Comparing Fact Finding Tasks and User Survey for Evaluating a Video Browsing Tool

Werner Bailer and Herwig Rehatschek*
JOANNEUM RESEARCH, Institute of Information Systems
Steyrergasse 17, 8010 Graz, Austria
werner.bailer@joanneum.at

ABSTRACT

There are still no established methods for the evaluation of browsing and exploratory search tools. In the (multimedia) information retrieval community evaluations following the Cranfield paradigm (as e.g. used in TRECVID) have been widely adopted. We have applied two TRECVID style fact finding approaches (retrieval and question answering tasks) and a user survey to the evaluation of a video browsing tool. We analyze the correlation between the results of the different methods, whether different aspects can be evaluated independently with the survey, and if a learning effect can be measured with the different methods. The results show that the retrieval task correlates better with the user experience according to the survey than the question answering tasks. It turns out that the survey rather measures the general user experience while different aspects of the usability cannot be analyzed independently.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*; H.5.2 [**User Interfaces**]: [Evaluation/methodology]; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Human Factors, Measurement

Keywords

video browsing, evaluation, user study

1. INTRODUCTION

With the increasing amount of multimedia data being produced, there is growing demand for more efficient ways of supporting exploration and navigation of multimedia data. Multimedia content abstraction methods are complementary to search and retrieval approaches, as they allow for exploration of an unknown content set, without the requirement to specify a query in advance. Interactive video browsing tools are examples of using multimedia content abstraction techniques for exploring large sets of video data.

As pointed out in [11] and also evident from [13] evaluation of browsing and exploratory search tools is still an open issue. In the information retrieval and multimedia information retrieval community evaluations following the Cranfield paradigm, which is also used by the TREC (text) and TRECVID (video) [9] retrieval benchmarks, has been widely adopted. This type of evaluation is system or component centric and answers a well defined information need, i.e. question answering or fact finding. In browsing or exploratory search the user's information need may not be well-defined [12]. User centric evaluation methods such as surveys take the context of the user's task when using a system into account. The issue of limited correspondence between these evaluation methods has been discussed for information retrieval systems [10].

Most of the literature on evaluation of exploratory search deals with text documents. In the multimedia domain rather evaluation approaches for summarization and skimming systems can be found, which often deal only with single multimedia documents but not with collections. The following classes of evaluation approaches have been proposed (of course combinations of the methods from different classes are sometimes used).

Survey The users are asked about their experience with the tool, their satisfaction with the results and the relevance of certain features of the tool (e.g. [7]). This type of evaluation does not require any ground truth or specific preparation of a data set.

Analysis of system logs This approach uses either server-side logs [1] or specific client applications that log user actions [4]. The main advantage is that evaluation does not interfere with the user's usage of the system and that the approach can be used for long-term studies. However, comparison across different types of tasks and systems might be difficult.

Question answering Users are asked fact finding questions about the content in order to evaluate whether they have found the correct segment of content or were able to extract information from the collection of multimedia documents (e.g. [5]). The questions can be open or in the form of a multiple choice test (quiz).

*Herwig Rehatschek is now with the Medical University of Graz.

MM'09, October 19–24, 2009, Beijing, China.

The correctness of answers to open questions needs to be checked by a human, while multiple choice tests can be very efficiently evaluated once the ground truth for a specific data set has been created.

Indirect evaluation The user performs a task using the tool or system. Based on the success of this task the effectiveness of the tool can be measured. The task can for example be a content retrieval task [2] or gathering information from a meeting archive browser [6]. Once ground truth for theses tasks has been created the answers can be checked automatically.

The aim of this paper is to apply different evaluation approaches to a video browsing tool and compare the results. As the focus is on the comparison of evaluation methods and the same browsing tool is used in all the experiments. Section 2 describes the research questions and the evaluation procedure, Section 3 presents the obtained results and Section 4 concludes the paper.

2. EVALUATION

2.1 Research Questions

Given the fact that there is no established method for the evaluation of multimedia browsing we have chosen to apply both TRECVID style approaches as well as a survey taking the user experience into account and intend to compare their results. The *retrieval tasks* contain a well defined need for video clips (motivated by scenarios in multimedia post-production), the *question answering tasks* are more goal oriented, leaving the description of the needed video clips more fuzzy. The *survey* asks about the experiences of the users when completing the two types of tasks. In particular, we want to answer the following questions:

Is there a correlation between the results of the retrieval and question answering tasks and the assessment of the users in the post-task questions of the survey? We want to know whether the results from the different approaches yield similar or complementary results.

