Exploratory Search in an Audio-Visual Archive: Evaluating a Professional Search Tool for Non-Professional Users

Marc Bron
ISLA, University of Amsterdam
m.m.bron@uva.nl

Jasmijn van Gorp
TViT, Utrecht University
j.vangorp@uu.nl

Frank Nack
ISLA, University of Amsterdam
nack@uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

ABSTRACT

As archives are opening up and publishing their content online, the general public can now directly access archive collections. To support access, archives typically provide the public with their internal search tools that were originally intended for professional archivists. We conduct a small-scale user study where non-professionals perform exploratory search tasks with a search tool originally developed for media professionals and archivists in an audio visual archive. We evaluate the tool using objective and subjective measures and find that non-professionals find the search interface difficult to use in terms of both. Analysis of search behavior shows that non-professionals often visiting the description page of individual items in a result list are more successful on search tasks than those who visit fewer pages. A more direct presentation of entities present in the metadata fields of items in a result list can be beneficial for non-professional users on exploratory search tasks.

Categories and Subject Descriptors

H.5.2 [**User interfaces**]: Evaluation/methodology

General Terms

Measurment, Performance, Design, Experimentation

Keywords

Exploratory search, Usability evaluation

1. INTRODUCTION

Traditionally, archives have been the domain of archivists and librarians, who retrieve relevant items for a user's request through their knowledge of the content in, and organization of, the archive. Increasingly, archives are opening up and publishing their content online, making their collections directly accessible for the general public. There are two major problems that these non-professional users face. First, most users are unfamiliar or only partially familiar with the archive content and its representation in the repository. The internal representation is designed from the expert point of view, i.e., the type of information included in the metadata, which does not necessarily match the expectation of the general public. This leads to an increase in exploratory types of search [5], as users are unable to translate their information need into terms that correspond with the representation of the content in the archive. The second problem is that archives provide users with professional search tools to search through their collections. Such tools were originally developed to support professional users in searching through the metadata descriptions in a collection. Given their knowledge of the collection, professionals primarily exhibit directed search behavior [3], but it is unclear to what extent professional search tools support non-professional users in exploratory search.

The focus of most work on improving exploratory search is towards professionals [1]. In this paper we present a small-scale user study where non-professional users perform exploratory search tasks in an audio-visual archive using a search tool originally developed for media professionals and archivists. We investigate the following hypotheses: (i) a search interface designed for professional users does not provide satisfactory support for non-professional users on exploratory search tasks; and (ii) users with high performance on exploratory search tasks have different search behavior than users with lower performance.

In order to investigate the first hypothesis we evaluate the search tool performance objectively in terms of the number of correct answers found for the search tasks and subjectively through a usability questionnaire. To answer the second hypothesis, we perform an analysis of the click data logged during search.

2. EXPERIMENTAL DESIGN

The environment. The setting for our experiment was the Netherlands Institute for Sound and Vision (S&V), the Dutch national audiovisual broadcast archive. In the experiment we used the archive's collection consisting of around 1.5 M (television) programs with metadata descriptions provided by professional annotators.

We also utilized the search interface of S&V.[1] The interface is available in a simple and an advanced version. The simple version is similar to search engines known from the web. It has a single search box and submitting a query results in a ranked list of 10 programs. Clicking on one of the programs, the interface shows a page with the complete metadata description of the program. Table 1 shows the metadata fields available for a program. Instead of

[1] http://zoeken.beeldengeluid.nl

the usual snippets presented with each item in a result list, the interface shows the title, date, owner and keywords for each item on the result page. Only the keywords and title field provide information about the actual content of the program while the other fields provide information primarily used for the organization of programs in the archive collection. The description and summary fields contain the most information about the content of programs but are only available by visiting the program description page.

We used the advanced version of the interface in the experiment which next to the search box offers two other components: search boxes operating on specific fields and filters for certain categories of terms. Fielded searches operate on specific fields in the program metadata. The filters become available after a list of programs has been returned in response to a query. The filters display the top five most frequent terms in the returned documents for a metadata field. The metadata fields displayed in the filter component of the interface are highlighted in bold in Table 1. Once a checkbox next to one of the terms has been ticked, programs not containing that term in that field are removed from the result list.

Table 1: All metadata fields available for programs. We differentiate between fields that describe program content and fields that do not. Bold indicates fields used by the filter component.

content descriptors		organizational descriptors	
field	explanation	field	explanation
description	program highlights	**medium**	storage medium
person	people in program	**genre**	gameshow; news
keyword	terms provided by annotator	**rights**	parties allowed to broadcast
summary	summary of the program format	owner	owner of the broadcast rights
organization	organization in program	**date**	broadcast date
location	locations in program	origin	program origin
title	program title		

Subjects. In total, 22 first year university students from media studies participated in the experiment. The students (16 female, 6 male) were between 19 and 22 years of age. As a reward for participation the students gained free entrance to the museum of the archive.

