A Systematic Review of Usability Evaluation in Web Development[1]

Emilio Insfran, Adrian Fernandez

ISSI Group, Department of Information Systems and Computation
Universidad Politécnica de Valencia
Camino de Vera, s/n, 46022, Valencia, Spain
{einsfran, afernandez}@dsic.upv.es

Abstract. The challenge of developing more usable Web applications has motivated the appearance of a number of techniques, methods and tools to address Web usability issues. Although there are many proposals for supporting the development of usable Web applications, many developers are not aware of them and many organizations do not properly apply them. This paper reports on a systematic review of the use of usability evaluation methods in Web development. The objective of the review is to investigate what usability evaluation methods have been employed by researchers to evaluate Web artifacts and how they were employed. A total of 51 research papers have been reviewed from an initial set of 410 papers. The results show that 45% of the papers reviewed reported the use of evaluation methods specifically crafted for the Web and that the most employed method is user testing. In addition, the results of the review have identified several research gaps. Specifically, 80% of the evaluations are still performed at the implementation phase of Web applications development and 47% of the papers did not present any validation of the usability evaluation method(s) employed.

Keywords: Usability Evaluation Methods, Web development, Systematic Review.

1 Introduction

Usability is a crucial factor in Web application development. The ease or difficulty that users experience with systems of this kind will determine their success or failure. As Web applications have become the backbone of business and information exchange, the need for usability evaluation methods specifically crafted for the Web – and technologies that support the usability design process – has become critical [21].

The challenge of developing more usable Web applications has motivated the appearance of a variety of techniques, methods and tools to address Web usability issues. Although there are many proposals for supporting the development of usable Web applications, many developers are not aware of them and many organizations do

[1] This work is funded by the META project (TIN2006-15175-C05-05), the Quality-driven model transformations project (UPV), and the CALIPSO research network (TIN2005-24055-E).

not properly apply them. To address this issue, several studies aimed at comparing usability evaluation methods for Web development were reported (e.g., [1], [11]). These studies often compare a reduced number of evaluation methods, and the selection of methods is normally driven by the expectations of the researcher. Therefore, there is a need to identify, in a more systematic way, what usability evaluation methods have been successfully applied to Web development.

In this paper, we present a systematic review for assessing what usability evaluation methods have been employed for Web usability evaluation and their relation to the Web development process. Systematic reviews are useful for summarizing all existing information about a phenomenon of interest (e.g., a particular research question) in an unbiased manner [14]. The goal of our review is, therefore, to examine the current use of usability evaluation methods in Web development from the point of view of the following research questions: *what usability evaluation methods have been employed by researchers to evaluate Web artifacts and how were they employed?*

This paper is organized as follows. Section 2 discusses related work. Section 3 presents the protocol we used to review the usability evaluation methods employed in Web development. Section 4 describes the results of the systematic review. Section 5 discusses the threats to the validity of the results. Finally, section 6 presents our conclusions and suggests areas for further investigation.

2 Related Work

A number of studies aimed at comparing usability evaluation methods for Web development have been reported in the last few years (e.g., [23], [1]).

One of the most complete studies was published by Ivory and Hearst [23] in 2002. They proposed a taxonomy for classifying automated usability evaluation methods. The taxonomy was applied to 128 usability evaluation methods, where 58 of them are suitable for Web user interfaces. The results of this survey suggest promising ways to expand existing methods to better support automated usability evaluation.

Another study by Alva et al. [1] presented an evaluation of seven methods and tools for usability evaluation in software products and artifacts for the Web. The purpose of this study was to determine the degree of similarity among the methods using the principles defined in the ISO 9241-11 standard [12]. However, this is an informal survey with no defined research questions and no search process to identify the methods that were considered.