Can the post-task questions of the survey be answered independently? Surveys aim at giving a more holistic picture of the user experience. We are thus interested whether the questions asking about different aspects of the tool can be treated independently.

Is there a learning effect? Is it evident in the task results that users achieve better results when using the tool for some time, and does that correspond to the user's experience as expressed in the survey?

2.2 Video Browsing Tool to be Evaluated

The application scenario is that of content management in the post-production phase of audiovisual media production, where users deal with large amounts of unedited audiovisual material. A detailed description of the video browsing tool being evaluated can be found in [3]. The basic workflow in the browsing tool is as follows: The user starts from the complete content set. By selecting one of the available features the content is clustered according to this feature. Depending on the current size of the content set, a fraction of the segments (mostly a few percent or even less) is selected to represent a cluster. The user can then decide to select a subset of clusters that seems to be relevant and discard the others, or repeat clustering on the current content set using another feature.

2.3 Materials and Procedure

The survey is independent of the data set used. For the retrieval and question answering tasks two data sets are used. The TRECVID BBC Rushes 2006 data set is the one used in the TRECVID 2006 rushes exploitation task and consists of about 25 hours of rushes of travel documentaries (in French). The SEMEDIA data set is a part of the data collected by BBC[1] and CCMA[2] in the context of the SEMEDIA project[3] and consists of about 10 hours of edited news stories and complete news, sports and talk show programs (in English and Catalan).

Each of the retrieval tasks consists of a one line description of the synopsis of a video segment. The task is to use the browsing tool to locate all segments that match the given textual description. The results are collected in the result list of the browsing tool and saved at the end of the task. The result lists are then matched against a list of ground truth segments created before.

Each question answering task is a multiple choice question with six statements of which one or more are true. The question is a description of a scene, where each of the options are a statement about the scene. The questions are chosen so that they share the set of relevant video segments with a corresponding retrieval task.

The survey consists of 3 questionnaires: The pre-test questionnaire is completed once for each individual user who takes part in the evaluation before they are trained in the use of the browsing tool. The post-task questionnaire is completed after each task that each user finishes during the experiment. The questions of the post-task questionnaire are listed in Table 1. The post-test questionnaire is completed once for each participant after completing the last task. The questionnaire is largely based on the one used for the TRECVID 2004 interactive search task [8]. Some questions that were too specific to retrieval systems have been discarded and two questions specific to video browsing have been added to the post-test questionnaire.

The retrieval and question answering tasks as well as the questionnaires are available at `http://semedia.joanneum.at/resources/`.

The evaluation session starts with an introduction of the browsing tool and an explanation of the evaluation procedure. Then the users have 10 minutes time for getting accustomed to using the browsing tool. Before starting to work on the tasks the users complete the pre-test part of the survey.

One evaluation session consists of a sequence of 4 retrieval tasks or a sequence of 4 question answering tasks. The participants are evenly divided into 4 groups with varying assignment of task types and data sets in order to avoid an effect of the order on the results. The working time for one task is 10 minutes including the time to complete the post-task questionnaire for each task. The users are allowed to ask staff for technical support about the use of the tool during the evaluation.

After the 4 tasks the users complete the post-test part of the survey. The total time for the session is thus about 60

[1] `http://www.bbc.co.uk`
[2] `http://www.ccma.cat`
[3] `http://www.semedia.org`

TVB1	I was familiar with the topic of the query.
TVB3	I found that it was easy to find clips that are relevant.
TVB4	For this topic I had enough time to find enough clips.
TVB5	For this particular topic the tool interface allowed me to do browsing efficiently.
TVB6	For this particular topic I was satisfied with the results of the browsing.

Table 1: Questions of the post-task questionnaire. Possible answers for each of the questions are: not at all, a little, fairly, quite a bit, very much.

minutes. Users can choose to do one or two sessions. In the latter case they work on a different type of task and a different data set in each of the sessions and complete only one pre-test survey in the first and one post-test survey in the second session.

2.4 Subjects

The tests have been performed with two groups of users. One group of users consists of staff of our institute (7 persons not involved with video browsing) and the second group consists of members of the different partners of the SEMEDIA project (12 persons: 6 persons from technical partners, 6 persons working in media production and archiving). 15 of them participated in one session consisting of 4 tasks, the other 4 in 2 sessions (i.e. completed 8 tasks).

