The Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis (Abstract)*

Bouke Huurnink
ISLA, University of Amsterdam
bhuurnink@uva.nl

Laura Hollink
Delft University of Technology
l.hollink@tudelft.nl

Wietske van den Heuvel
Netherlands Institute for Sound and Vision
wvdheuvel@beeldengeluid.nl

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

ABSTRACT

Finding audiovisual material for reuse in new programs is an important activity for news producers, documentary makers, and other media professionals. Such professionals are typically served by an audiovisual broadcast archive. We report on a study of the transaction logs of one such archive. The analysis includes an investigation of commercial orders made by the media professions, as well as a characterization of sessions, queries, and the content of terms recorded in the logs. We identify a strong demand for short pieces of audiovisual material in the archive. Also, searchers are generally able to quickly navigate to a usable audiovisual broadcast, but it takes them longer to place an order for a subsection of a broadcast than it does for them to order an entire broadcast. Queries are found to predominantly consist of (parts of) broadcast titles and of proper names. Our observations imply that it may be beneficial to increase support for fine-grained access to audiovisual material, for example, through manual segmentation or content-based analysis.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation, Measurement

Keywords

Transaction log analysis, audiovisual archive

1. INTRODUCTION

Documentary makers, journalists, news editors, and other media professionals routinely require previously recorded audiovisual material for reuse in new productions. For example, a news editor might wish to reuse footage shot by overseas services for the evening news. To complete production, the media professional must locate audiovisual material that has been previously broadcast in another context. One of the sources for reusable broadcasts is the audiovisual archive, which specializes in the preservation and management of audiovisual material [1]. Where audiovisual material was once primarily stored on analog carriers, in recent years audiovisual archives have started making their content available in digital format and enabling online access [2, 3]. In such an environment, the media professional can search for and purchase multimedia material. In addition, with audiovisual acquisition being done through a digital interface, the archive can record information about the media professional's information seeking process. Despite the fact that an increasing amount of audiovisual programming is produced digitally, little is known about the search behavior of media professionals locating material for production purposes.

We aim to characterize the behavior of users of an audiovisual archive, and in addition to give insight into the content of their searches. We work in the context of a large national audiovisual broadcast archive, actively used by media professionals from a range of production studios. The archive presents a rich source of information because of its specialist nature. Our study is performed through an analysis of *transaction logs* — the electronic traces left behind by users interacting with the archive's online retrieval and ordering system. The transaction log analysis is enhanced by leveraging additional resources from the audiovisual broadcasting field, which can be exploited due to the specialist nature of the archive. In particular we analyze purchase orders of audiovisual material, and we use catalog metadata and the a structured audiovisual thesaurus to investigate the content of query terms.

2. METHOD

Our study takes place within the context of the *Nederlands Instituut voor Beeld en Geluid* — the Netherlands Institute for Sound and Vision, a large audiovisual archive, which we will refer to below as "the archive." The archive functions as the main provider of archive material for broadcasting companies in the Netherlands. The collection contains more than 700,000 hours of radio, television, documentaries, films and music. Nowadays, all digitally broadcast television and radio programs made by the Dutch public broadcasting companies are automatically ingested in the archive's digital asset management system. The digital multimedia items available in the archive can be divided into two types: video and audio. The video items consist largely of television broadcasts, but also include movies, amateur footage, and internet broadcasts. The audio portion of the collection consists primarily of radio broadcasts and music recordings. Each catalog entry contains multiple fields which contain either freely entered text or structured terms.

*The full version of this paper appeared in *Journal of the American Society for Information Science and Technology*, 61(5):994-1014, May 2010.

DIR'11, February 4, 2011, Amsterdam.
ACM xxx-x-xxxxx-xxx-x/xx/xx.

The archive is primarily used by media professionals who work for a variety of broadcasting companies, public and commercial, and are involved in the production of a range of programs. Once purchased, ordered audiovisual material may be re-used in many types of programs, especially news and current affairs programs. In addition, it is sometimes used for other purposes, for example to populate online platforms and exhibitions.

Search through audiovisual material in the archive is based on manually created catalog entries. These entries are created by the archival staff, and include both free text as well as structured terms contained in a specialized audiovisual thesaurus. The thesaurus is called the *GTAA* — the Dutch acronym for "Common Thesaurus for Audiovisual Archives."

3. EXPERIMENTAL DESIGN

Transaction logs from *Beeld en Geluid*'s online search and purchasing system were collected between November 18, 2008 and May 15, 2009. The logs were recorded using an in-house system tailored to the archive's online interface.