Batra and Bishu [3] reported the results obtained with two usability evaluation studies for Web applications. The objective of the first study was to compare the efficiency and effectiveness between user testing and heuristic evaluation. The results showed that both methods addressed very different usability problems and are equally efficient and effective for Web usability evaluation. The objective of the second study was to compare the performance between remote and traditional usability testing. The results indicate that there is no significant difference between the two methods.

Although several comparisons about usability evaluation methods have been reported, we are not aware of any existing systematic review published on the field of

Web usability. The majority of the published studies are informal literature surveys or comparisons with no defined research questions, no search process, no defined data extraction or data analysis process. We only found two systematic reviews conducted in related fields [9], [19]. Freire et al. [9] presented a systematic review on Web accessibility to identify existing techniques for developing accessible content in Web applications. This review includes 53 studies, and it also proposes a classification of these techniques according to the processes described in the ISO/IEC 12207 standard [13]. Mendes [19] presented a systematic review to investigate the rigor of claims of Web engineering research.

3 Research Method

A systematic review is a means of evaluating and interpreting all available research that is relevant to a particular research question, topic area, or phenomenon of interest [14]. It aims at presenting a fair evaluation of a research topic by using a trustworthy, rigorous, and auditable methodology.

A systematic review involves several stages and activities. In the *planning the review* stage, the need for the review is identified, the research questions are specified, and the review protocol is defined. In the *conducting the review* stage, the primary studies are selected, the quality assessment used to include studies is defined, the data extraction and monitoring is performed, and the obtained data is synthesized. Finally, in the *reporting the review* stage, the dissemination mechanisms are specified, and the review report is presented. The activities concerning the planning and the conducting of our systematic review are described in the following subsections. The reporting the review stage is presented in Section 4.

3.1 Research Question

We have carried out a systematic literature review using the approach suggested in [14]. The goal of our study is to examine the current use of usability evaluation methods in Web development from the point of view of the following research question: *What usability evaluation methods have been employed by researchers to evaluate Web artifacts and how were they employed?* The criteria used to classify the evaluation methods are presented in Section 3.3.

This research question will allow us to summarize the current knowledge about Web usability evaluation and to identify gaps in current research in order to suggest areas for further investigation. The study's population and intervention is as follows:

- **Population**: Web usability full research papers
- **Intervention**: Usability evaluation methods
- **Outcome**: No focus on the outcome itself
- **Experimental design**: Any design

Our review is more limited than a full systematic review as suggested in [14] since we did not follow up the references in papers. In addition, we did not include other

references such as technical reports, working papers and PhD theses. This strategy has been used in another systematic review conducted in the Web Engineering field [19].

3.2 Identifying and Selecting Primary Studies

The main sources we used to search for primary studies are IEEExplore and ACM digital libraries. In addition, we have included the proceedings of the following special issues and conferences:

- World Wide Web conference proceedings – WWW (2003, 2004, 2007), Usability and accessibility & Web engineering tracks [26] [7], [27].
- International conference on Web Engineering proceedings – ICWE (2003-2007) [16], [15], [17], [25], [2].
- IEEE Internet Computing Special issue on "Usability and the Web" (1 volume published in 2002) [21].
- A book on Web Engineering by Springer (LNCS) published in 2005 [20].
- International Web Usability and Accessibility workshop proceedings – IWWUA (2007) [24].

The search string defined for retrieving studies is as follows: *usability AND web AND development AND (evaluation OR experiment OR study OR testing)*

We experimented with several search strings and this one retrieved the greatest amount of relevant papers. This search string was used in the IEEExplore and the ACM digital libraries as well as in the other sources that were inspected manually. The period reviewed was the last 10 years, i.e., studies published from 1998 to 2008. With respect to the digital libraries, we ensured that our search strategy was applied to magazines, journals and conference proceedings.

3.3 Inclusion Criteria and Procedures

Each identified study was evaluated the researchers conducting the systematic review to decide whether or not it should be included. The discrepancies were solved by consensus. The studies that met the following conditions were included:

- Studies presenting usability evaluation method(s) that are applied to Web development. Only studies that presented a "formal" method (e.g., heuristic evaluation, cognitive walkthrough) were selected.
- Full research papers.

