MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-224

User Tests for Assessing a Medical Image Retrieval System: A Pilot Study

Dimitrios Markonis[a], Frederic Baroz[b], Rafael Luis Ruiz De Castaneda[b], Celia Boyer[b], Henning Müller[a]

[a] *University of Applied Sciences Western Switzerland, Sierre, Switzerland*
[b] *Health On the Net Foundation (HON), Geneva, Switzerland*

Abstract

Content-based image retrieval (CBIR) has often been proposed to assist medical decision making in complement to textual information search. However, applications of this novel technology have rarely reached the end users. The study presented in this paper describes the design and setup for performing pilot user tests in order to assess a medical information retrieval system that supports CBIR with the goal of having more detailed tests with an updated system. Five individuals with medical education participated in the study at the University hospitals of Geneva. They were recorded and observed while interacting with the system, and then provided feedback on the usability of the system. Participants seemed to understand the concept and practical uses of the new tools, and needed 10-15 minutes to feel confident with the system. The results of this pilot study will be used for improving the system functionalities as well as an input for designing a new iteration of larger-scale user tests among radiologists.

Keywords:

Usability tests, user-centered design, medical informatics applications, content-based image retrieval.

Introduction

Images are produced in a quickly increasing variety and quantity, and are essential in many aspects of medical diagnoses and treatment planning. Much of the knowledge stored in images is little exploited as access to the visual image information is rarely possible. Content-based image retrieval (CBIR) uses the visual content of example images to retrieve other images or cases. Over the past 15 years, CBIR has been considered promising for assisting information search in the medical fields and several systems have been developed [1]. However, most systems were technology-driven and very few applications reached the end users or integrated into the medical professionals' workflows [2].

User-centered design (UCD) [3] has been used for several decades in industry [4, 5], and medical applications [6]. It is driven by user requirements and feedbacks to improve the product's usability, and user experiences. A few aspects of UCD have also been used for CBIR [7].

UCD in software development includes key elements involved user feedbacks to the design and development of the application. The first element is investigation and understanding of the user requirements [8] which are needed to identify the general design directions. User-centered evaluation is another important part of UCD, which needs to be performed in the early stages of the development [10] and is seen as an iterative process throughout the development cycle [5]. The key elements are also described in the ISO standard for the Human-centered design for interactive systems (ISO 9241-210, 2010)[1].

User-centered evaluation is often performed in the form of empirical usability tests, having a number of target users to interact with the system. Usability of the system is assessed with factors such as learnability, efficiency, effectiveness, memorability, and satisfaction [10]. Various methods exist for conducting these tests, including thinking aloud, direct or recorded observation of the interaction, survey forms, and log analysis. A survey on common usability testing techniques and tools is presented in [11]. A more detailed description of important aspects of usability test design can be found in [12].

An important aspect when designing a usability test is the number of participants required. Early studies have discovered that a single individual is not able to detect all usability problems but 3-4 are sufficient [13]. In [14], it is suggested that five users are enough, while other studies have questioned this choice [15, 16]. The exact number of participants remains an open question, although in [17], it is explained that five participants are indeed enough for each iteration of an iterative user-centered evaluation.

In this article, the design choices, the setup, and the preliminary results of the first round of the user-centered evaluation of the Khresmoi[2] search engine are presented. This system aims at assisting general practitioners, the general public, and radiologists in accessing trustable biomedical information. These three target groups have differing search behaviors, goals and information requirements. Thus, the system is divided into three subsystems, designed to correspond to the needs of each target groups. Following the same concept, usability tests are designed and conducted separately in each target groups, concentrating on domain-specific research questions.

This study focuses on the pilot tests on the Khresmoi subsystem developed for radiologists. The system combines text and CBIR search for finding and navigating through scientific biomedical articles and their images. The prototype design is based on the investigation of the image use behavior of radiologists [18]. The development is based on the Parallel Distributed Image Search Engine (ParaDISE [19]) and ezDL [20].

[1] http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075

[2] http://www.khresmoi.eu/

Three research questions that iterative user-centered evaluation tries to answer for the particular subsystem are:

- Does the Khresmoi system improve current search for information in radiology (which is mainly patient-centered or using Google on the Internet)?
- Does it cover unmet information needs and to what extend?
- Which functionalities are useful and which tools need to be improved/changed/added?

The objective of this first iteration is an initial assessment of the integrated system to identify the most important usability problems and missing functionalities and to redefine the user study protocol for larger scale user tests.

