Usability Testing Tools for Web Graphical Interfaces

Carlos Teixeira, Bernardo Santos and Ana Respício[1]
Department of Informatics, University of Lisbon 1749-016 Lisboa, Portugal
[1]Operations Research Center, University of Lisbon 1749-016 Lisboa, Portugal
E-mail: cjteixeira@fc.ul.pt

Keywords: usability testing tools, user-centered interaction design, graphical web, multi-criteria decision analysis

Received: February 16, 2013

Software design and development following a user-centered approach can benefit from the adoption of adequate usability testing tools. However, the choice of a suitable tool for a particular purpose can be a difficult task, due to the multiplicity of such tools, each one offering a variety of different features. This paper surveys usability testing tools for web graphical interfaces, selects a set of appropriate tools and evaluates them. A set of relevant evaluation features is identified and aggregated into criteria. A multi-criteria additive utility function and the Analytical Hierarchy Process are proposed as evaluation methods and for establishing a ranking of a selected set of usability testing tools. Results of both methods are presented and compared.

Povzetek: Prispevek predstavlja pregled orodij za spletne grafične vmesnike.

1 Introduction

The user-centered design process relies on the involvement of users in every dimension that could be related to the success of the product. As human issues are always a main source of complexity for engineering, the size and heterogeneity of designers' team is often a requirement and another source of problems in itself. In order to overcome this small additional source of complexity, designers should cooperate according to some common guidelines built on their experience and a vast literature of recommendations, in a productive way that should provide convergence of results toward the final product (Norman, 2002).

Long lasting design teams have their own stabilized strategies, tactics and tools, partly established on the acquired experience with previous projects. New teams or teams with several new collaborators can take extra benefits from commercial off-the-shelf, well documented frameworks of integrated computer tools. When it concerns user-centered design of web interfaces, advanced prototypes, the final product and the users, can be directly accessed by robust common frameworks. These frameworks are repeatedly used, project after project, by the same teams. Even if teams are often remixed in their composition, a reliable framework, well understood by all the personal, will decrease the distance in the gulf that separates the evaluation protocols and the corresponding collected data from the team intuition about the problems and the innovations for their solutions.

Evaluation of a product relying on users tests (usability testing) is an irreplaceable technique in user-centered design (Shneiderman, 1998; Nielsen, 1993), since it gives direct input on how real users interact with the system (Nielsen, 1993).

There are many usability testing tools (UTTs) available nowadays, with different features and capacities. This paper is an attempt to organize the concerned information and choose a suitable usability testing tool for web interfaces (Nielsen, 1999; Dix et al., 2003), with particular emphasis on graphical interaction.

The evaluated UTT issues and features and the corresponding preferences were established by a restricted number of experts with the aim of conveying the usability tests of interfaces designed for prototypes developed by the World Search Project (World Search Project, 2010). This is a Portuguese project of almost 2 million euros investment which is responsible for the design of search interfaces for dedicated areas of public concern, namely in the health area. The goal of the World Search Project is the research and development of innovative web search technologies in Portugal as well as the research and development of generic and business information with semantic relevance and with the proper knowledge of the Portuguese language, culture and market.

The second section presents the issues and features considered for evaluation and comparison of UTTs. The third section surveys usability testing tools and presents the selected set of UTTs. The evaluation methods adopted are described in Section 4. The fifth section presents and discusses the results obtained insofar. Final section presents conclusions and some directions of future work.

2 Main issues and features for UTT evaluation

Many issues and features are relevant for building a comprehensive usability testing tool. Figure 1 is a tentative graphical representation of the main issues considered. These were represented as a flow as close as possible from the temporal order where designer's plans must be implemented.

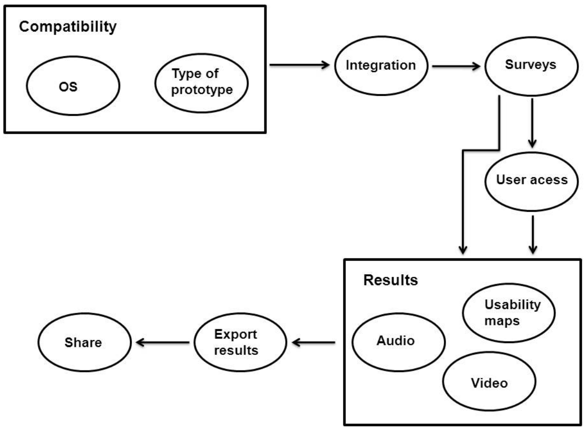

Figure 1: Usability testing tools issues.