The following types of papers were excluded:

- Papers presenting recommendations and principles for Web design.
- Papers presenting techniques on how to aggregate usability measures.
- Papers presenting usability metrics.
- Introductory papers for special issues, books, and workshops.
- Papers not written in English.

3.4 Data Extraction Strategy

The data extracted were compared according to the research questions stated, which are decomposed into the following criteria:

1. What usability evaluation methods (UEMs) have been employed by researchers to evaluate Web artifacts?
    i. Is it a new evaluation method or an existing method from the HCI field? (New, Existing)
    ii. What is the type of usability evaluation method employed? (Inspection method, User testing, Other)
2. What is the phase in which the evaluation method is applied? (Requirements, Design, Implementation)
3. What is the type of evaluation? (Manual, Automated)
4. Was the evaluation method evaluated? (Yes, No). If yes:
    i. What type of evaluation was conducted? (Survey, Case study, Experiment)
5. Was the evaluation conducted with the intention to provide feedback to the design? (Yes, No)

With regard to the first criterion, the paper is classified as *new* if it presents at least one evaluation method that is specifically crafted for the Web. Otherwise, it is classified as *existing* if the paper uses existing methods from the HCI field.

In addition, the evaluation method is classified according to the following types: inspection method, user testing, or other. The paper is classified as *inspection method* if it reports an evaluation based on expert opinion (e.g., heuristic evaluation, guideline reviews, standards inspection, cognitive walkthroughs). Otherwise, the paper is classified as *user testing* if it reports an evaluation that involves the user's participation. Such evaluations typically focus on lower-level cognitive or perceptual tasks. In this category, we also consider the several protocols that exist to conduct user testing (e.g., thinking aloud, question-asking). Finally, the paper is classified as *others* if it reports the use of other methods (e.g., focus group, web usage analysis).

With regard to the second criterion (the phase in which the evaluation is conducted), each paper is classified into one or more ISO/IEC 12207 high-level processes: Requirements, Design, and Software Construction (Implementation). The paper is classified at the *requirements* phase if the artifacts used as input for the evaluation include high-level specifications of the Web application (e.g., task models, uses cases, scenarios). The paper is classified at the *design* phase if the evaluation is conducted on the intermediate artifacts of the Web application (e.g., navigational models, abstract user interface models, dialog models). Finally, the paper is classified at the *implementation* phase if the evaluation is conducted in the Web application.

With regard to the third (the type of evaluation conducted), the paper is classified as *manual* if it presents a usability evaluation that is manually performed. Otherwise, it is classified as *automated.* The fourth criterion is related to the evaluation of the usability evaluation methods. Depending on the purpose of the evaluation and the conditions for empirical investigation, three different types of strategies can be carried out [8]: survey, case study and experiment. A *survey* is an investigation performed in retrospect, when the method has been in use for a certain period of time. A *case study*

is an observational study and data is collected for a specific purpose throughout the study. An *experiment* is a formal, rigorous and controlled investigation. Experiments provide a high level of control and are useful for comparing usability evaluation methods in a more rigorous way. For evaluations of this type, statistical methods are applied in order to determine which method is better.

Finally, the fifth criterion is to determine whether or not the evaluation method provides feedback to the designer. The evaluation method is classified as *No* if it is aimed at only reporting usability problems. The method is classified as *Yes* if it also provides recommendations on how the problems can be fixed.

3.5 Conducting the review

The search to identify primary studies in the IEEExplore and ACM digital libraries was conducted on the 22nd of March 2008. The application of the review protocol yielded the following results:

- The bibliographic database search identified 338 potentially relevant publications (181 from the IEEExplore and 157 from the ACM digital library). After applying the exclusion criteria documented in Section 3.3, 37 publications were finally selected (11 from IEEExplore and 26 from ACM digital library).
- The manual bibliographic review of the other sources identified another 72 potentially relevant publications. After applying the exclusion criteria, the following publications were finally selected: 14 papers (3 from WWW, 3 from ICWE, 3 from the IEEE Internet Computing special issue, 4 from IWWUA, and 1 chapter from the book).