Materials and Methods

User study protocol

After defining the research questions, appropriate methods for recording measures of the usability of the system need to be identified. Next, tasks that the participants perform need to be carefully chosen regarding the research questions and the assessment aspects. Finally, a step-by-step outline of a session is prepared. A combination of the proposed guidelines of [10] and [12] was followed in this study.

In order to assess the usability of the system, the following measures were used: efficiency, effectiveness, and user satisfaction. For efficiency, the time required to find the first relevant result during each task was measured. For effectiveness, the number of relevant documents found during each tasks was measured. The participants' computer screen and facial expressions were observed and video recorded during their interaction with the system. Finally, for evaluating the user satisfaction, survey forms and open discussion with the participants were used.

The recruitment of participants was done via personal contacts and people who volunteered to take part in the study are at the radiology department of the University hospitals of Geneva.

Session outline

Each session of the user tests consisted of the following steps:

1. Introduction to the Khresmoi project, the existing search system, and the user test goals.
2. Tutorial video on the system tools and functionalities.
3. Demographic survey.
4. Guided user tests in clear scenarios.
5. Survey on the satisfaction with the tools and functionalities.
6. Free possibility to use the system.
7. Survey on the satisfaction with the system, and open discussion.

The introduction intended to help the participant understand the concept of the system and to motivate him/her to do the test. Then, the video demonstration of the system introduced the tools offered by the application. During steps 3-7, the participant was observed by the observer to identify potential shortcomings of the system or the user study design itself. The observer was instructed to have a neutral attitude, allowed to help only when the participant was blocked and could not proceed with a task.

Task design and description

The design of the tasks took into account that the participants need to use most of the system tools and functionalities and cover the information needs of the target user group. They had to describe realistic scenarios that appear in clinical and academic workflows. For this reason, two groups of tasks were used: Four 2D image search tasks and two article search tasks. A subset of the ImageCLEF2012[3] medical image-based and case-based retrieval task topics was used respectively. The topics for the image-based task were selected after the log analysis of queries to a radiology image search engine [9], while case-based topics consisted of cases included in an educational database [21].

Session setup and tools used

For observation and recording, the commercial Morae[4] mentioned in [11] was used. Morae allows screen and face video recording, remote observation, and inclusion of introductory text, questionnaires, and task descriptions on screen. It is also compatible with commonly used statistical packages for result analysis and presentation.

A combination of a modified version of the System Usability Scale (SUS) [22] and the Questionnaire for User Interaction Satisfaction (QUIS) [23] was used for the user satisfaction survey forms. Open questions for providing comments on specific aspects of the system and suggestions for improvements were added. To get preliminary answers to the research goals, questions about the novelty, usefulness, and intention of use of the system were added.

The setup of the session includes hardware, software preparation, and training sessions of the observer to get familiar with the recording tool and the study purpose. The hardware used in each session includes two Windows computers – one for the participant and one for the observer. The Khresmoi client was downloaded to the participant's computer and Morae was installed on both computers.

At the end of each session, the files containing the recordings, the answers to the surveys, and the observer's notes were acquired. The details of preparing, setting up and running a session were added into a document to facilitate the experiment reproducibility.

Results

Demographics

Five individuals (2 females, 3 males) participated in two sets of parallel sessions at the University hospitals of Geneva. All participants were below 30 years old, with two of them below 25. Two participants had radiology background (one specializing in bones), one was a non-radiology intern and two were final year students in medicine. All participants declared frequent computer use. Three persons answered to search for medical info more than once per day, one answered once per day, and one answered once per week.

[3] http://www.imageclef.org/2012

[4] http://www.techsmith.com/morae.html

Efficiency - Effectiveness - User satisfaction

The mean time for retrieving the first relevant result during the 2D image search tasks was 158 seconds. This time included choosing image examples, investigating the results, and judging a result as relevant. This time includes only the cases when a relevant result was found. For case-based retrieval tasks, the respective mean time was 179 seconds.

The mean number of results selected as relevant was 5 for the 2D image search tasks and 2.6 for the case-based search. One participant (one still studying medicine) did not select any relevant results for all tasks.

User satisfaction on the specific system aspects was measured in a Likert scale where 1 was strongly negative and 5 was strongly positive. Results are given in Figure 1.