In the pre-test part of the survey we have collected information about the subjects. About 75% of the users are researchers or faculty staff and two thirds of the users search the Web or information systems more than once per day. More than one third of the users never use video retrieval systems, while about half of the users use them between once week and once a day, the others more frequently. More than half of the users were unfamiliar with the tool to be evaluated, only 10% were fairly or more familiar with it. Two thirds had no or little knowledge of the data sets used, only 17% were fairly or more familiar with all of the data.

3. RESULTS

3.1 Correlation between methods

In order to compare the different evaluation approaches we analyze the correlation between the results. The assumption is that the F1 scores of a retrieval task and the corresponding question answering task are correlated, as well as the F1 measures with the answers to the questions TVB3-6 in the post-task questionnaire of the respective task (cf. Table 1).

The correlation coefficients between the F1 measures of the retrieval and question answering tasks are $r = -0.45$ and $\rho = -0.33$ (p-value 0.41). There is a slightly negative correlation between the results but no significant one. A t-test also shows at a significance level of 0.0001 that the two distributions have different means. We can conclude that retrieval and question answering tasks are not directly comparable, even if the users need a very similar result set to answer each of them.

Table 2 shows the correlation between the task results (again F1 measures) and the answers to the post-task questions of the respective task. The retrieval results are only correlated with question TVB6 at a significance level of 0.10, i.e. the user's satisfaction with the browsing result is positively correlated with the actual retrieval performance.

There is also only one strong correlation for the question answering task. The F1 measure is correlated with question TVB4 at a significance level of 0.10. But this is a negative correlation, i.e. the users scored worse on the question answering tasks for which they felt to have more time. A possible explanation is that users feel stressed in cases where they encounter many video segments that match the query but think they have more time than when they only encounter few relevant ones.

		TVB1	TVB3	TVB4	TVB5	TVB6
R (F1)	r	-0.06	0.27	0.33	0.42	0.62
	ρ	-0.05	0.21	0.36	0.40	0.71
	p	0.91	0.61	0.38	0.33	0.05
Q (F1)	r	0.23	-0.46	-0.80	-0.40	-0.46
	ρ	0.33	-0.41	-0.68	-0.16	-0.15
	p	0.42	0.31	0.06	0.70	0.73

Table 2: Correlation between F1 measures of retrieval (R) and question answering (Q) task results and the post-task questionnaire (r denotes Pearson's product-moment correlation coefficient, ρ denotes Spearman's rank correlation coefficient and p the associated p-value).

3.2 Independence of post-task questions

After each retrieval or question answering task the users answer the questions listed in Table 1. The correlation among the questions is shown in Table 3. There is a strong correlation among the questions TVB3, TVB4, TVB5 and TVB6, that can be accepted at a significance level of 0.10, in two cases even at a level of 0.01. These results show that it is difficult for the user to judge certain aspects separately (e.g. whether the tool was helpful in this case). Instead a general impression of the browsing experience is rated, including the satisfaction with the tool and with the results and the impression to have sufficient time.

The familiarity with the topic is not or only very weakly correlated with the other aspects. There is only a correlation with the perceived easiness of the task ($\rho = 0.67$ at significance level 0.10), i.e. the task seems easier for users who are familiar with the topic. However, they do not feel that they have more time or achieve more satisfying results than others.

3.3 Learning effect

We analyze this by looking for trends in the results achieved for the 4 tasks done in one session. As we have four different sequences of tasks, two different data sets and the tasks are done in different order by different users, the difficulty of individual tasks does not influence the trend. As expected, some of the measures (such as the familiarity with the search topic) do not show a clear pattern, but some seem to have a trend. If we fit a linear trend function to the data we get a clear trend for two of the questions: TVB4 (sufficient time,

		TVB3	TVB4	TVB5	TVB6
TVB1	r	0.33	0.27	0.43	0.36
	ρ	0.67	0.50	0.54	0.33
	p	0.07	0.20	0.17	0.43
TVB3	r		0.90	0.86	0.86
	ρ		0.86	0.86	0.64
	p		0.01	0.01	0.08
TVB4	r			0.67	0.81
	ρ			0.64	0.74
	p			0.09	0.03
TVB5	r				0.84
	ρ				0.67
	p				0.07

Table 3: Correlation among questions of the post-task questionnaire (r denotes Pearson's product-moment correlation coefficient, ρ denotes Spearman's rank correlation coefficient and p the associated p-value). The questions TVB1 through TVB6 are listed in Table 1.

slope 0.22) and TVB6 (satisfaction with results, slope 0.19). The longer users work with the tool the higher is their satisfaction with the results and they perceive the working time as more sufficient.

