User Evaluation Outside the Lab: The Trial of Físchlár-News

Hyowon Lee[1], Alan F. Smeaton[1], Barry Smyth[2]

[1] Centre for Digital Video Processing, Adaptive Information Cluster,
Dublin City University, Glasnevin, Dublin 9, Ireland
{hlee, asmeaton}@computing.dcu.ie

[2] Smart Media Institute, University College Dublin,
Belfield, Dublin 4, Ireland
barry.smyth@ucd.ie

Abstract: A user study of Físchlár-News system was conducted in Spring 2004 with 16 users, each user using the system for a 1-month period. Físchlár-News is an experimental online news archive that incorporates various automatic content-based video indexing techniques and a news story recommender algorithm to process and index the daily 9 o'clock broadcast news from TV and allows its users to browse, search, be recommended, and play news stories on a conventional web browser. Pre- and post-trial questionnaires, interaction logging and incident diary methods collected both qualitative and quantitative usage data during the trial period. While the details of the findings from this evaluation is reported elsewhere, in this paper we report the details of the methodology taken and our experience of conducting this evaluation.

1. Introduction

In computing or engineering research laboratories, experimental systems are built for the purpose of technical experimentation to measure underlying functionality, or to be able to effectively illustrate the overall approach of a system to other research colleagues. In research labs in the information retrieval field, the "systems" developed are in most cases in the form of software that implements a particular retrieval algorithm taken, and experiments conducted to measure its effectiveness. Because this line of information retrieval systems are "experimental" and thus still at a very early stage of development, naturally the main focus of concern tends to be on the accuracy of underlying retrieval algorithms.

For example, in the TRECVid activity (http://www-nlpir.nist.gov/projects/t01v/t01v.html), the annual video retrieval evaluation forum to promote progress in content-based retrieval from digital video via open, metrics-based evaluation, participating groups from around the world develop their own video retrieval systems and compare the effectiveness of their approach with other participants. In this activity, the developed systems detect and retrieve various features of video content such as camera shot boundaries and panning/zooming, the existence of a crowd, building, or animal. Evaluation is measured in terms of precision and recall, determining the effectiveness of each participant's system. In the task called Interactive Search, each of the participating groups develops a complete system with its retrieval engine and also a front-end user-interface so that a group of users could conduct a search task with the system interactively. Recruited users are given an introduction, training and asked to perform a set of search tasks in a given time limit. Many novel and interesting user-interfaces and their supporting underlying technologies are featured in this activity, opening a whole window of possibilities for video information retrieval systems for end users. Although real users are involved to interact with the system thus allowing the experimenters to be able to collect various data including pre- and post-session questionnaires, interaction log data and users' search performance, this type of evaluation is highly artificial for well-known reasons including the controlled environment and the fact that these capture only the very first impressions of the users.

In the real world users use information systems in their own home and workplaces, over varying time spans and in varying contexts: frequent interruption during the task is a normal phenomenon; background noise, telephone ringing, swapping among different systems (applications) - for checking emails, for copying text from a word processor, for getting author information from a website, etc. - are the *main characteristics* of their work. An information system's true value cannot be evaluated in an isolated lab, safely cleared of these "distracting" factors. If we are to evaluate a system and see how usefully, effectively, efficiently and easily it supports the particular information access it was designed to support, we need to bring the system out to the real world, to somehow integrate it with other technologies in use, and get users use it in their own environments and context.

With the recent emphasis on the user side of the system in the information retrieval field, with the recognition of the importance of Human-Computer Interaction and usability in the system development process, and with the awareness of the social impact of new technology, quite a few user evaluations are now moving out of their labs and are conducted in the users' natural, real environment. These new lines of study, often focusing on a small number of users but each user's case investigated in minute detail, tend to last long enough to be able to witness the changing nature of the use, and are highly qualitative rather than quantitative in the nature of their inquiry. They try to gain understanding and insight the newly developed information system in use and the circumstances in which it is being put to use, which will guide further improvement, re-direct their overall approach and re-evaluate the priorities within the project.

Following a similar philosophy, we have conducted a user evaluation at a work place in 2004. The system evaluated was Físchlár-News, a web-based TV news archive that incorporates various content-based video technology to automatically capture and analyse the daily broadcast TV news and prepare an easily searchable, browsable, and playable web interface for its users. Sixteen users freely used the system on their own PC at their workplace for a 1-month period in their own free time, in their own way, and for their own benefit. During the 1-month period, we captured usage data in the form of questionnaires, interaction logging and an incident diary, resulting in a large amount of qualitative and quantitative data on users' real usage. In this paper we report the details of this trial and the methods of evaluation we used.