When adopting any new software, anyone first concern will go into compatibility issues such as OS compatibility. Specifically for the design process, it is important to integrate several types of possible prototypes, giving a wide space of freedom to the designers, while facilitating several options of integration with all kind of surveys, questionnaires and alerts. A flexible integration can promote high quality testing. For instance, integrating the tests within the application (ex: using Javascript) can increase the dynamics of the usability tests as well as the quality of possible tests when compared to submitting screenshots to the UTT. Another relevant issue is the type of surveys produced by the UTT, the extent and the kind of questions allowed in the surveys that will be used to produce results. Our aim is to perform usability tests, having access to users located across the country or even abroad. Thus, user access is also a main issue to be considered. Concerning collecting results, three types of input are relevant: usability maps, which contribute to the analysis of users' interaction with the application; video recording, that is fundamental for tracing users' actions in the display and simultaneously recording facial expressions while interacting; and, audio recording, for collecting voice information produced by the user along with the interaction and consequently producing annotations (essential for the think-aloud protocol). As our goal is to evaluate interfaces with graphical interaction, a higher importance is given to features concerning collecting video from display, as well as generating usability maps including clicks and mouse movements. Finally, it is aimed that the format used to export the results is adequate for the subsequent analysis. Features concerning results' formats are aggregated by the issue "Export results", which also includes features related with the possibilities of sharing results ("Share") with the developers and designers teams (project partners). The survey of Vraa (Vraa, 2009) identifies important features and functionalities relevant for UTT evaluation. Many of these were also considered in present contribution.

To summarize, the following lines enumerate main issues (criteria) considered and the features (sub-criteria) within each of them:

1. OS Compatibility: Windows; Linux; Mac OS.
2. Supported types of prototypes: Applications; Prototypes; Screenshots of the interface; Wireframes; Mock-up's.
3. Interface integration with the UTT: Offline program (off-line test generation and managing); Website post (the URL to be tested is submitted to the UTT website); Uploaded images (screenshots submission); JavaScript code (that forwards information to an on-line account of the UTT website); Online wizard (all details of the interface; associated tasks are submitted to the UTT website in a pre-specified order).
4. User access (to the usability tests): Local; Remote; On-line.
5. Creation and submission of surveys and tasks for the users: Complete survey; Screen aligned questions (kind of pop-up with questions during specific passages of the usability test); Screen aligned text (kind of pop-up with questions during specific passages of the usability test);
6. Collecting audio: Record (both user and wizard-of-Oz /prototypes/ etc.); Annotations.
7. Collecting video: Display; Facial Expressions; Eye Tracking; Annotations.
8. Usability maps: Clicks; Mouse move; Scroll reach; Attractive zones; Interest zones; Attention zones; Form inputs.
9. Export: XLS/CSV/TSV; XML; Database; Share (online access management to results for the development team).

3 Selected UTT

This section describes the process of selecting the UTT candidates for the present study, which was inspired by several interesting web articles starting with Vraa (Vraa, 2009), Fadeyev (Fadeyev, 2009) and Tomlin (Tomlin, 2009). In the following years related articles were also published on-line by Walker (Walker, 2010), Gube (Gube, 2011), Jules (Jules, 2011) and LeMerle (LeMerle, 2012).

Table 1 displays in the first row our list of 23 candidates and the considered UTT reviews in the first column. Each UTT discussed by a given review is highlighted with an 'x' mark in the corresponding cell.