The question is whether this trend can also be measured in the task results. When fitting a trend function to the results of the question answering tasks, we get slopes of -0.06 for precision and -0.04 for recall, i.e. no clear trend can be seen, especially not a positive trend as in the survey answers. For the retrieval tasks the trend function for precision has a slope of 0.12 and that for recall of 0.01. The user's perception is supported by the precision values of the retrieval task, although the increase is not as strong as in the survey answers. The fact that the satisfaction is more related to precision than to recall can be explained as follows. The users know only about the video segments they have found, i.e. not about the correct ones not found that would be measured by recall. Thus the perceived quality of the results depends on how well the segments in the result set match the query, which correlates with precision.

4. CONCLUSION

In this paper we have applied two TRECVID style fact-finding approaches and a user survey to the evaluation of a video browsing tool. We analyze the correlation between the results of the different methods, whether different aspects can be evaluated independently with the survey, and if a learning effect can be measured with the different methods.

In general the results show (not unexpectedly) that especially the recall scores are rather low in such an application. This is definitely an issue that needs to be addressed in future work in video browsing. We are especially interested in comparing the different evaluation approaches.

We can conclude that the retrieval task correlates better with the user experience according to the survey than the question answering tasks. As retrieving relevant content is also closer to the real-world application of the tool than finding facts about the content, it seems to be the more appropriate evaluation method in this case, although it is a costly method due to the efforts for data set and ground truth preparation.

It turns out that the survey rather measures the general user experience while different aspects of the usability cannot be analyzed independently. This means that surveys are rather suitable for comparing the general usability of tools for certain applications than for getting information about strengths and weaknesses of a certain tool. However, it would be interesting to apply surveys to a comparative evaluation of two tools that differ in one aspect and see if the influence of this aspect on the results can be measured.

5. ACKNOWLEDGMENTS

The research described here has been partially supported by the European Commission under the contracts FP6-045032, "SEMEDIA" (`http://www.semedia.org`) and FP7-215475, "2020 3D Media" (`http://www.20203dmedia.eu/`).

6. REFERENCES

[1] S. F. Adafre and M. de Rijke. Exploratory search in wikipedia. In *SIGIR Workshop on Evaluating Exploratory Search Systems*, 2006.

[2] W. Bailer, C. Schober, and G. Thallinger. Video content browsing based on iterative feature clustering for rushes exploitation. In *Proc. TRECVID Workshop*, pages 230–239, Gaithersburg, MD, USA, Nov. 2006.

[3] W. Bailer and G. Thallinger. A framework for multimedia content abstraction and its application to rushes exploration. In *Proc. ACM CIVR*, Amsterdam, NL, Jul. 2007.

[4] B. J. Jansen, R. Ramadoss, M. Zhang, and N. Zang. Wrapper: An application for evaluating exploratory searching outside of the lab. In *SIGIR Workshop on Evaluating Exploratory Search Systems*, 2006.

[5] V. Jijkoun and M. de Rijke. A pilot for evaluating exploratory question answering. In *SIGIR Workshop on Evaluating Exploratory Search Systems*, 2006.

[6] W. Kraaij and W. Post. Task based evaluation of exploratory search systems. In *SIGIR Workshop on Evaluating Exploratory Search Systems*, 2006.

[7] Y. Qu and G. W. Furnas. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Inf. Process. Manage.*, 44(2):534–555, 2008.

[8] A. Smeaton and P. Wilkins. TRECVID 2004: Interactive search questionnaires. http://www-nlpir.nist.gov/projects/tv2004/questionnaires.html.

[9] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. ACM MIR*, pages 321–330, Santa Barbara, California, USA, 2006.

[10] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proc. ACM SIGIR*, 2001.

[11] R. W. White, B. Kules, S. M. Drucker, and m. schraefel. Supporting exploratory search. *Commun. ACM*, 49(4):36–39, 2006.

[12] R. W. White, G. Marchionini, and G. Muresan. Editorial: Evaluating exploratory search systems. *Inf. Process. Manage.*, 44(2):433–436, 2008.

[13] R. W. White, G. Muresan, and G. Marchionini. Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems. *SIGIR Forum*, 40(2):52–60, 2006.