above methods would take many hours of interviews and simulations of use cases. A comparative study carried out in the field confirms that automated tools for analytical evaluation are very efficient in terms of execution time, objectivity and reliability of the results obtained [11]. They can help designers to understand what usability problems may arise and how the interface should be improved, with respect to a given set of guidelines. However, most existing methods rely on interface source code to discover usability problems and offer advice on how to fix them.

In [1] we presented USherlock, a tool of GUI usability evaluation based on a front-side approach to derive the structure of the interface starting from what the user sees on the display. It allows to determine the nature of the elements of a graphical interface according to the changes produced on the interface itself. Depending on the type of interaction (input/output; pause; double click; click and double click; insertion of a character) and in case a visual feedback is found, the tool identifies some typologies of "dynamic" elements: button, link, text area, etc., and adds a new node to a tree which represents the hierarchical structure of each frame of the interface. For each element or set of elements classified the evaluation process will run all the usability controls. At the end of the evaluation process, each node is assigned a list of the inconsistencies identified and a score (rating between 0 and 1), which indicates the "quality" of the node.

In this paper, we present the results of the empirical study we have conducted to validate the proposed automatic usability evaluation technique with respect to its effectiveness, compared to the results of a traditional test, as it is usually performed during summative evaluation. We wanted to verify whether USherlock could be reliably adopted to evaluate usability of given artefacts with an acceptable error coverage. We decided to compare its performance with the outcomes of a heuristic evaluation performed by evaluators with a medium level of expertise. Therefore, the participants we recruited for our experiment were 24 students from the graduate course of Computer Science, who had successfully passed the exam of HCI and Software Usability and were well-trained on heuristic evaluation techniques. We analyzed four typologies of web sites and a stand-alone application. A control group of four expert evaluators (two external independent evaluators and two of the co-authors) was in charge of providing data on the real usability issues characterizing the analyzed artefacts. The effectiveness of each usability evaluation method could then be measured as the ratio between the number of detected problems and the number of real problems (as specified by the control group).

The results of the empirical study show that a statistically significant improvement in terms of effectiveness is indeed achieved using USherlock with respect to heuristic evaluation performed by evaluators having a medium level of expertise, with obvious advantages in terms of time and cost with respect to the canonical manual tests. To show that the improvements are unlikely to have occurred by chance, we have applied a one-tail t-test with a significance p-value < 0.05 on the collected results for the pair of usability evaluation methods.

A major lesson learnt from this research is that the actual adoption by practitioners of an automatic usability evaluation technique can be greatly encouraged if a rigorous empirical analysis is performed to prove its reliability in terms of problems discovered. The automatic detection of specific usability issues, with the derived advantages in terms of time and cost, can then be effectively combined with other forms of usability evaluation, such as user testing and heuristic evaluation, meant to measure other aspects of usability especially related to users' degree of satisfaction.

The rest of the paper is organized as follows. Section 2 discusses about some related work. Section 3 summarizes the architecture of the implemented tool USherlock and describes some algorithmic aspects. In Section 4 we describe the empirical evaluation process developed to validate the tool analysis results. Section 5 contains some conclusions and final remarks.

2. Related work

The field of usability evaluation has been widely investigated over the last years. Several methods have been proposed both for the evaluation of web applications and of interactive graphical applications. However, many of such methods have failed to meet practitioners' expectations with respect to their usability evaluation goal, resulting in a low adoption rate. The need has therefore emerged among HCI researchers to perform appropriate studies meant to evaluate the usefulness of a usability evaluation method and the benefits gained by its adoption. The importance of rigorous studies to validate usability evaluation methods was first claimed by Gray and Salzman in [6]. The authors reviewed several experimental studies performed on usability evaluation methods and concluded that most suffer from the lack of meticulousness in proving the statistical validity of the achieved results. Similar claims were later discussed by Hartson et al., who also addressed the lack of standard criteria and a clear understanding of the factors being measured as major problems when comparing different usability evaluation methods [8]. A few years later Hornbæk summarized current practice of measuring usability and critically reviewed that practice, after reviewing usability measures employed in 180 studies published in core HCI journals and proceedings [9]. The goal of their study was to understand what