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 been often proposed to assist medical information search. However, applications of this novel technology have rarely reached the target end users. This study describes the design and setup of performing user tests in order to assess a medical information retrieval system that supports CBIR. Five persons of different levels of medical background participated in the study at the Hospital of Geneva. They were recorded and observed while interacting with the system and provided feedback on the usability of the system. Participants seemed to understand the concept and the practical usefulness of the new tools provided and needed 10-15min to feel confident with the system. The results of this pilot study will be used both for improving the system functionalities and as an input for designing a new iteration of user tests among radiologists.

Keywords: Usability tests, User-centered design, Medical Informatics Applications, Content-based image retrieval.

Introduction

Content-based image retrieval (CBIR) is an information retrieval method that uses a set of images as positive or negative examples and retrieves images with similar visual content from a database. In the early years, it was considered promising for assisting information search in the medical field and several systems were developed [MMB2004]. However, being a rather technology-driven research field, very few applications have reached the end users and never achieved to be fully integrated into the medical professionals' workflow.

User-centered design (UCD) [VMS2002] has been used for several decades in industry [KoS1999, KaK2005], but also in medical informatics applications [DCM2010] and as it is driven by the user requirements and feedback, it is considered to improve the product's usability and the user experience. Some aspects of UCD have also been rarely used for systems supporting CBIR [Fag2006].

UCD in software development includes some key elements in order to involve users' feedback to the design and the development of the application. Firstly, investigation and understanding of the user requirements [MDH2006, TMK2012] should be achieved to identify the general design directions. User-centered evaluation is another important part of UCD which should be performed in the early stages of the development [Hol2005] and should be seen as an iterative process throughout the whole development cycle [KaK2005, DCM2010]. These 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, which include having a number of target end users to interact with the system. Usability of the system is assessed using factors such as learnability, efficiency, effectiveness, memorability and satisfaction [Hol2005]. 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 the common usability testing techniques and tools is presented in [Bas2010] and a more detailed description of aspects that should be taken into account when designing a usability test can be found in [Kel2009].

An important aspect when designing a usability test is the number of participants needed. Early studies have discovered that a single individual is not able to detect all usability issues but 3-4 are sufficient [NiM1990]. In [NiL1993] it is suggested that five users are enough, while studies have questioned this choice of number for some cases [SpS2001, WoC2001]. The number of participants remains an open question, though in [Nie2012] it is explained that five participants are indeed enough for each iteration of an iterative user-centered evaluation.

In this work, 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 in accessing trustable biomedical information to general practitioners, the general public and radiologists. These three main target groups have different search behavior, goals and information requirements, so the system is divided in integrated subsystems, designed to correspond to the target group's needs. Following the same concept, usability tests were designed and conducted separately, concentrating on domain-specific research questions.

This study is focused on the pilot user-tests on the Khresmoi subsystem that was developed to be used by radiologists. The system combines text and CBIR search for finding and navigating through images and articles in the medical literature. The prototype system design was based on the investigation of radiologists' image use behavior and requirements done in [MHD2012] and the development was based on the Parallel Distributed Image Search Engine (ParaDISE) first used in [SME2012] and ezDL [BDF2012].

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

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

The purpose of this first iteration of user tests is an initial assessment of the integrated system, identifying the most important usability issues and missing functionalities as well as redefining the user study protocol for the coming user test iterations.

Materials and Methods

User study protocol

The initial step for designing a user study is to define the research questions that it should attempt to answer as well as the usability aspects to be assessed. Then, appropriate methods for recording the qualitative and quantitative measurements of the usability of the system should be found. Tasks that the participants will be required to perform should be carefully chosen with regard to the research questions and the assessment aspects. Finally, a step-by-step outline of a session should be decided. In this study, a combination of the proposed guidelines of [Kel2008] and [Hol2005] was followed.

The general research questions that the 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 more useful and which tools need to be improved/changed/added?