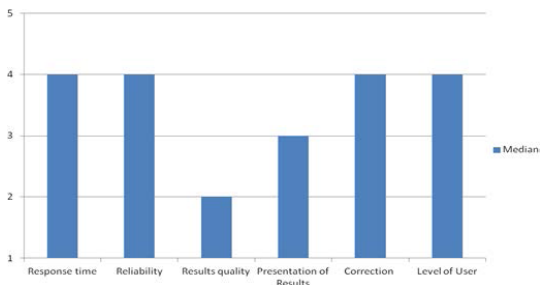

Figure 1- Median of measuring user satisfaction over specific system aspects in a Likert scale (1=strongly negative, 5=strongly positive).

The median for system response time was 4 in a Likert scale. The same median was obtained for system reliability. In terms of results quality and presentation, the median was 2 and 3 respectively, while both ability to correct mistakes and system design to be used by all levels of users obtained a median of 4.

All questions about the user intention in academic, research, and clinical work obtained medians of 4. Finally, questions regarding the practical usefulness of the novel features of the system obtained a median of 5 out of 6 due to a design error. This was excluded from the global user satisfaction evaluation. User satisfaction results over general aspects of the system are presented in Figure 2.

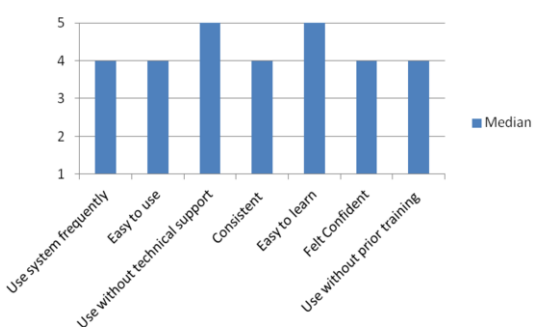

Figure 2- Median values of measuring general user satisfaction about the system in Likert scale.

The median for the questions about intention to use the system frequently was 4. The same median was obtained for easiness to use and consistency. The median for using the system without technical support was 5 and the same score was achieved for easiness to learn. Finally, the participants answered that they felt confident when they used the system and that they could use the system without prior training giving a median grade of 4.

In order to assess the global satisfaction of each participant, the mode over the general satisfaction questions was taken, measuring the most frequent grade given (Figure 3). Also, for measuring the consistency of this satisfaction, the frequency of mode was given (Figure 4).

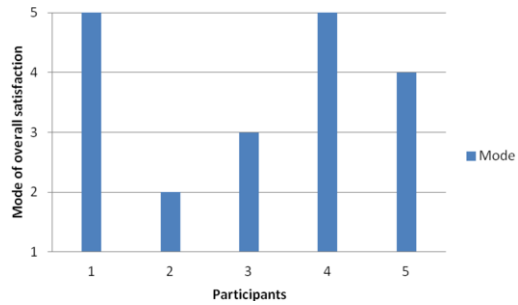

Figure 3- Mode values for each participant over the global satisfaction question in a Likert scale.

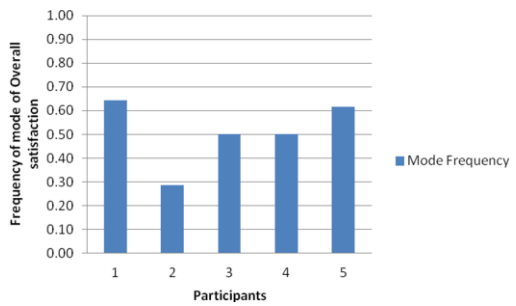

Figure 4- Mode frequency for each participant over the global satisfaction question.

Open questions - propositions

Much feedback was given on the open questions on specific aspects of the system as well as the propositions section. All open responses were grouped into similar comments. Redundant comments were removed and all comments were transmitted to the development team. Frequent comments include:

- Complaints about CBIR performance were frequent as often several irrelevant results were ranked highly;
- Zooming in/out images and level/window settings were considered important additional features;
- Displaying more information about the images in the result lists was also requested;
- Radiology related functionalities (contrast adjusting etc.) were proposed;

Below are some of the comments given in their raw form (translated from French):

- The search for associated articles is interesting at this stage; the search by images would also be useful if visual results were more relevant.
- As a student, search results have to be extremely relevant because we do not have the knowledge to exclude bad images on our own.
- It seems reliable more or less, I feel like it has difficulties distinguishing CT scan images from MRIs.
- More information on the description of images could be interesting to narrow down searches. A zoom in on an image in the 'details' section would be useful.
- There is no text below images in the result list. Difficult to get a good idea of the image «at a glance» when they are small.
- The tool reacts very well to its use, no delay, no bug, tasks we are asked to do are rapidly performed.
- Takes 15 minutes to be comfortable.