The list of candidates elected for evaluation was mainly based on the review of Tomlin (Tomlin, 2009) that extensively describes UTT in terms of features, presenting several plans of prices. Some of the Tomlin UTTs are not included in our candidates. The *Clixpy* and *Simple Mouse Track* websites were not found. The *Google Website Optimizer* and the *UserVue* were merged

UTT review	*Concept Feedback*	*Chalkmark*	*ClickHeat*	*ClickTale*	*Crazyegg*	*Ethnio*	*Feng-GUI*	*Fivesecondtest + NavFlow + ClickTest*	*Feedback Army*	*Loop 11*	*Mechanical Turk*	*Morae (include User Vue)*	*Open Hallway*	*Silverback*	*Usabilla*	*Userfly*	*User Testing*	*Google Analytics (WebSiteOptimizer)*	*Intuition HQ*	*4Q Survey*	*Mouse Flow*	*Attention Wizzard*	*Click density*
(Tomlin, 2009)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x					
(Vraa, 2009)										x						x							
(Fadeyev, 2009)		x	x	x				x	x					x		x	x	x					
(Walker, 2010)	x			x	x	x	x	x	x			x	x	x	x	x		x					
(Gube, 2011)		x						x	x	x					x				x				
(Jules, 2011)	x	x	x	x	x		x	x	x	x		x		x	x	x	x			x	x	x	x
(LeMerle, 2012)		x	x	x				x	x					x		x	x						

Table 1: Usability testing tools reviews and candidates selected for evaluation.

into *Google Analytics* and the *Morae*, respectively. The *Website Grader* was conceived in order to enhance online marketing websites, which is not within the scope defined in this paper introduction. *Fivesecondtest* is now available with two complementary applications *NavFlow* and *ClickTest*, which can be seen as a single UTT (the *UsabilityHub* from *Angry Monkeys*).

Vraa (Vraa, 2009) presents and discusses the best "Do's and Don'ts for Web Design and Usability" naming "16 crucial web design and usability best practice compilations and tools".

Though Vraa only reviews two UTT, the extended discussion on crucial UTT features inspired us in the identification of evaluation criteria and relevant features.

In the same year, Fadeyev (Fadeyev, 2009) surveys ten affordable UTT, claiming that "testing for usability is the only reliable way to find out how well a website works". Walker (Walker. 2010) also describes some of the already reviewed UTT and added a few more, whose main goals were to improve the visibility of websites for marketing purposes and thus were not included in our list of candidates. Gube (Gube, 2011) reviews the "22 essential tools for testing your website's usability" by classifying them into six categories.

1. User Task Analysis: *Intuition HQ*, *Usabilla*, *Loop11* and *Fivesecondtest*.
2. Readability: "*Juicy Studio: Readability Test*", *WordsCount* and *Check My Colours*.
3. Site Navigability: *Websort.net*, *OptimalSort*, *Chalkmark*, *WriteMaps*, *NavFlow* and *PlainFrame*;
4. Accessibility: "*Juicy Studio: Local Tools*", *VisCheck*, *W3C Markup Validation Service*, *WebAnywhere* and *Browsershots*.
5. Website Speed: *Pingdom Tools* and *Page Speed Online*.
6. User Experience: *Feedback Army* and *UserVoice*.

OptimalSort was already considered as part of the *Chalkmark* package. Other UTT referred were discarded, mainly because they were designed to evaluate specific aspects and not to support a significant coverage of all required usability issues.

Jules (Jules, 2011) presents the "best website usability testing tools and services", reviewing four UTT of our list that hadn't been previously discussed. The ten "essential website usability tools" discussed by LeMerle (LeMerle, 2012) were also analysed during this study.

Besides the preliminary analysis of the descriptions in web pages articles, the official websites for each of the selected candidates were also analysed. In order to assure the presence (or absence) of the features under assessment, all the content available was analysed, namely the videos demonstrating the UTT features.

4 Evaluation method

A simple additive utility function was used for providing a score on each UTT.

$$UF(UTT) = \sum_{j=1}^{m} w_j \sum_{k=1}^{n_j} w_{j,k}\, s_{j,k}\,(UTT)$$

This function linearly weights binary attributes $s_{j,k}(UTT)$ corresponding to the presence of elementary UTT features (0 for inexistent / 1 for implemented) using a two level hierarchy of weights. The second level $w_{j,k}$ weights the presence of feature *k* within the main issue *j*. Considering that issue *j* aggregates n_j features $\sum_{k=1}^{n_j} w_{j,k} = 1$. The first level aggregates the evaluation of *m* main issues where w_j is the weight determining the impact of the *j*-th main issue on the evaluation of the given UTT, where $\sum_{j=1}^{m} w_j = 1$.

The highest values found for this function should indicate the most suitable UTTs for our usability evaluations.