The paper is organised as follows: in Section 2, a review of similar user evaluations are described that adopt a long-term and qualitative approach conducted at users' place. In Section 3, the system under evaluation, Físchlár-News, is briefly introduced. Section 4 describes the details of the evaluation: the goals, experimental setup, the procedure taken, and a brief summary of the analysis and findings. Section 5 comments some of our thoughts after having conducted this evaluation, and finally Section 6 concludes the paper with a call for more user evaluations of this type.

2. Long-Term User Study outside the Lab

In the early development stage of an information system or product, an evaluation task focused on the understanding of its use in the context of users' work is important, often in the form of an ethnographic study. The main focus is usually not to prove/disprove a very specific hypothesis but to learn the overall and specific circumstances where the system is used from which some hypotheses might be thought of for further investigation. The starting approach of this type of evaluation is usually quite open. The style of these evaluations, having its origin from social sciences, can be quite alien to the research community in computing and information science and have not yet been adopted as a major evaluation type even though the concept of inclusion of the user in the whole system diagram has been adopted. However, more commercially-oriented studies have also been started at the potential consumers' home and workplace environment and there is growing number of studies conducted in this fashion.

In [O'Brien *et al*. 1999], a usage and its context of usage of a set-top box was investigated in 11 households, with the focus on gaining deeper understanding of how the technology device is integrated into the ordinary domestic environment. The study used interview/discussion with the family members and observation occurred during home visits, and highlighted the importance of the management issues of a technology product (e.g. who uses when, guidance from parents, and payment methods) for it to be successfully integrated into a family environment.

In [Lee 2000] the use of a Web TV set-top box was investigated in 10 family homes. Each family was visited 4 times throughout a 1-month period in which naturalistic observation and semi-structured interview, diary-keeping and video recording were used as data-gathering methods. The findings highlighted the social aspects among the

family members in the use of the device and different life-stages of the families, and useful design implications were suggested including the need for user guidance to ensure that conflict of usage time among family members is minimised.

In [Perry *et al*. 2001] the use of mobile devices that support work (mobile phones and laptop computers) was studied with 17 types of mobile workers (including sales staff, consultants, medical workers, civil servants and media) who were interviewed before and after their travel as well as providing travel diaries and an inventory of their documents and communications. The highly qualitative analysis showed the actual and detailed circumstances where the mobile workers were supported by the devices they brought, including using a mobile phone to ask co-workers at home office to fax a material to a customer. They provided design implications for future mobile device developers from this study.

In [Petersen *et al*. 2002] two families were visited four times with an interval of 1month between each visit, to see their natural *development of the use* over time of a new TV set with an integrated video recorder. Observation, interviews and video recording were used at each visit, and the changes in their use from previous visits were highlighted.

As mentioned earlier, these studies are highly qualitative in the nature of their inquiry with a relatively small number of users but each case is investigated in depth, and aim to find the intricate nature of their new products' usage in a real environment, the social aspects when many people are to use it together, and changing usage patterns and their opinions. With the appearance of new information technology and the fast pace with which it becomes wide-spread, more and more studies are focusing on this aspect of new system use. Having higher ecological validity than the lab-based studies and more potential to find out unexpected use and problems, in this line of qualitative studies "the results are not numbers but understanding, not percentages but insights," as described as the main characteristic of the "new computing [Shneiderman 2002, p238]" and similarly a core concept of the "new usability [Thomas and Macredie 2002]."

In our work, we take this approach in our user evaluation, but with a difference that the product in concern is an experimental, although complete, system whose constituent technologies are still at research and development stage. Our aim in conducting this evaluation was to see how the currently imperfect technology, when it becomes more reliable in the near future, could be made into a stable product and be used by normal users, integrated into their real environment. In this sense, we were trying to jump ahead the current stage of research and peep into the situation where these research outputs could be applied in the near future. Sixteen users were visited in their natural working environment and use of a web-based, video news archive system was then monitored for a 1-month trial period by various observation methods. The next section briefly describes the system under evaluation and Section 4 details the evaluation conducted.