Therefore, a total of 51 research papers were selected by our inclusion criteria. Some studies had been published in more than one journal/conference. In this case, we selected only the most complete version of the study. Other studies appeared in more than one source. These publications were taken into account only once. The searching results revealed that research papers about Web usability are published in several conferences/journals from different fields, such as Human-Computer Interaction (HCI), Web Engineering (WE), and other related fields.

4 Results

The results of our study are presented in Table 1. They have been organized by selection criteria and publication source. The list of papers containing all the data extracted from the studies was not included in this paper due to space restrictions.

These results indicate that 45% of the papers reviewed presented new evaluation methods specifically designed for the Web (see Fig. 1 (a)). For instance, Blackmon et al. [5] proposed the cognitive walkthrough for the web (CWW) method. When compared to the traditional method, this method was found to be superior for evaluating how well websites support user navigation and information search tasks. In another study, Bolchini and Garzotto [6] proposed a usability inspection method for Web applications called MiLE+. The method was evaluated through two studies that

measured the efficiency, performance, and the perceived difficulty of learning the method. The remaining 55% of the studies reported the use of existing evaluation methods (e.g., cognitive walkthrough, heuristic evaluation, user testing).

Table 1. Systematic Review Results

Selection criteria		IEEE	ACM	WWW	ICWE	IE3IC	Book	IWWUA	Total
Usability Evaluation Method	New	4	9	2	3	3	0	2	23
	Existing	7	17	1	0	0	1	2	28
Type of Usability Evaluation Method	Inspection method	4	5	0	1	1	1	1	13
	User testing	7	17	1	0	0	1	0	26
	Other	4	11	2	2	2	1	3	25
Web development phase	Requirements	1	1	0	0	0	0	1	3
	Design	4	4	0	1	3	1	3	16
	Implementation	7	25	3	3	1	1	1	41
Type of evaluation	Manual	9	19	0	1	1	1	4	35
	Automated	2	7	3	2	2	0	0	16
Validation?	Survey	0	3	0	0	0	0	0	3
	Case study	1	3	2	1	0	1	3	11
	Experiment	2	10	0	0	0	0	1	13
	No	8	10	1	2	3	0	0	24
Feedback to design?	Yes	4	6	0	0	2	0	3	15
	No	7	20	3	3	1	1	1	36

IEEE – IEEExplore electronic database
ACM – ACM digital library
WWW – World-Wide Web conference from 2003 to 2007
ICWE – International Conference on Web Engineering from 2003 to 2007
IE3IC – IEEE Internet Computing Special Issue on Usability and the Web
Book – A book on Web Engineering by Springer
IWWUA – International Workshop on Web Usability and Accessibility 2007

The results also revealed that the most frequently used type of evaluation method is user testing, i.e., 41% of the papers reviewed reported some kind of testing involving users (see Fig. 1 (b)). This may indicate that most evaluations are performed mainly during late stages of the Web development lifecycle. Inspections accounts for 20% of the studies, whereas 39% of the studies reported the use of other methods (e.g., paper prototype, remote user testing, survey). An example of the use of inspection methods is described in Sutcliffe [22]. The author proposed a set of heuristics for assessing the attractiveness of Web user interfaces. The heuristics were tested by evaluating three airline websites. The results of the study show that aesthetics may play an important role for initial visits but content issues may be dominant for repeat visits.