4.1 Utility model

The preferences (scores) for the main issues as well as for the features were set using an integer quantitative scale. Table 2 displays the correspondence between the quantitative values used and their qualitative importance.

Quantitative	Qualitative
5	Crucial
4	Important
3	Significant
2	Minor
1	Irrelevant

Table 2: Quantitative versus qualitative scale for setting preferences.

Weights were obtained by normalizing preferences into the interval [0;1]. Considering the preference for feature k within an issue j, represented by $p_{j,k}$, the corresponding weight is obtained by $w_{j,k} = p_{j,k} \,/\, \sum_{k=1}^{n_j} p_{j,k}$, where n_j is the number of features aggregated in issue j. This normalization ensures the equality $\sum_{k=1}^{n_j} w_{j,k} = 1$. Similarly, the weight for a main issue was computed as its relative contribution for the sum of all issues' preferences, thus ensuring $\sum_{j=1}^{m} w_j = 1$.

Preferences were obtained in two rounds by a team of three experts working for the project and having responsibilities in the task of interface design. All of them have a large experience in the development of software (ten or more years). In the first round each expert set up his/her own preferences in a printed form. The resulting printed forms were shared among the team. In a second round all the experts together discussed their scores until they agreed in a final number according to the quantitative scale of Table 2. In the remaining text we will refer to the above described scoring system as the Utility Model (UM).

4.2 Analytical hierarchy process

UM assumes criteria to be preferentially independent. The Analytical Hierarchy Process (AHP) (Saaty, 2005) also uses a linear additive model, but instead of giving absolute weights, the experts are questioned for pairwise comparisons of criteria and alternatives. This seems to be a much reasonable approach, namely because absolute values given in a single evaluation have very few references for providing the desired overall balanced result. Our AHP results were computed using a free trial version of commercial software (*Expert Choice Comparison*, 2012). This software considers all scores and makes all weights computations using a percentual scale. The pairwise comparison scale uses a judgment of preferences including nine categories: "extremely" preferred, "very strongly to extremely ", "very strongly", "strongly to very strongly", "strongly", "moderately to strongly", "moderately", "equally to moderately" and "equally" preferred.

A rating scale was used to score sub-criteria: the null value was assigned whenever a feature is absent; otherwise the score was set to 1. Though the AHP model has been criticized due to inconsistencies that can arise from weighting and scoring, we found easy to overcome them through a careful analysis and comparison setting. Again the preferences were set up in a collaborative meeting.

5 Results

5.1 Utility model

Table 3 presents the most significant results obtained by using the UM. The first column displays the main issues considered and the features aggregated under each issue. The second column presents the preferences specified for issues and features, on a 1-5 scale according to Table 2.

The UTTs under evaluation are presented in the first line and have been ranked according to their final scores, which were computed using the utility function and normalized to a 1-10 scale (last line). The column for each UTT also displays information about the presence or absence of each feature, represented by a 1 or a null value in the corresponding cell, respectively; and the values of relative scores for issues.

The best scored UTT, *Morae*, although providing limited user access was not excluded from our analysis because it presents good scores in almost all the other issues. However, this limitation may restraint remote or online usability tests, which is a major requirement in this project. Final decision about the election of the UTT to adopt should be based on testing the UTT since, at the present stage, our evaluation was mainly supported by industrial advertising information. Analogously, *Loop 11*, ranked in second place presents high preferences in the majority of issues. It was not excluded from the evaluation, despite not offering features for collecting audio – another important feature. The best ranked next three UTTs, *User Testing*, *Userfly* and *Usabilla*, also present good scores, offering all the required functionalities, even in a limited way. *Usabilla* is an exception as it does not provide audio collecting or video recording, which can be too confining.

Collecting additional information and testing the UTTs would be advantageous to support a final decision, as this study was mainly supported by industrial advertising information. Even considering the limitations above, Table 3 still provides a fair ranking suggestion for UTT selection, but then we present a new model based on the results of comparison.