In total the logs contained $290,429$ queries after cleaning. The transaction logs often reference documents contained in the archive's collection. In order to further leverage the log information, we obtained a dump of the catalog descriptions maintained by the archive on June 19, 2009. The size of the catalog at that time was approximately $700,000$ unique indexed program entries.

We define five key units that will play a role in our analysis below: the three units common in transaction log analysis of session, query, term; and two units specific to this study, facet and order. The specialized knowledge sources available within the archive allowed (and motivated) us to develop the last two units of analysis.

4. MAIN FINDINGS

Our analysis was structured around four main research questions, the answers to which we summarize here. With respect to our first question, *what characterizes a typical session at the archive?*, we found the typical session to be short, with over half of the sessions under a minute in duration. In general, there were also few queries and result views in a session, with a median value of one query issued and one result viewed. Sessions resulting in orders had a considerably longer duration, with over half of the sessions having a median duration of over seven minutes, but no increase in terms of the number of queries issued and results viewed.

In answer to our second question, *what kinds of queries are users issuing to the audiovisual archive?*, we found nearly all of the queries contained a keyword search in the form of free text, while almost a quarter specified a date filter. The advanced search option, for searching on specific catalog fields, was used in 9% of the queries. The most frequently occurring keyword searches consisted primarily of program titles. Advanced search on specific catalog fields, when utilized, frequently specified the media format or copyright owner of the results to be returned, for example that only results available in high-quality digital format should be returned.

In addressing our next research question, *what kinds of terms are contained in the queries issued to the archive?*, we performed a content analysis of the query terms. This was accomplished by using catalog information as well as session data; terms in a query were matched to the titles and thesaurus entries of the documents that were clicked during a user session. This allowed us to leverage the thesaurus structure for identifying different kinds of query terms. The approach does have limitations, as terms can only be identified in sessions where users click at least one result, and even then, a term can only be identified if it is present as a title or thesaurus entry. Of all the queries where users clicked a result during the session, 41% contained a title term. Thesaurus terms were identified in 44% of the queries. Approximately one quarter of the thesaurus terms consisted of general subjects such as *soccer*, *election*, and *child*. Another quarter consisted of the names of people, especially of politicians and royalty. The remaining terms were classified as locations, program makers, other proper names, or genres.

To answer our final research question, *what are the characteristics of the audiovisual material that is ordered by the professional users?*, we isolated the orders placed to the archive. Orders were for recent and historical material, with 46% of orders for items that were broadcast over one year before the order date. We identified three units of ordering: *programs*, *stories*, and *fragments*. We saw that less than a third of orders placed to the archive were for entire broadcasts, while 17% of the orders were for subsections of broadcasts that had been previously defined by archivists. Nearly half of the orders were for audiovisual fragments with a start and end time specified by the users. The fragments were typically on the order of a few minutes in duration, with 28% of fragments being one minute or less. In these cases, where users specified the fragment boundaries manually, sessions typically took more than two and half times as long as when ordering an entire broadcast.

5. CONCLUSION

Our main contributions in this paper include: a description of the search behavior of professionals in an audiovisual archive in terms of sessions and queries, and orders; a categorization of their query terms by linking query words to titles and thesaurus terms from clicked results; and an analysis of the orders made from the archive in terms of their size relative to the broadcast length and the time taken to get from query to purchase. Our study is significant in that there is a relatively large time span covered (almost half a year), and in that the users are specialists in audiovisual search, looking for broadcasts and fragments of broadcasts for reuse in new productions. In addition, we utilize catalog annotations to provide additional detail about the data recorded in the transaction logs. The results of the study can serve to give researchers and archives insight into aspects of multimedia search related to the specific use case of media professionals. They may also be used by audiovisual broadcast archives to better adjust their services to the user.

Acknowledgments. This research was 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 (GALATEAS), by the 7th Framework Program of the European Commission, grant agreement no. 258191 (PROMISE), by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.-066.512, 612.061.814, 612.061.815, 640.004.802, 380-70-011 and by the Center for Creation, Content and Technology (CCCT).

References

[1] R. Edmondson. *Audiovisual Archiving: Philosophy and Principles*. UNESCO, Paris, France, 2004.

[2] J. Oomen, H. Verwayen, N. Timmermans, and L. Heijmans. Images for the future: Unlocking value of audiovisual heritage. In *Museums and the Web 2009: Proceedings*, Toronto, Ontario, Canada, 2009. Archives & Museum Informatics.

[3] R. Wright. Annual report on preservation issues for European audiovisual collections. Deliverable PS_WP22_BBC_D22.4_Preservation Status_2007, BBC, 2007.