Experiment setup. In each of the five studios available at S&V either one or two subjects performed the experiment at a time in a single studio. In case two subjects were present, each of them worked on machines facing opposite sides of the studio. We instructed subjects not to communicate during the experiment. During the experiment one instructor was always present in a studio. Before starting, the subjects learned the goals of the experiment, got a short tutorial on the search interface and performed a test query. During this phase the subjects were allowed to ask questions.

In the experiment each subject had to complete three search tasks in 45 minutes. If after 15 minutes a task was not finished, the instructor asked the subject to move on to the next task. Search tasks are related to matters that could potentially occur within courses of the student's curriculum. Each search task required the subjects to find five answers before moving on to the next task. A correct answer was a page with the complete metadata description of a program that fulfilled the information need expressed by the search task. Subjects could indicate that a page was an answer through a submit button added to the interface for the experiment.

We used the following three search tasks in the experiment: (i) For the course "media and ethnicity" you need to investigate the role of ethnicity in television-comedy. Find five programs with different comedians with a non-western background. (ii) For the course "television geography" you need to investigate the representation of places in drama series. Find five drama series where location plays an important role. (iii) For the course "media and gender" you need to give a presentation about the television career of five different female hosts of game shows broadcasted during the 1950s, 1960s or 1970s. Find five programs that you can use in your presentation.

Subjects received the search tasks in random order to avoid any bias. Also, subjects were encouraged to perform the search in any means that suited them best. During the experiment we logged all search actions, e.g., clicks, performed by each subject. After a subject had finished all three search tasks, he or she was asked to fill out a questionnaire about the experiences with the search interface.

Methodology for evaluation and analysis. We performed two types of evaluation of the search interface: a usability questionnaire and the number of correct answers submitted for the search tasks. The questionnaire consists of three sets of questions. The first set involves aspects of the experienced search behaviour with the interface. The second set contains questions about how useful users find the filter component, fielded search component, and metadata fields presented in the interface. The third set asks subjects to indicate the usefulness of a series of term clouds. The primary goal is not to evaluate the term clouds or their visualization but to find preferences for information from certain metadata fields. We generated a term cloud for a specific field as follows. First, we got the top 1000 program descriptions for the query "comedian." We counted the terms for a field for each of the documents. The cloud then represented a graphical display of the top 50 most frequent terms in the fields of those documents, where the size of a term was relative to its frequency, i.e, the higher the frequency the bigger the term. In the questionnaire subjects indicate agreement on a 5 point Likert scale ranging from one (not at all) to five (extremely). The second type of evaluation was based on the evaluation methodology applied at TREC [2]. We pooled the results of all subjects and let two assessors make judgements about the relevance of the submitted answers to a search task. An answer is only considered relevant if both assessors agree. Performance is measured in terms of the number of correct answers (#correct) submitted to the system.

For the analysis of the search behavior of subjects we looked at (i) the number of times a search query is submitted using any combination of components (#queries); (ii) the number of times a program description page is visited (#pages); and (iii) the number of times a specific component is used, i.e., the general searchbox, filters and fields. A large value for #queries indicates a look up type search behavior. It is characterized by a pattern of submitting a query, checking if the answer can be found in the result list and if it is not, to formulate a new query. The new query is not necessarily based on information gained from the retrieved results but rather inspired by the subject's personal knowledge [4]. A large value for #pages indicates a learning style search behavior. In this search strategy a subject visits the program description of each search result to get a better understanding of the organization and content of the archive. New queries are then also based on information gained from the previous text analysis [4]. We check the usage frequency of specific components to see if performance differences between subjects are due to alternative uses of interface components.

3. RESULTS

Search interface evaluation. Figure 1 shows the distribution of the amount of *correct* answers submitted for a search task, together with the distribution of the amount of answers (correct or incorrect) submitted. Out of the possible total of 330 answers, 173 are actually submitted. Subjects submit the maximum number of five

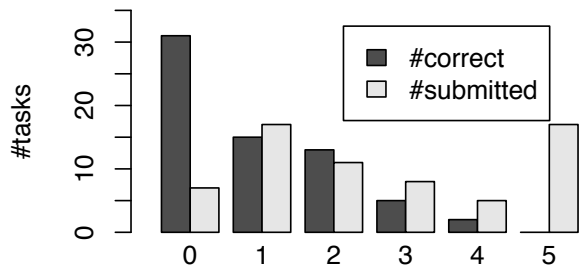

Figure 1: Distribution of amount correct/submitted answers.

answers for 18 of the tasks. This suggests that subjects have difficulties in finding answers within the given time limit. Subjects find no *correct* answers for 31 of the tasks, five subjects find no *correct* answer for any of the tasks, and none of the subjects reaches the maximum of five *correct* answers for a task. In total 64 out of 173 answers are *correct*. This low precision indicates that subjects find it difficult to judge if an answer is correct based on the metadata provided by the program description. Table 2 shows questions about the satisfaction of subjects with the interfaces. Subjects indicate their level of agreement from one (not at all) to five (extremely). For all questions the majority of subjects find the amount of support offered by the interface on the exploratory search tasks marginal. This finding supports our first hypothesis that the search interface intended for professional users does not provide satisfactory support to non-professional users on exploratory search tasks.