3. Físchlár-News

Físchlár-News is one of the series of digital video experimental systems developed within the Centre for Digital Video Processing at Dublin City University. The system represents an accumulation of some research developments within our Centre and incorporates them as a main part of its functionality, demonstrating functional output of these research streams and at the same time demonstrating how these could make a complete, usable system.

The system records daily 9 o'clock broadcast TV news along with its Closed Caption text and automatically indexes and archives this into a news stories database with which users can search, browse, play and recommended news stories they are interested. The system is fully-automatic, incorporating underlying technology including Shot Boundary Detection and keyframe extraction [Browne *et al*. 2000], news story segmentation by anchorperson detection and shot clustering [O'Hare *et al*. 2004], and automatic story recommendation based on collaborative filtering [O'Sullivan *et al*. 2004]. Playback uses streaming technology from an Oracle Video Server, and requires a plug-in on the web browser. The system has been engineered to be able to easily plug-in a new development and upgrade, and uses MPEG-7 in its internal exchange of data implemented with the Cocoon XML publishing framework. More background and technical details of the system can be found in [Smeaton *et al*. 2004].

The system has been deployed within the campus of Dublin City University since 2001, and has seen incremental developments, resulting in user-interface feature refinement at each stage of upgrade. Since April 2003 the system has allowed users to access news video by individual story units for searching, browsing, playing and recommendation. During the evaluation period reported in this paper, the system had more than 4,000 news stories available (April 2003 – March 2004) for users. The user can select a date from a calendar feature to view that day's news summary, see the details of each story with the full transcript of what was spoken along with images extracted from each shot, search the transcript of the whole archive to find a story of interest by typing in query terms, play a story or any part of a story, browse a list of related stories to jump to other similar stories or to a thread of development of a topic appeared in the past. The user can also indicate a rating for each story while browsing, which will be used to locate other news stories past and current that the user might be interested in, and be presented with them as recommended stories for further browsing and playback.

In summary, the system automatically converts a daily broadcast news video into a collection of news stories all easily searchable, browsable and playable in a way a conventional video access cannot provide. There is no such system in the market yet as the underlying technology for automatic, content-based video analysis is not mature enough to ensure reliable indexing.

4. User Evaluation

With the Físchlár-News system already deployed with a number of users, we conducted a user evaluation that spanned about two months in 2004. In this section we describe the detail of the approach and the procedure we have taken in the evaluation.

4.1 Experimental Setup

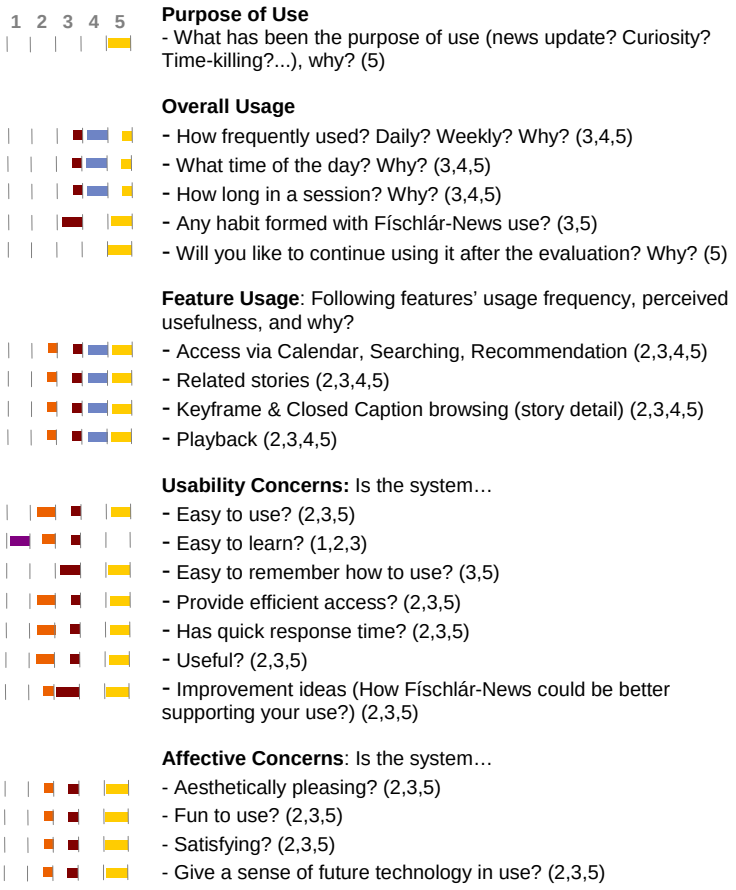
The high-level goals of the evaluation were:

(1) To draw the picture of real usage of the system;
(2) To find out the system's usability and acceptance level by users over time;
(3) To identify how the system could be further improved.