In this pilot study, in order to assess the usability of the system, the following axes were used as suggested in [Hol2005]: efficiency, effectiveness, learnability and satisfaction. Memorability was not evaluated as this is the first iteration of the evaluation. 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. Assuming the tasks had approximately the same difficulty, comparing the efficiency and effectiveness in tasks with respect to the chronological order of the tasks can be used as an indicator of the learnability of the system. 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 free discussion with the participant were used.

Session outline

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

1. Introduction to the Khresmoi project, the existing Khresmoi search system and the user tests goals (5 minutes)
2. Tutorial video on the system tools and functionalities (3 minutes)
3. Demographics survey (5 minutes)
4. Guided user tests in clear scenarios (30-40 minutes)
5. Survey on the satisfaction with the tools and functionalities (10 minutes)
6. Free possibility to use the system (5 minutes)
7. Survey on the satisfaction with the system, free discussion (10 minutes)

The introduction intended to help the participant understand the concept of the system and 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 being observed by the test facilitator to identify potential shortcomings of the system or the user study design itself. The facilitator was instructed to have a neutral attitude and was 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 they should 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 workflow. 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 [TMK2012] while case-based topics consisted of cases included in an educational database [MGK2012].

Session setup and tools used

For observation and recording purposes, the commercial product Morae[4] mentioned in [Bas2012] was used. Morae allows for screen and face video recording, remote observation and inclusion of introductory text, questionnaires and task descriptions on screen. Also, it is compatible with commonly used statistical packages and presentation software for result analysis and presentation.

A combination of modified versions of the System usability scale (SUS) [Bro1996] and the Questionnaire for User Interaction Satisfaction (QUIS) [CDN1988] was used for the user satisfaction survey forms. Open questions for providing comments on specific aspects of the system and suggestions for improvements and additions were added. In an attempt to get some preliminary answers to the research questions, questions about the novelty, usefulness and intention of use of the tools provided were also added.

The setup of the session included hardware and software preparation but also training sessions of the test facilitators to get familiar with the recording tool, their role and the study purposes. This process also helped in refining the study protocol. The hardware used in each session included two Windows-based computers - one for the participant and the other for the facilitator. The Khresmoi client was downloaded to the participant's computer and Morae software was installed in both computers.

At the end of each session, the file containing the recordings and the answers to the questionnaires and the facilitator's log file with observation notes were acquired. The details of preparation, setting up and running a session were put into a document to ensure that the experiment would be reproduced under the same conditions by new facilitators.

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

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

Results

Demographics

Five persons (2 females, 3 males) took part in two sets of parallel sessions at the hospital of Geneva. All were below 30 years old, with two of them being below 25 years old. Two of the participants had radiology background (one in bone specialization), one was a non-radiology intern and two were students in Medicine. All the participants declared frequent computer use for personal, educational and professional purposes. Three persons answered that they search for medical info more than once per day, one once per day, and one answered once per week.

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 included only the cases that 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 in Medicine) did not select any relevant results for any of the tasks.

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

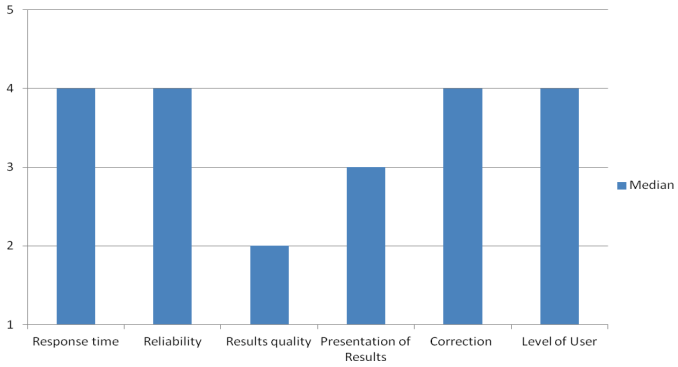

Figure 1- Median values of measuring user satisfaction over specific system aspects in Likert scale (1=strong negative, 5=strong positive).