Search behavior analysis. Although all subjects are non-experts with respect to search with this particular interface, some perform better than others. We investigate whether there is a difference in the search behavior of subjects that have high performance on the search tasks and users that have lower performance. We divide subjects into two groups depending on the average number of correct answers found aggregated over the three tasks, i.e., 2.9 out of the possible maximum of 15. The group with higher performance (group G) consists of 11 subjects with 3 or more correct answers, whereas the group with lower performance (group B) consists of 11 subjects with 2 or less correct answers.

Table 3 shows the averages of the search behavior indicators for each of the two groups. We first look at the usage frequency of the filter, field, and search box components by subjects in group G vs. group B. There is no significant difference between the groups, indicating that there is no direct correlation between performance on the search tasks and use of specific search components. Next we look at search behavior as an explanation for the difference in performance between the groups. Our indicator for lookup searches, i.e., #queries, shows a small difference in the number of submitted queries. That subjects in both groups submit a comparable number of queries suggests that the difference in performance is not due to one group doing more lookups than the other. The indicator for learning type search, i.e., #pages, shows that there is a significant difference in the number of program description pages visited between subjects of the two groups, i.e., subjects in group G tend to visit program description pages more often than subjects of group B. We also find that the average time subjects in group G spend on a program description page is 27 seconds, while subjects from group B spend on average 39 seconds. These observations support our hypothesis that there are differences in search behavior between subjects that have high performance on exploratory search tasks and subjects with lower performance.

Usefulness of program descriptions. One explanation for this difference in performance is that through their search behavior subjects from group G learn more about the content and organization of the archive and are able to assimilate this information faster from the program descriptions than subjects from group B. As subjects process more program descriptions they learn more about the available programs and terminology in the domain. This results in a richer set of potential search terms to formulate their information need. To investigate whether subjects found information in the program descriptions useful in suggesting new search terms, we analyse the second set of questions from the questionnaire. The top half of Table 4 shows subjects' responses to questions about the usefulness of metadata fields present on the search result page. Considering responses from all subjects the genre and keyword fields are found most useful and the title and date fields as well, although to a lesser degree. The fields intended for professionals, i.e., origin, owner, rights, and medium are found not useful by the majority of subjects. Between group B and G there are no significant differences in subject's judgement of the usefulness of the fields.

Table 2: Questionnaire results about the satisfaction of subjects with the search interface. Agreement is indicated on a 5 point Likert scale ranging from one (not at all) to five (extremely).

question	mode	avg
To what degree are you satisfied with the search experience offered by the interface?	2	2.3
To what degree did the interface support you by suggesting new search terms?	2	2.4
To what degree are you satisfied with the suggestions for new search terms by the interface?	2	2.3

Table 3: Analysis of search behavior of subjects. Significance is tested using a standard two-tailed t-test. The symbol $^{\blacktriangle}$ indicates a significant increase at the $\alpha < 0.01$ significance level.

	filter	field	searchbox	#queries	#pages
B avg	21.3	29.5	44.8	35.2	21.2
G avg	15.2	44.0	42.0	34.3	$35.7^{\blacktriangle}$

Table 4: Questions about the usefulness of metadata fields on program description pages and the mode and average (avg) of the subjects responses: for all subjects, the good (G) and bad (B) performing group. We use a Wilcoxon signed rank test for the ordinal scale. The symbol $^{\triangle}$ ($^{\blacktriangle}$) indicates a significant increase at the $\alpha < 0.05$ (0.01) level.

		all	B		G	
question	field	mode	mode	avg	mode	avg
Degree to which fields on the result page were useful in suggesting new terms	date	3	2	2.2	3	3.0
	owner	1	1	1.6	1	2.0
	rights	1	1	1.3	1	1.4
	genre	4	1	2.8	4	3.9
	keyword	4	1,5	3.1	4	3.5
	origin	1	1,2	1.7	1	2.0
	title	3,4	2	2.2	4	3.0
	medium	1	1	1.5	1,2	1.6
Degree to which fields in program descriptions were useful in suggesting new terms	summary	4	1,4	2.8	5	3.8
	description	4	4	3.3	$4^{\triangle}$	4.1
	person	4	1,3,4	2.8	$4^{\blacktriangle}$	3.8
	location	1,3,4	1,3	2.0	$4^{\blacktriangle}$	3.0
	organization	1	1	1.8	1,2	2.0

The bottom part of Table 4 shows subject's responses to questions about the usefulness of metadata fields only present on the program description page and not already shown on the search result page. Based on all responses, the summary, description, person and location metadata fields are considered most useful by the majority of the subjects. These findings further support our argument that program descriptions provide useful information for subjects to complete their search tasks.