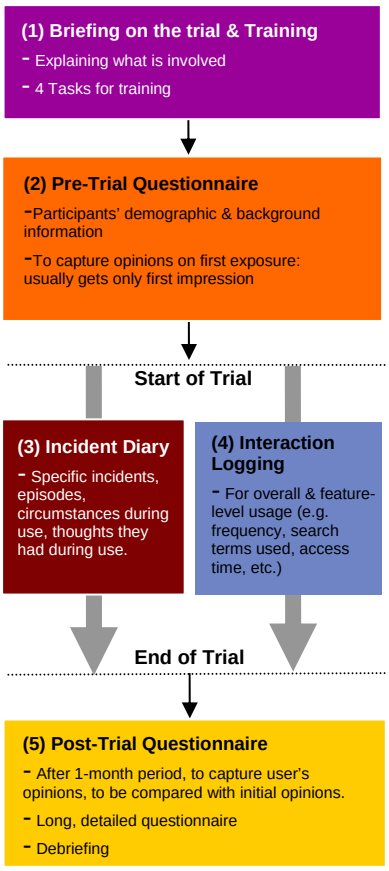
In establishing these goals, we took a more open approach to find usage patterns or trends that might emerge by collecting and looking into detailed daily usage, rather than setting a specific, single hypothesis to prove or disprove - this is one of the major differences this line of qualitative studies has from the traditional TRECVid-style user experiments.

We listed detailed questions we wanted answers to, from which any usage patterns/trend could be revealed, for example, what our users thought of the related stories feature, how frequently they played a clip and how long they watched a news story. The full list of questions is listed in Figure 1 (a). For each of these questions, we decided what evaluation method or what combination of methods would be most suitable to answer the question, that is, the information gathering method – for example, a questionnaire method to answer what users thought of the related stories feature, interaction logging to answer how frequently they played a clip and how long they watched. Because some questions were only answerable by getting users to use our system from their own desk and in their own time, setting up a field trial period has become an important part of this evaluation. The kinds of evaluation method and the procedure taken is illustrated in Figure 1 (b).

The questions in Figure 1 (a) can be answered by one or more than one method in Figure 1 (b), and in most cases the answers can be found by some combination of the methods rather than a single method potentially providing stronger evidence or a different perspective on the same finding. The numbers in the brackets at the end of each question indicate the evaluation methods in Figure 1(b) that we thought would answer the question. On the left side of each question in Figure 1 (a) we colour-coded the relative contribution of each of the methods in answering the question. Our expectation was that if we find out all answers to these questions, we will be able to draw the real usage picture of the system, users' opinions after using it in their own context, and to find out how the system can be improved, thus achieving the three high-level goals mentioned earlier. The details of each of the evaluation methods in Figure 1(b) are described in the following subsection.



(a) Questions to be answered



(b) Evaluation procedure

Figure 1: Full list of questions and the evaluation procedure/methods to answer these questions

4.2 The Procedure Taken

We started with two pilot users recruited from within our research Centre but who were not directly related to the development of the system. The pilot users were administered exactly same procedure as the rest of the main users, but started 3 weeks earlier to troubleshoot any flaws or practical problems that we might have overlooked. Real users were recruited by visiting Postgraduate Research Laboratory in the School of Computing, where about 60 research students have their own allocated desk, PCs and other equipment.

INTRODUCTION AND BRIEFING

One evaluator dealt with all users one by one. The evaluator visited a prospective user at his/her desk and introduced Físchlár-News on his/her own web browser, explaining each feature, installing the video streaming plug-in if not installed already, and explaining how the system works if s/he showed interest in the technical aspect. In most cases s/he liked the idea of being able to find and play last night's TV news on his/her own PC at their workplace, many of them expressing much interest in its technology as well as in use. A total of 25 prospective users were visited one by one this way.

For four people whose PC could not access Físchlár-News properly (e.g. using the Linux operating system on which Físchlár-News is currently unable to stream the video), the session finished at this point: to them it was just an interesting introduction to Físchlár-News. For the rest of the people whose PC could enable all features of Físchlár-News properly, the evaluator introduced the trial period and the activities involved, and asked his/her interest in participation. Everyone (21 in total) who was asked for this replied positively. S/he was asked to read and sign a 1-page informed consent form which was clearly and concisely written about the user's rights to ask

questions and to terminate his/her trial at any point without any penalty, and about the evaluator's responsibility to answer any questions and assist the participants. It took approximately 30-40 minutes up to this point.