The median grade for the response time of the system was 4. The same median grade was obtained in the question about the system reliability. In terms of results quality and presentation the median grades were 2 and 3 respectively, while ability to correct one's mistakes using the system and system's design to be used by all levels of users both obtained a median grade of 4.

Questions about the user's intention of use in academic, research and clinical work respectively obtained medians of 4. Finally a question regarding the practical usefulness of the novel features of the system obtained a median of 5 out of 6 due to a design error, so was excluded from the global user satisfaction evaluation.

User satisfaction results over general aspects of the system are presented in Figure 2.

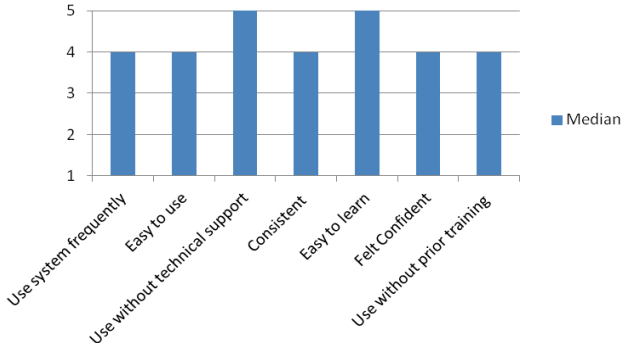

Figure 2- Median values of measuring general user satisfaction about the system in Likert scale.

The median value of grades on the question about intention to use the system frequently was 4. The same median value was obtained for easiness to use and consistency. The median grade for using the system without technical support was 5 and the easiness to learn, too. 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).

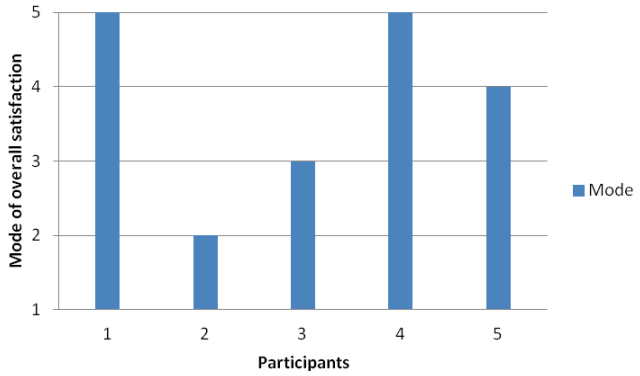

Figure 3- Mode values for each participant over the global satisfaction question in Likert scale.

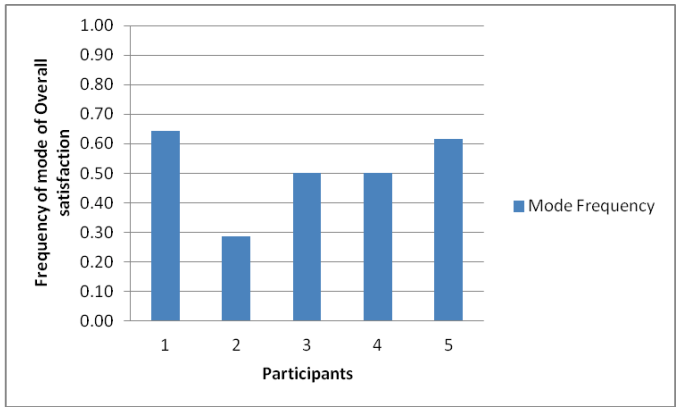

Figure 4- Mode Frequency values for each participant over the global satisfaction question.

Open questions - propositions

Much feedback was given on the open questions on specific aspects of the system as well as the propositions section. Some commonly received comments were:

- Complaints about CBIR performance
- Requests about being able to zoom in the result images
- Requests of displaying more information about the images in the result lists
- Other propositions about functionalities such as backspace usage, radiology related functionalities (contrast adjusting etc.)