Criteria and features	**preferences / UTT**	*Morae*	*Loop 11*	*User Testing*	*Userfly*	*Usabilla*	*Silverback*	*Click density*	*Intuition HQ*	*4Q Survey*	*Mouse Flow*	*Open Hallway*	*Google Analytics*	*ClickTale*	*Chalkmark*	*Fivesecondtest*	*Ethnio*	*Mechanical Turk*	*Concept Feedback*	*Crazyegg*	*Feng-GUI*	*Feedback Army*	*Attention Wizzard*	*ClickHeat*
OS compatibility	**3**	**4**	**10**	**10**	**10**	**10**	**3**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
- windows	**4**	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
- mac os	**3**	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
- linux	**4**	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Types of interfaces supported	**4**	**5**	**5**	**5**	**5**	**8**	**5**	**5**	**8**	**5**	**5**	**5**	**9**	**5**	**5**	**8**	**5**	**5**	**9**	**5**	**1**	**5**	**5**	**5**
- applications	**4**	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1
- prototypes	**5**	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1
- screenshots of website	**2**	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0
- wireframes	**2**	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0
- mockups	**4**	0	0	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0
Interface integration	**4**	**3**	**2**	**2**	**3**	**3**	**3**	**3**	**1**	**3**	**3**	**2**	**3**	**3**	**1**	**1**	**3**	**2**	**1**	**3**	**1**	**1**	**1**	**3**
- offline program	**4**	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- online post /URL submission	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0
- upload images	**1**	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0
- Javascript code	**4**	0	0	0	1	1	0	1	0	1	1	0	1	1	0	0	1	0	0	1	0	0	0	1
- online wizard	**3**	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
Usability test access	**5**	**1**	**9**	**5**	**5**	**9**	**1**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
- local	**1**	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- remote	**4**	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- online	**5**	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Surveys	**4**	**10**	**10**	**6**	**6**	**6**	**6**	**0**	**3**	**6**	**0**	**0**	**0**	**0**	**6**	**3**	**7**	**6**	**0**	**0**	**0**	**3**	**0**	**0**
- complete survey	**5**	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
- screen aligned questions	**4**	1	1	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1	0	0	0	1	0	0
- screen aligned text	**3**	1	1	1	1	1	1	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0
Collecting audio	**4**	**10**	**0**	**10**	**6**	**0**	**10**	**0**	**0**	**0**	**0**	**6**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
- audio record	**5**	1	0	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
- annotations	**4**	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Collecting video	**5**	**10**	**3**	**5**	**3**	**0**	**8**	**3**	**5**	**0**	**3**	**3**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
- display	**5**	1	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
- facial expressions recording	**4**	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- eye tracking	**4**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- annotations	**4**	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Usability map types supported	**4**	**6**	**5**	**0**	**5**	**2**	**0**	**5**	**2**	**0**	**7**	**0**	**0**	**6**	**2**	**2**	**0**	**0**	**0**	**5**	**5**	**0**	**1**	**2**
- clicks	**5**	1	1	0	1	1	0	1	1	0	1	0	0	1	1	1	0	0	0	1	0	0	0	1
- mouse move	**5**	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
- scroll reach	**4**	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
- attractive zones	**4**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
- interest zones	**4**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
- attention zones	**4**	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0
- form inputs	**5**	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
Export results	**5**	**6**	**6**	**2**	**2**	**3**	**0**	**5**	**2**	**5**	**2**	**2**	**5**	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**3**	**0**	**2**	**0**
- XLS / CSV / TSV	**5**	1	1	0	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
- XML	**4**	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
- database	**3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
- online mng. results access	**3**	0	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	0
UM		**8**	**7**	**6**	**6**	**5**	**5**	**5**	**5**	**5**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**3**	**3**	**3**	**3**	**3**	**3**

Table 3: Main issues and features preferences in a 1-5 scale. Relative and final scores in a 0-10 scale.

5.2 Analytical hierarchy process

Table 4 presents the results obtained from AHP study. All numbers are displayed as percentages. The first column ranks the UTTs according to AHP results. Each of the columns 1-9 displays a criterion (main issue), its weight (second row) and the importance of each UTT in this criterion. Column "AHP" displays the relative importance of the UTT obtained by AHP, while "AHP (%)" displays the corresponding normalization considering 100% for the best alternative scores. Their counterparts "UM" and "UM (%)" display the same numbers obtained by the UM.