The analysis of the results confirmed that the evaluations are mainly performed at the implementation level (68%) of the Web application (see Fig. 1(c)). Around 27% of the studies describe evaluations performed using the Web application's intermediate artifacts (e.g., abstract user interface, navigational model). Only 5% of the evaluations were performed at the requirements specification level (e.g., laboratory user testing of paper mock-ups or prototypes). Therefore, there is a need for usability evaluation methods that can be used at early stages of Web development.

With regard to the type of evaluation, 69% of the studies performed the evaluations manually (see Fig. 1 (d)). Around 31% of the studies reported the existence of some kind of automated tool to support the proposed method. For instance, Becker and

Berkemeyer [4] proposed a technique to support the development of usable Web applications. The technique is supported by a GUI-based toolset called RAD-T (rapid application design and testing) that allows early usability testing at the design stage.

We also verified whether the studies reported some kind of empirical evaluation. The results revealed that 47% of the studies did not conduct any type of evaluation (see Fig. 1 (e)). However, it was surprising to observe that, from the papers that did perform evaluations, 25% of them reported on controlled experiments. The majority of these studies were published in HCI conferences and journals; hence, experimentation is a common research method used in this field. An example of this is the study conducted by Hornbæk and Frøkjær [11], where two psychology-based inspection techniques (cognitive walkthrough (CW) and metaphors of human thinking (MOT)) were compared. The results show that the participants identified 30% more usability problems using MOT. Around 22% of the studies report case studies. For instance, Matera et al. [18] presented a case study in which three methods were applied to the evaluation of a Web application: design inspections to examine the hypertext specification, web usage analysis to analyze the user behavior, and heuristic evaluation to analyze the released prototypes and the final Web application.

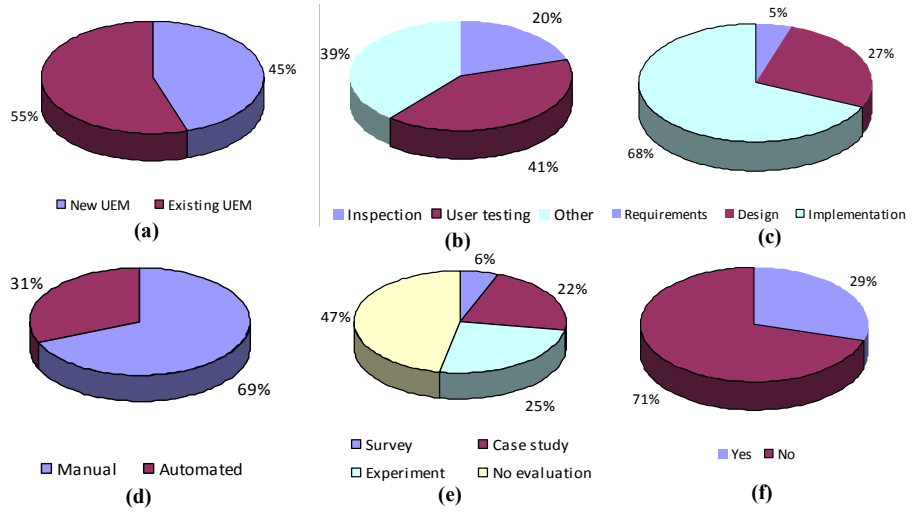

Fig. 1. Percentage of coverage by criteria used for data extraction

Finally, 71% of the studies reported only on usability problems giving no feedback on the corresponding design artifacts (see Fig. 1 (f)). The remaining studies (29%) also offered suggestions for design changes based on the usability problems detected. For instance, Hornbæk and Frøkjær [10] reported an experiment aimed at comparing the assessment of both usability and utility of problems and redesign suggestions. The results of the experiment showed how redesign proposals were assessed by developers as being of higher utility than just problem descriptions. Usability problems were seen more as a help in prioritizing ongoing design decisions.

Figure 2 shows the number of selected publications on Web usability evaluation methods by year and source. The analysis of the number of research studies on Web

usability showed that there has been a growth of interest on this topic. Most of the studies about Web usability were found at the ACM digital library.