Discussion

Lessons learned: user tests

The user tests presented are the first iteration of the user-centered evaluation, so focus was given on evaluating the user test design as well. Research questions have to be clearly defined and evaluation indicators carefully chosen.

One of the main outcomes is that a video tutorial alone is not enough and a user often requires exploring by himself the new functionalities before proceeding to complex information search tasks. This can limit the effectiveness of information finding during the early tasks and makes them less appropriate for performance comparison (text search vs. visual plus text search). For this purpose, the inclusion of a tutorial task after the video may be useful, where users will be asked to perform very simple tasks using the tools.

Some task descriptions and questions of the survey were not completely clear and this may cause misunderstanding results retrieved by the participants. It was also observed that participants often did not read the tasks in full detail and often performed slightly different actions than the ones the task asked. This resulted in inaccurate evaluation of some users. Therefore, to improve accuracy, the task descriptions need to be short and clear and even an additional oral description needs to be given, pointing out the important parts of the task. Misunderstanding will thus be less likely to affect the effectiveness of the participants.

The use of a commercial recording and observation software such as Morae has both advantages and drawbacks. All information that the participant needs for performing the test can be found on his/her screen and no transition to paper is needed. It also provides results in a unified digital format that is easy to transfer to statistical packages to analyze, and present in a meaningful way. It allows indirect observation (as the observer can remotely observe the user's screen and face), which removes some of the subjects' stress of being observed. On the other hand, the use of such a tool increases the hardware and software requirements and is prone to software crashes. Moreover, purchasing a commercial product depends on the available resources. It needs to be noted that all parts of this user test can be performed without the use of such software but doing so will require additional manual work.

A general feeling expressed by a few participants was that they felt they were being evaluated instead of the system. This feeling can affect the subject's behavior, performance, and answers; therefore, this aspect has to be clarified in the introduction.

Lessons learned: system usability

This pilot study was considered partly internal because participants were chosen among those partly known by the interviewers. Therefore, user satisfaction measurements should be taken with skepticism; however, feedback on improvements and proposed additions continued to be fully valid. Main satisfaction tendencies of the system could be observed. Overall, system satisfaction is high as can be seen in Figures 1, 2, and 3, with the majority of participants having a mode of satisfaction measurement above neutral and mode frequency above 0.5. However, there is a clear drop in satisfaction in certain aspects, such as the results quality and presentation (with median 2 and 3 out of 5).

In order to feel confident with the system, it took the users approximately 2 to3 tasks (10 to15 minutes) as it was recorded by the observers and commented by participants. This performance is considered satisfactory with regard to the inexperience of the users with such a novel technique as CBIR. Participants seemed to agree on the learnability aspect (median satisfaction of 5 in two related questions) and seemed generally satisfied with the response time of the system (median satisfaction of 4). The answers on the novelty, usefulness, and intention of uses showed that participants understood the concept of the new tools and the practical usefulness in their workflow (median of 5 out of 6). This was particularly encouraging, considering that the system is still in development and this aspect can be hidden by usability dissatisfaction.

Some participants explicitly complained about the results acquired by mixed queries (text + image example) expecting the system to give results that would correspond more to the text query or the same modality with the query image. This gives solid directions for the next steps of the development process. System bugs, inconsistencies, and usability problems that were identified during these tests were also communicated to the development team. Another interesting finding of this study is that participants were familiar with using advanced query options, such as AND, OR and quotes, and explicitly asked if the system supports these kind of queries.

Conclusion

The design, setup, and results of a pilot usability study for a medical information retrieval system were presented. Most importantly, the lessons learned about the difficulties and design choices of such a study were shared.

An iterative user-centered evaluation can assist in directing the development process towards a system that covers real needs. The user test design depends on the research questions, the available resources, and the development stage. During this iterative process, the study tasks and questions need to be evaluated and refined.

In terms of the evaluation on the system, the feedback was generally positive, but certain aspects were identified to require improvements. Systematic inconsistencies and bugs were discovered. Taking into account these facts, the pilot study

accomplished its goals and obtained encouraging results about the direction of the system development.

Acknowledgments

This work was partly funded by the European Union in FP7 via the Khresmoi project (grant agreement 257528).