TRAINING WITH SHORT TASKS

The user was then asked to conduct a series of short tasks with the system. Four tasks were given one after another, covering the major features of the system and in the order as to disambiguate any possible confusion of use. This was to make sure the user knows how to use the system by him/herself, and provide a kind of training. While the evaluator assisted if s/he was stuck, mostly s/he was encouraged to explore and discover features him/herself. The following are the 4 tasks:

1. Find recent news on Michael Jackson's child molestation charges and see what happened, then rate the stories;
2. Find the news about Special Olympics on 27 June 2003, then rate the stories;
3. Check your recommended stories for today, then rate the stories;
4. Watch the video clip of Saddam Hussein body check-up by a doctor just after his capture, then rate the stories.

The evaluator asked each task verbally, then watched the user conducting the task. By the time the user completed the 4th task and was playing a video clip showing Saddam Hussein's body check-up, it was notable that s/he was realising what the system provides and what s/he can do with the system. The task completion provided an important motive in which the user was able to relate the system to his/her own interest or possible benefit. This part took approximately 10-20 minutes.

PRE-TRIAL QUESTIONNAIRE

After the task completion, the user was given a pre-trial questionnaire (designed to answer some of the questions in Figure 1 (a)), to be returned any time on the following day. The questionnaire included questions on the user's demographic details (age, research area, news watching habits, etc.) and detailed likert scale questions asking user's opinions on different features of the system. This was equivalent to questionnaires often administered by the user evaluations of information retrieval systems – capturing the first impression of the system after exposure to the system for 2-4 hours and having conducted a few, focused tasks. In our 1-month trial evaluation, however, this was only the start: a separate, post-trial questionnaire would be administered at the end of the trial to capture the impression of the system after they have used the system for a month.

INCIDENT DIARY

To be able to capture the context and circumstances when the user accesses the system in their normal daily activity, an incident diary was given to the user. As an indirect way to obtain the usage of the system in a long-term evaluation, the incident diary is a very useful method that supplies an account of the reasons for what the user did, complementing the interaction log data. The diary was a small, hardcover notebook with 50 pages ring-bound to be filled in by a user whenever s/he uses the system, about any problems encountered, any success in looking for news, or in general any ideas s/he came across during use. An alternative was the use of an online logging method but because that meant opening another web browser along with Físchlár-News on the screen we believed this would prevent the notetaker from easily drawing their ideas or mixing text and drawing, and so we settled for a paper-based diary. We considered the diary as one of the major sources of usage data collection, and much consideration was given to how to encourage the user to write this as often and as much as possible. While a high-quality, feel-good notebook was chosen to give positive feeling about using the dairy, a more practical incentive was a "small gift" promised to each user if s/he fills in more than half of the diary by the end of the 1-month period. Every user's reaction to this offer was either smile or laugher, and all liked the idea of the possible gift at the end of the trial period. It was expected that, along with the post-trial questionnaire that would be administered at the end of the trial, the diary would provide rich contextual data at the time the user was actually using the system. Other user evaluations that used a diary method can be found in [Preece *et al*. 2002, p377].

1-MONTH TRIAL PERIOD

After the user is given the pre-trial questionnaire and the dairy, the 1-month trial period started. The evaluator left the user and checked the interaction log data for that user, to note the exact time at this point. The official interaction logging time started 2 hours after the evaluator left the person: this was to exclude from the interaction log data the 4 tasks conducted with the evaluator and also probable experimental access by the user just after the

evaluator left him/her out of curiosity or further testing. For most users, the first use was the following day when they received the first email reminder (email reminder will be explained below).

The crucial thing during the 1-month period was to make sure the system was working without any failure at all times, that the daily recoding and news story indexing is done accurately, the system web server running OK, and that the users have not forgotten about the system.

For our users, there were three things during the 1-month period that reminded them to use the system: the diary sitting on their desk; regular but short visits by the evaluator during the trial to ask if everything is okay and to briefly discuss their more recent experiences (at least 3 times per user); and a daily email sent to the user with the top 3 news stories summarised and a hyperlink to the system website to allow an easy access to the system from the email. A daily email reminder was a part of the system that provides a selective summary of the day's news as well as a convenient link to the system website.