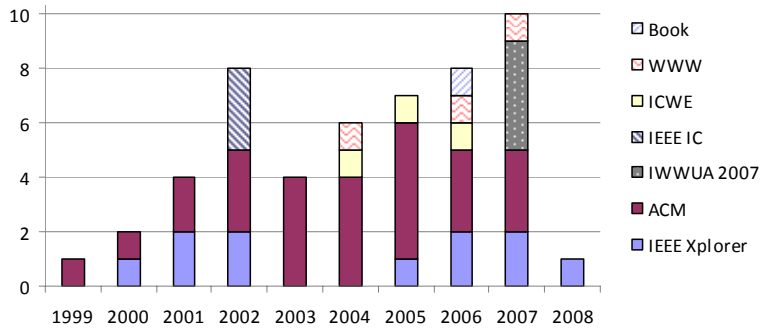

Fig. 2. Number of Publications on Web Usability by Year and Source

5 Threats to Validity

The main limitations of this study are publication selection bias, inaccuracy in data extraction, and misclassification. Publication bias refers to the problem that positive results are more likely to be published than negative results [14]. We believe that we have alleviated this threat, at least to some extent, by scanning relevant journal special issues and conference proceedings. However, we did not consider grey literature or unpublished results. With regard to publication selection, we chose the sources where papers about Web usability are normally published. However, we have excluded some journals in the Web engineering field from this systematic review (i.e., Journal of Web Engineering and International Journal of Web Engineering and Technology) since we had no access to these journals. This fact could affect the validity of our results. We attempted to alleviate the threats of inaccuracy in data extraction and misclassification by conducting the classifications of the papers with three reviewers.

6 Conclusions and Future Work

This paper has presented a systematic review of usability evaluation methods for Web development. The results of the review have identified several research gaps.

In particular, usability evaluations should be performed early in the Web development process and should occur repeatedly throughout the design cycle, not just when the product has been completed. The majority of the papers reported on evaluations at the implementation phase. It also reveals that the evaluations are mainly performed in a single phase of the Web application development. Usability evaluation at each phase of the Web application development is critical for ensuring that the product will actually be used and be effective for its intended purpose(s). In addition, the majority of the methods reviewed only allowed the generation of a list of

usability problems. New proposals for redesign that address usability problems as an integral part of the evaluation method are needed.

Although our findings may be indicative of the field, further reviews are needed to confirm the results obtained. Future work includes the extension of this review by including other sources (e.g., Science Direct and Scopus databases). We also want to analyze more in-depth the level of integration of the usability evaluation methods into the different processes of the Web application lifecycle. Finally, we plan to collect more information about the empirical evidence of the effectiveness of usability evaluation methods for the Web.