When we contrast responses of the two groups we find that group G subjects consider the description, person, and location metadata fields significantly more useful than subjects from group B. This suggests that group B subjects have more difficulties in distilling useful information from these fields (recall also the longer time spent on a page). This does not say that these users cannot understand the provided information. All that is indicated is that the chosen modality, i.e., text, might not be the right one. A graphical representation, for example as term clouds, might be better.

Fields as term clouds. In response to the observations just made, we also investigated how users would judge visual representations of search results, i.e., in the form of term clouds directly on the search result page. Here the goal is not to evaluate the visualization of the clouds or the method by which they are created. Of interest to us is whether subjects would find a direct presentation of information normally "hidden" on the program description page useful.

Recall from §2 that we generate term clouds for each field on the basis of the terms in the top 1000 documents returned for a query. From Table 5 we observe that subjects do not consider the description and summary clouds useful, while previously these fields were judged most useful among the fields in the program description. Both clouds contain general terms from the television domain, e.g., program and series, which do not provide subjects with useful search terms. Although this could be due to the use of frequencies to select terms, these fields are inherently difficult to visualize without losing the relations between the terms. The genre, keyword, location and, to some degree, person clouds are all considered useful, but they support the user in different ways. The genre field supports the subject in understanding how content in the archive is organized, i.e., it provides an overview of the genres used for categorization. The keyword cloud provides the user with alternative search terms for his original query, for example, satire or parody instead of cabaret. The location and person clouds offer an indication of which locations and persons are present in the archive and how prominent they are. For these fields visualization is easier, i.e., genre, keywords or entities by themselves are meaningful without having to represent relations between them. Subjects consider the title field only marginally useful. For this field the usefulness is dependent on the knowledge of the subject as titles are not necessarily descriptive. The subjects also consider the organization field marginally useful, probably due to the nature of our search tasks, i.e., two tasks focus on finding persons and in one locations play an important role. We assume though that in general this type of information need occurs when the general public starts exploring the archive. Together, the above findings suggest that subjects find a direct presentation of short and meaningful terms, i.e., categories, keywords, and entities, on the search results page useful.

Table 5: Questions about the usefulness of term clouds based on specific metadata fields. Agreement is indicated on a 5 point Likert scale ranging from one (not at all) to five (extremely).

cloud	mode	avg	cloud	mode	avg
title	2	2.8	description	1	2.5
person	2,3	2.9	genre	4	3.4
location	4	3.3	summary	1	2.3
organization	2	2.2	keyword	4	3.8

4. CONCLUSION

We presented results from a user study where non-professional users perform exploratory search tasks with a search tool originally developed for media professionals and archivists in an audio visual archive. We hypothesized that such search tools provide unsatisfactory support to non-professional users on exploratory search tasks. By means of a TREC style evaluation we find that subjects achieve low recall in the number of correct answers found. In a questionnaire regarding the user satisfaction with the search support offered by the tool, subjects indicate this to be marginal. Both findings support our hypothesis that a professional search tool is unsuitable for non-professional users performing exploratory search tasks.

Through an analysis of the data logged during the experiment, we find evidence to support our second hypothesis that subjects perform different search strategies. Subjects that visit more program description pages are more successful on the exploratory search tasks. We also find that subjects consider certain metadata fields on the program description pages more useful than others. Subjects indicate that visualization of certain fields as term clouds directly in the search interface would be useful in completing the search tasks. Subjects especially consider presentations of short and meaningful text units, e.g., categories, keywords, and entities, useful.

In future work we plan to perform an experiment in which we present non-professional users with two interfaces: the current search interface and one with a direct visualization of categories, keywords and entities on the search result page.

Acknowledgements. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

REFERENCES

[1] J.-w. Ahn, P. Brusilovsky, J. Grady, D. He, and R. Florian. Semantic annotation based exploratory search for information analysts. *Inf. Proc. & Management*, 46(4):383 – 402, 2010.

[2] D. K. Harman. The TREC test collections. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and evaluation in information retrieval*. MIT, 2005.

[3] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. Search behavior of media professionals at an audiovisual archive. *J. Am. Soc. Inf. Sci. and Techn.*, 61:1180–1197, 2010.

[4] G. Marchionini. Exploratory search: from finding to understanding. *Comm. ACM*, 49(4):41 – 46, April 2006.

[5] R. White, B. Kules, S. Drucker, and M. Schraefel. Supporting exploratory search: Special issue. *Comm. ACM*, 49(4), 2006.