DEBRIEFING AND POST-TRIAL QUESTIONNAIRE

One day before the end of the 1-month trial period of each user, s/he was visited by the evaluator and discussion on his/her 1-month experience with the system was conducted. Just before ending this discussion, s/he was given a post-trial questionnaire, to be returned on the following day. The questionnaire contained some of the questions listed in Figure 1(a) and had overlapping questions with the pre-trial questionnaire, but as this one was filled in *after* the 1-month use, this captured not only the user's impression at first exposure and thus was weighted higher in the analysis.

FOLLOW-UP QUESTIONNAIRE

One week after the end of the trial, a short follow-up questionnaire was composed and administered to the users by email, to further clarify some of the answers users gave in the post-trial questionnaire and diary comments. This was the official end of the data collection for this evaluation.

SUMMARY OF THE DATA GATHERING METHODS USED

In summary, for each user following data were obtained:

- Pre-trial questionnaire (first exposure to the system)
- Incident diary (during 1-month trial)
- Interaction log data (during 1-month trial)
- Post-trial questionnaire (at the end of the trial)
- Follow-up questionnaire (1 week after the end of the trial)

In addition to the above, other smaller, informal methods included feedback obtained from the occasional chat at the users' desks during the 1-month trial period. This is illustrated in Figure 2.

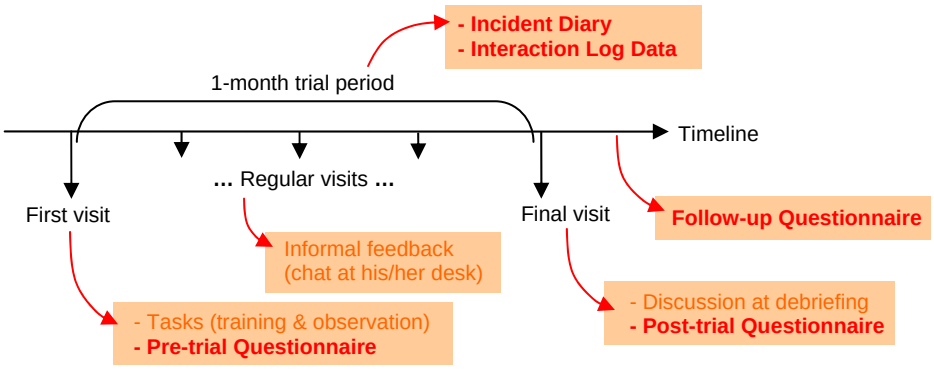


Figure 2: Summary of data gathering methods by period

4.3 Collected Usage Data and Analysis

Figure 3 shows a 3-month timeline in 2004 and the 1-month trial period of each of the 21 users.

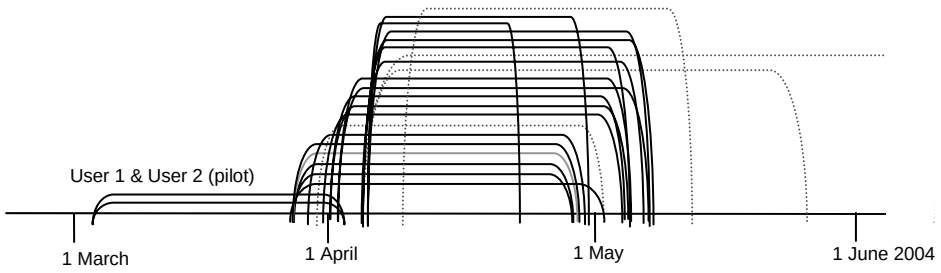


Figure 3: Trial period for the 21 individual users

From the collected 21 sets of pre-/post-trial and follow-up questionnaires, interaction logging data, and incident diaries, 5 users' data were discarded due to extremely low frequency of usage (less than twice during the 1-month period), one of them due to a technical problem in streaming the video, and 4 of them due to particular work pressure during the trial period, and these 5 users' periods are drawn in dotted lines in Figure 3. Having little interaction with the system during their daily work environment, we considered the usage data and comments from these users would not be suitable in capturing the use and opinions coming from a long-term usage. Consequently, the data analysis was carried out with the remaining 16 users' usage data.