Discussion

Lessons learned: user tests

The user tests presented were the first iteration of the user-centered evaluation, so focus was given on evaluating the user tests 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 the user needs to explore himself the new functionalities of the tools provided before proceeding to complex information search tasks. This way the effectiveness of information finding during the early tasks is hindered and makes them inappropriate to use for in tool performance comparison (text search vs. visual plus text search). This was also responsible together with the different difficulty level of the tasks for not being able to measure the learnability of the system. For this purpose, the inclusion of a Tutorial task after the video may be necessary, where it will be asked from the user to perform very simple tasks using the tools.

Some task descriptions and questions of the questionnaires were not completely clear and this caused some errors on the results retrieved by the participants. It was also observed that participants didn't read the tasks in full detail and often performed different actions than the ones the task asked. This was responsible for some measurements failing to accurately represent the participants' efficiency and effectiveness. This indicates that the task description should be short and clear. Even an oral description should be given, pointing out the important points of the tasks. This way misunderstanding of the tasks will 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. For example, all of the information that the participant needs for performing the test can be found on his screen and no transition to paper is needed. It provides results in a unified digital format that is easy to transfer to statistical packages, analyze and present in a meaningful way. It allows for indirect observation (as the facilitator can remotely observe the user's screen and face from his computer) which takes away some of the subject's stress of being observed. On the other hand, the use of such software increases the hardware and software requirements and limitations of a session and is prone to software crashes. Moreover, purchasing a commercial product depends on the available resources. It should be noted that all of the parts of the presented user tests can be performed without the use of such sophisticated software but would require additional manual work.

A general feeling that was expressed by some of the participants was that they felt they were being evaluated instead of the system. This feeling can greatly affect the subject's behavior, performance and answers, so this aspect should be explicitly clarified when the purposes of the study are explained in the introductory speech.

Lessons learned: system usability

This pilot study was considered as partly internal because participants were chosen among acquaintance circles. For this reason, user satisfaction measurements were taken with skepticism, while feedback on improvements and proposed additions continued to be fully valid. Still, main satisfaction tendencies of the system could be observed. Overall system satisfaction is positive as it can be seen in Figures 1, 2 and 3, with the majority of the participants having a mode above average and frequency of mode above 0.5. However there is a clear drop in satisfaction about certain aspects, such as the results quality and presentation (with median values 2 and 3 out of 5 respectively).

Feeling confident with the system took approximately 2-3 tasks for the user (10-15minutes) as it was recorded by the observers and commented by some participants, which is considered satisfactory with regard to the inexperience of the users with novel techniques such as CBIR. Participants seemed to agree on the learnability aspect (median value of 5 in two related questions) and seemed satisfied in general with the response time of the system (median value of 4). The answers to the questions about the novelty, usefulness and intention of use showed that participants understood the concept of the new tools and the practical usefulness in their workflow (median value of 5 out of 6). This was particularly encouraging, considering that the system is still in development stages and these aspects 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 of the same modality with the query image. This gives solid directions for the next steps of the development process. System bugs, inconsistencies and usability issues 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 that kind of queries.

Conclusion

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

An iterative user-centered evaluation can greatly assist on directing the development process towards a system that will cover real needs in an effective and efficient manner. The user tests design and the choice of methods, tools depend on the research questions, the available resources and the development stage. During this iterative process the clarity and usefulness of the study tasks and questions should also be evaluated and refined.

In terms of the evaluation on the system the feedback was generally positive, but also certain aspects were identified to need improvement and system inconsistencies and bugs were discovered. Taking into account these facts, the pilot study accomplished its goals, with encouraging results about the direction the system development has taken.

Acknowledgments

References

[1] Markonis, D., Holzer, M., Dungs, S., Vargas, A., Langs, G., Kriewel, S., et al. (2012). A survey on visual information search behavior. *Methods of information in Medicine* .

Address for correspondence

Dimitrios Markonis, Msc, dimitrios.markonis@hevs.ch