References

[1] Müller H, Michoux N, Bandon D, Geissbuhler A. "A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions. Int J Med Inform. 2004; 73(1): p. 1-23.

[2] Aisen AM, Broderick LS, Winer-Muram H, Brodley CE, Kak AC, Paulopoulou C, Dy J, Shyu C, Marchiori A. Automated storage and retrieval of thin-section CT images. Radiology. 2003;(228): p. 265—270.

[3] Vredenburg K, Mao JY, Smith PW, Carey T. A survey of user-centered design practice. In Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves; 2002: ACM. p. 471-478.

[4] Hertzum M. User Testing in Industry: A Case Study of Laboratory, Workshop, and Field Tests. In A. Kobsa and C. Stephanidis (Eds.), User interfaces for all, Proceedings. 5. ERCIM workshop; 1999. p. 59-72.

[5] Kaikkonen A, Kekalainen A, Cankar M, Kallio T, Kankainen A. Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing. JUS. 2005 November; 1(1): p. 4-17.

[6] De Vito Dabbs A, Myers BA, Mc Curry KR, Dunbar-Jacob J, Hawkins RP, Begey A, Dew MA. User-centered design and interactive health technologies for patients. Comput Inform Nurs. 2009 May-June; 27(3).

[7] Fagan JC. Usability testing of a large, multidisciplinary library database: basic search and visual search. ITAL. 2005 September; 25(3): p. 140-150.

[8] Garcia Seco de Herrera A, Markonis D, Eggel I, Müller H. The medGIFT Group in ImageCLEFmed 2012. In CLEF working notes; 2012; Rome, Italy.

[9] Tsikrika T, Müller H, Kahn CEJ. Log Analysis to Understand Medical Professionals' Image Searching Behaviour. In Proceedings of the 24th European Medical Informatics Conference (MIE2012); 2012; Pisa, Italy.

[10] Holzinger A. Usability engineering methods for software developers. Communications of the ACM. 2005; 48(1): p. 71-74.

[11] Bastien CJ. Usability testing: a review of some methodological and technical aspects of the method. Int J Med Inform. 2010; 79: p. 18-23.

[12] Kelly D. Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval. 2009; 3(1 - 2): p. 1-224.

[13] Nielsen J, Molich R. Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people; 1990. p. 249-256.

[14] Nielsen J, Landauer JK. A mathematical model of the finding of usability problems. In CHI '93 Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems; 1993; Ney York, USA. p. 206-213.

[15] Spool J, Schroeder W. Testing web sites, Five users is nowhere near enough. In CHI'01 extended abstracts on Human factors in computing systems, ACM; 2001. p. 285-186.

[16] Woolrych A, Cockton G. Why and when five test users aren't enough. In Proceedings of IHM-HCI 2001 conference; 2001; Toulouse, France. p. 105-108.

[17] Nielsen J. Alertbox. [Online]. 2012 [cited 2012 Dec 09]. Available from: http://www.useit.com/alertbox/number-of-test-users.html.

[18] Markonis D, Holzer M, Dungs S, Vargas A, Langs G, Kriewel S, Müller H. A survey on visual information search behavior and requirements of radiologists. Methods Inf Med. 2012; 51(6).

[19] Müller H, Despont-Gros C, Hersh W, Jensen J, Lovis C, Geissbuhler A. Health care professionals' image use and search behaviour. In Proceedings of the Medical Informatics Europe Conference (MIE 2006); 2006; Maastricht, The Netherlands. p. 24-32.

[20] Beckers T, Dungs S, Fuhr N, Jordan M, Kriewel S, Tran VT. ezDL: An Interactive Search and Evaluation System. Open Source Information Retrieval. 2012; 9.

[21] Müller H, Garcia Seco de Herrera A, Kalpathy-Cramer J, Demner-Fushman D, Antani S, Eggel I. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In CLEF 2012 working notes; 2012; Rome, Italy.

[22] Brooke J. SUS-A quick and dirty usability scale. Usability evaluation in industry. 1996; 189: p. 194.

[23] Chin JP, Diehl VA, Norman KL. Development of an instrument measuring user satisfaction of the human-computer interface. In Proceedings of the SIGCHI conference on Human factors in computing systems, ACM; 1988. p. 213-218.

Address for correspondence

Dimitrios Markonis, Msc,
HES-SO Valais
Rue du TechnoPole 3, 3960 Sierre, Switzerland
dimitrios.markonis@hevs.ch