The 16 users were all postgraduate students in the School of Computing, ages between 20 – 34, of which 9 were male and 7 female. Thirteen users said their main source of keeping up-to-date with current affairs was browsing Internet news, some of them also TV news, 4 from newspaper reading, and 2 said radio was their main source of news. The users accessed Físchlár-News in their own time for their own reasons throughout the trial period. The following is an overview of their usage data:

- Total access: **149** times (average 9 access / person)
- Average time spent / access: **14.1** minutes (assuming session is terminated when a user becomes inactive for 20 minutes)
- Story detail viewed **251** times (average 15.7 / person)
- News stories played **376** times (average 23.5 / person)
- Closed Caption text search **72** times (average 4.5 / person)
- Story rating **327** times (average 20.4 / person)
- Story recommendation list viewed **80** times (average 5 / person)

Figure 4 shows the access dates and frequencies of individual users throughout the trial period.

A total of 142 pages of diary comments were collected, which were transcribed and each sentence categorised into the emerged themes. These comments along with the questionnaire comments were collated and related to the interaction log data. This provided very rich, qualitative usage data. Some of the findings from the data analysis include:

- The main purpose of the access was to get up-to-date news during free time at work, such as in the early morning just before starting the day's work, just after lunch, or just before finishing work. Due to this, Físchlár-News was often used together with user's favourite news websites such as BBC news (http://news.bbc.co.uk/) or Unison (http://www.unison.ie/);
- The perceived strength of Físchlár-News was its playback of news stories, and much praised by the users and was the most

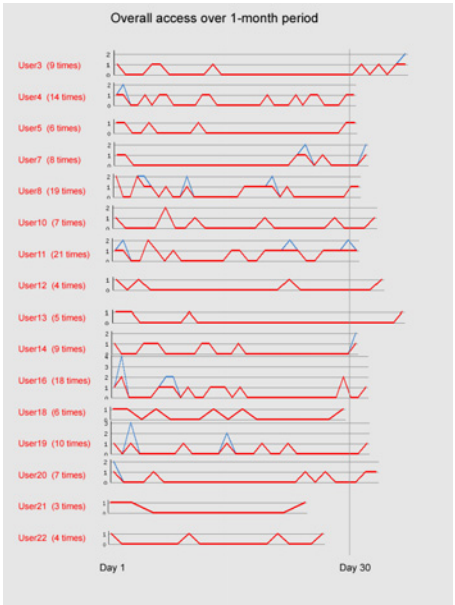


Figure 4: Access throughout 1-month period

frequently used feature of the system;
- Limited news coverage of the RTE 9 o'clock news of Físchlár-News was a major point of complaint, causing the users to prefer other news websites;
- Our users wanted to access news stories by broad categorisation such as politics, entertainment, international and sports, as they do in reading newspapers and browsing online news websites.

Full details of the usage analysis are published elsewhere.

5. Discussion

In conducting this evaluation, many aspects of what we could consider as the characteristics of user-oriented evaluation have been noticed. Dealing with real users was a time-consuming task, as their circumstances and availability are all varied and we did not force them to make any prior appointments for the sake of our evaluation. There were times when the evaluator had to visit a user's desk for 4 consecutive days to actually meet him once; visiting 3 consecutive times just to get back the completed questionnaire was common; absence, late response, repeated failure to return the questionnaire were the normal interaction with the users – this is understandable considering the fact that to our users there were more important things (their own work and business) to do during the day and the tasks involved in this evaluation or indeed accessing Físchlár-News was only a minor sideline. Keeping the evaluation task as minimal as possible on the users' part (thus bothering our users as little as possible during the trial period) while at the same time trying to extract as much data as possible from them was a matter of delicate trade-off and required careful judgement.

Constant note-taking was required throughout the trial period to ensure that all small talks, discussions, complaints, minute details surrounding the comments, each user's overall aptitude and affective tendency towards computer use in general and Físchlár-News in specific, were recorded so that the understanding of the system usage could take these factors into account. The notes taken after each visit, comments in the questionnaires and diaries meant an overwhelming amount of qualitative data, requiring connection to the more-straightforward, easier-to-analyse quantitative data mostly from interaction logging. At the data analysis stage, each of the 16 users' data had to be looked at case by case, which was highly time-consuming.