The best scored UTT, *Loop 11*, presents high preferences for criteria considered crucial (4, 7 and 9). In addition, it reached satisfactory scores for the other criteria. It does not offer the features of criterion 6, however, this will not exclude it from our choice. The second UTT, Morae, provides limited user access, which may restraint remote usability tests. However, this UTT presents good scores in almost all the other criteria and, consequently, was not excluded. Considering that this evaluation was mainly supported by industrial advertising information, additional information is needed.

The next four UTT, *Usabilla*, *Click Density*, *Userfly* and *User Testing* present good scores, offering all the required functionalities, even in a limited way, with the exception of *Usabilla* that does not provide audio and video recording.

Sensitivity analysis allowed us to conclude that the "User Access" weight strongly influences the relative importance of *Morae*.

AHP produced results finer tuned than the previously obtained by the UM, highlighting the relative differences between UTTs. This is also disclosed by the standard deviation values displayed in the last line. The pairwise comparison of criteria is also more comprehensive than the normative assignment of marks, either in a quantitative or qualitative scale. Though small differences were found in the relative positions, the most significant difference concerns the first two UTTs, which can be explained by the tuned comparison of criteria preferences. These results should be interpreted carefully. Besides the limited type of sampling, most of the features were reduced to binary evaluation.

Scalability, for instance in the number of surveys or usability tests, seems often just a question of pricing. However, some of the features, even when present, may have some limitations when compared to a similar implementation in another UTT. Ultimately, some very specific features which can be highly valuable are only provided by few UTT. It should also be noted that all the preferences were defined by a small number of experts and considering the requirements of a specific project (World Search Project). Pricing can obviously be an important restriction for any product, which in this case was decided to be considered separately. It is still interesting to find some correlation between the price and the number of features or their specificity. Again, scalability can produce very significant pricing differences.

Criteria	1. OS compatibility	2. Supported types of prototypes	3. Interface integration w/ the UTT	4. User access	5. Creation & submission of surveys	6. Collecting audio	7. Collecting video	8. Usability maps	9. Export	AHP	UM	AHP (%)	UM (%)
UTT /Weights	3	7	7	27	7	7	18	7	18				
Loop 11	5	4	3	8	13	0	8	9	15	8	7	100	90
Morae	2	4	7	1	13	24	16	12	15	7	8	93	100
Usabilla	5	6	7	8	8	0	0	4	11	6	5	77	70
Click density	5	4	7	4	0	0	8	7	12	6	5	71	62
Userfly	5	4	7	4	8	14	8	11	2	6	6	70	74
User Testing	5	4	3	4	8	24	11	0	2	6	6	70	78
Mouse Flow	5	4	7	4	0	0	8	13	2	5	4	60	58
Intuition HQ	5	5	1	4	4	0	11	4	2	5	5	59	61
4Q Survey	5	4	7	4	8	0	0	0	12	5	5	59	59
ClickTale	5	4	7	4	0	0	8	12	0	4	4	57	51
Google Analytics	5	6	7	4	0	0	0	0	12	4	4	56	53
Open Hallway	5	4	3	4	0	14	8	0	2	4	4	55	54
Silverback	1	4	7	1	8	24	14	0	0	4	5	55	63
Fivesecondtest	5	6	1	4	4	0	0	4	2	3	4	44	49
Ethnio	5	4	7	4	9	0	0	0	0	3	4	42	49
Crazyegg	5	4	7	4	0	0	0	9	0	3	3	41	42
Mechanical Turk	5	4	3	4	8	0	0	0	0	3	4	40	46
Chalkmark	5	2	1	4	8	0	0	4	2	3	4	39	49
ClickHeat	5	4	7	4	0	0	0	4	0	3	3	38	38
Concept Feedback	5	6	1	4	0	0	0	0	2	3	3	38	43
Feng-GUI	5	0	1	4	0	0	0	7	5	3	3	37	40
Feedback Army	5	4	1	4	4	0	0	0	0	3	3	36	40
Attention Wizzard	5	4	1	4	0	0	0	1	2	3	3	36	39
							Std deviation			1,4	1,3	18	16

Table 4: AHP results – compared with previous UM results.