References

1. Alva O. M. E., Martínez Prieto, A. B., Cueva Lovelle, J. M., Sagástegui Ch. T. H., López, B. Comparison of Methods and Existing Tools for the Measurement of Usability in the Web. Proc. Int. Conf. on Web Engineering 2003, Spain. Springer Verlag, pp. 386-389.
2. Baresi L., Fraternali P., Houben G. (Eds.): Proc. of the International Conference on Web Engineering 2007, Como, Italy, July 16-20, 2007, LNCS 4607 Springer 2007.
3. Batra S.,Bishu R.R. Web Usability and Evaluation: Issues and Concerns. Usability and Internationalization. HCI and Culture, LNCS 4559, 2007, pp. 243-249.
4. Becker S. A., Berkemeyer A., Rapid Application Design and Testing of Web Usability. IEEE Multimedia, 9(4): 38-46, October/December 2002.
5. Blackmon M. H., Polson P. G., Kitajima M. and Lewis C. Cognitive walkthrough for the web. Proc. of the CHI 2002, Minneapolis, Minnesota, USA, pp. 463 – 470, 2002.
6. Bolchini D., Garzotto F. Quality of Web Usability Evaluation Methods: An Empirical Study on MiLE+, Proc. of the IWWUA 2007 , Nancy, France, 2007, pp. 481-492.
7. Feldman S. I., Uretsky M., Najork M., Wills C. E. (Eds.): Proc. of the International Conference on World Wide Web 2004, New York, USA, May 17-20, 2004. ACM 2004.
8. Fenton, N., and Pfleeger, S. L. Software Metrics: A Rigorous and Practical Approach, Second Edition. International Thomson Computer Press, 1996.
9. Freire A. P., Goularte R., Fortes R. P. M. Techniques for Developing more Accessible Applications: a Survey Towards a Process Classifications, Proc. of the 25th ACM Int. Conference on Design of communication, El Paso, Texas, USA, 2007, pp. 162 – 169.
10. Hornbæk K., Frøkjær E. Comparing Usability Problems and Redesign Proposals as Input to Practical Systems Development. Proc. of the CHI 2005, Portland, USA, pp. 391 – 400.
11. Hornbæk K., Frøkjær E. Two psychology-based usability inspection techniques studied in a diary experiment Proc. of the 3rd Nordic conference on Human-computer interaction (NordCHI'04), Tampere, Finland, pp. 3-12, 2004.
12. ISO – International Standard Organization, ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, 1998.
13. ISO – International Standard Organization, ISO/IEC 12207: Standard for Information Technology – Software Lifecycle Processes, 1998.
14. Kitchenham B. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Version 2.3, EBSE Technical Report, Keele University, UK.
15. Koch N., Fraternali P., Wirsing M. (Eds.): Proc. of the International Conference on Web Engineering 2004, Munich, Germany, July 26-30, 2004,. LNCS 3140, Springer.
16. Lovelle J. M. C., Rodríguez B. M. G., Aguilar L. J., Gayo J. E. L., Ruiz M. P. P (Eds.): Proc. of the Int. Conf. on Web Engineering 2003, Oviedo, Spain, LNCS 2722, Springer.
17. Lowe D., Gaedke M. (Eds.): Proc. of the International Conference on Web Engineering 2005, Sydney, Australia, July 27-29, 2005, LNCS 3579, Springer.

18. Matera M., Rizzo F., Carughi G. T. Web Usability: Principles and Evaluation Methods, In Web Engineering, Mendes E., Mosley N (eds.), pp. 143-179.
19. Mendes E. A Systematic Review of Web Engineering Research, Proc. of the International Symposium on Empirical Software Engineering (ISESE'05), 2005, pp. 498-507.
20. Mendes E., Mosley N (eds.), Web Engineering, 2005, Springer.
21. Neuwirth C. M., Regli S. H. IEEE Internet Computing Special Issue on Usability and the Web, Vol. 6, No. 2, March/April 2002.
22. Sutcliffe, A. Assessing the Reliability of Heuristic Evaluation for Website Attractiveness and Usability. Proc. of the HICSS 2002, Volume 5, pp. 137-141.
23. Yvory, M., Hearst, M. The State of the Art in Automating Usability Evaluation of User Interfaces. ACM Computing Surveys, 33(4):470-516, 2001.
24. Weske M., Hacid M. S., Godart C. (Eds.): Web Information Systems Engineering - WISE 2007 Workshops Proceedings, Nancy, France, December 3, 2007, LNCS 4832, Springer.
25. Wolber D., Calder N., Brooks C. H., Ginige A. (Eds.): Proc. of the 6th International Conference on Web Engineering 2006, Palo Alto-CA, USA, July 11-14, ACM 2006.
26. WWW03, Proc. of the Twelfth International World Wide Web Conference 2003, Budapest, Hungary, 20-24 May 2003. ACM, 2003, http://www2003.org.
27. WWW07, Proc. of the Twelfth International World Wide Web Conference 2007, Banff, Alberta, Canada, May 8-12, 2007. ACM, 2007, http://www2007.org.