Finally, we were ambitious conducting this evaluation in the sense that the system deployed for the evaluation was still an experimental system with its constituent technologies at varying stages of research and development. For example, shot boundary detection is relatively mature in the field and achieves accuracy of over 95% for hard cuts; news story segmentation is more difficult with various alternative methods proposed and experimented today, with varying accuracy of 40 – 80%. When these constituent technology are not yet perfect, users will notice these – in their daily news story list and in their browsing screen – and their perception and comments will be clouded by poor functional effectiveness instead of help revealing more useful usage concerns. For this reason, during this user evaluation the poor performance of news story segmentation had to be covered by the evaluator's daily checking and subsequent manual correction if required. Thus, in addition to the costly nature of dealing with users and obtaining and analysing qualitative data, this was another cost in our attempt to peep into the future usage of a new system by constructing a complete system with some still immature underlying technology.

However, we believe the cost paid off. For one thing, the considerable amount of qualitative data collected could be, in a way, referred to as our users' "wish list" of things that can be further improved in Físchlár-News, answering the 3rd goal we have set out to achieve (see the start of Section 4.1). After all, a product should be improved by using conventional usability engineering. But equally importantly, we were able to see how a combination of technologies that are still at the research stage could be put together and perceived by real users who are busy, complaining, judgemental, and enthusiastic when the system looked futuristic and fancy but highly intolerant when the system is in the way of what they want to do. This re-adjusts our priorities of what we research and develop, and re-directs and re-focuses future direction of our work.

6. Conclusion

In this paper we described our experience of conducting a 1-month trial to be able to observe our users' experience with our system for accessing broadcast TV news. Applying usability engineering methods from the HCI field, careful planning and constant observation and interaction with the users, were the important factors in successful completion of an evaluation of this size. We hope more diverse and specific guidelines and methods

for this type of evaluation will become available for researchers developing systems that allow information access for users, and we also hope that our study reported in this paper is of help in shaping such guidelines and methods.

As we already know and also experienced in this study, user evaluation outside of lab is difficult in terms of the required robustness of the system and is costly in terms of preparing and managing the evaluation process. However, the results we get from such a study can reveal important directional clues and plausible visions of where the technology we are in the middle of researching might be valuably applied and used in the real world.

Acknowledgements

Part of this material is based on work supported by Science Foundation Ireland under Grant No. 03/IN.3/I361. The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

References

Browne, P., Smeaton, A.F., Murphy, N., O'Connor, N., Marlow, S. and Berrut, C. (2000) Evaluating and combining digital video shot boundary detection algorithms. *Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP 2000)*, Belfast, Northern Ireland, 31 August - 2 September 2000.

Lee, W.O. (2000) Introducing Internet Terminals to the home: interaction between social, physical, and technological spaces. *Proceedings of the 14th Annual Conference on HCI*, pp119-132.

O'Brien, J., Rodden, T., Rouncefield, M. and Hughes, J. (1999) At home with the technology: an ethnographic study of a set-top-box trial. *ACM Transactions on Computer-Human Interaction*, **6**(3), pp282-308.

O'Hare, N., Smeaton, A., Czirjek, C., O'Connor, N., and Murphy, N. (2004) A generic news story segmentation system and its evaluation. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Montreal, Quebec, Canada, 17-21 May 2004.

O'Sullivan, D., Smyth, B., Wilson, D., Mc Donald, K. and Smeaton, A.F. (2004) Improving the quality of the personalized electronic program guide. *Journal of User Modeling and User-Adapted Interaction*, **14**(1), pp 5-36.

Perry, M., O'Hara, K., Sellen, A., Brown, B. and Harper, R. (2001) Dealing with mobility: understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction*, **8**(4).

Petersen, M., Madsen, K. and Kjaer, A. (2002) The usability of everyday technology - emerging and fading opportunities. *ACM Transactions on Computer-Human Interaction*, **9**(2), pp74-105.

Preece, J., Rogers, Y. And Sharp, H. (2002) *Interaction Design: Beyond Human-Computer Interaciton*. John Wiley & Sons.

Shneiderman, B. (2002) *Leonardo's Laptop: Human Needs and the New Computing Technologies*. MIT Press, Cambridge.

Smeaton, A.F., Gurrin, C., Lee, H., Mc Donald, K., Murphy, N., O'Connor, N., O'Sullivan, D., Smyth, B. and Wilson, D. (2004) The Físchlár-News-Stories System: personalised access to an archive of TV news. *RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, Avignon, France, 26-28 April 2004.

Thomas, P. and Macredie, R. (2002) Introduction to the new usability. *ACM Transactions on Computer-Human Interaction*, **9**(2), Special Issue on the New Usability.