6 Conclusions and future work

Our team main concern in the World Search Project (World Search Project, 2010) is to enforce a user-centered design approach in a set of advanced information search demonstrators for specific domains. This kind of approach can benefit from using integrated usability testing tools (UTTs) for new applications design and development. Experience teams working regularly with a suitable UTT can better concentrate on solving usability issues and proposing innovative products. New team members can also find a good reference for integration by sharing such UTT capabilities with more experienced member teams. To the best of our knowledge, our study is the first quantitative evaluation and comparison of a significant number of UTTs within the context of Web graphical interfaces design. A special effort was given to include in our list all UTTs adequate to this context. A simple linear utility function and AHP model was used to score and rank 23 UTTs. Weighting and scoring was performed by a small team of experts.

The presented results should be considered with caution, due to the limited type of evaluation, namely almost exclusively based on the vendor's descriptions. Future work is expected in three different directions. The first direction will investigate and test other suitable multiple criteria decision analysis methods (Cechich et al., 2003; Figueira et al., 2004). A second direction will increase the number of experts for getting more reliable preferences and perhaps including new features. A third direction will verify features in lab for the preferred set of candidates. There will be an extra concern on usability tests/ UTTs features for applications running in mobile devices.

7 Acknowledgement

This work was partially supported by project QREN – I&D / ADI Nº 11495 (World Search Project, 2010) and by National Funding from FCT - Fundação para a Ciência e a Tecnologia, under the project: PEst-OE/MAT/UI0152.

References

[1] Cechich A., Piattini M., and Vallencillo A., 2003. *Component-Based Software Quality*, Springer Verlag, Berlin, Heidelberg.

[2] Dix A., Finlay J., Abowd G., Beale R., 2003. *Human Computer Interaction*, Prentice-Hall, Upper Saddle River, NJ, USA.

[3] Expert Choice, 2012. [online]. Available: http://expertchoice.com [10 October 2012].

[4] Fadeyev, D.,2009. *10 Tools to Improve Your Site's Usability on a Low Budget*, [Online], Available: http://www.webdesignerdepot.com/2009/06/10-tools-to-improve-your-site%E2%80%99s-usability-on-a-low-budget/ [23 July 2012].

[5] Figueira J., Greco S., Ehrgott M., 2005. *Multiple Criteria Decision Analysis*, Springer Verlag, Boston Dordrecht, London.

[6] Gube, J., 2011. *22 Essential Tools for Testing Your Website's Usability*, [Online], Available: http://mashable.com/2011/09/30/website-usability-tools/ [24 July 2012].

[7] Jules, 2011. *Best Website Usability Testing Tools and Services*, [Online], Available: http://www.quertime.com/article/arn-2011-04-06-1-best-website-usability-testing-tools-and-services/ [24 July 2012].

[8] LeMerle, R., 2012. *10 Essential Website Usability Tools*, [Online], Available: http://blog.ineedhits.com/tips-advice/10-essential-website-usability-tools-045711224.html [24 July 2012].

[9] Nielsen J., 1993. *Usability Engineering*, Academic Press, Boston.

[10] Nielsen J., 1999. *Designing Web Usability*, New Riders Publishing, Thousand Oaks, CA, USA.

[11] Norman D., 2002. *The Design of Everyday Things*, Basic Books, New York.

[12] Preece J., Rogers Y., Sharp H., Benyon D., Holland S., Carey T., 1994. *Human Computer Interaction*, Addison Wesley. Reading, Massachusetts.

[13] Saaty, T.L., 2005. The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making. In J. Figueira, S. Greco, and M. Ehrgott (eds.) MCDA: State of the Art Surveys, Springer Verlag.

[14] Shneiderman B., 1998. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Addison Wesley, Reading, Massachusetts.

[15] Tomlin, W., 2009. *24 Usability Testing Tools*, [Online], Available: http://www.usefulusability.com/24-usability-testing-tools/ [23 July 2012].

[16] Vraa, L., 2009. *16 Crucial Webdesign and Usability Best Practice Compilations and Tools*, [Online], Available: http://www.tripwiremagazine.com/2009/06/16-crucial-webdesign-and-usability-best-practice-compilations-and-tools.html [23 July 2012].

[17] Walker, T., 2010. *20 Fantastic Usability & Conversion Analysis Tools*, [Online], Available: http://spyrestudios.com/usability-conversion-analysis-tools/ [24 July 2012].

[18] World Search Project, 2010. [Online], Available: http://www.microsoft.com/portugal/mldc/worldsearch/en/ [March 2011].