Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis

Bouke Huurnink
Intelligent Systems Lab Amsterdam, University of Amsterdam, Science Park 107, Amsterdam, the Netherlands

Laura Hollink
Computer Science Department, FEW, VU University Amsterdam, De Boelelaan 1081A, Amsterdam, the Netherlands

Wietske van den Heuvel
Research and Development Department, Netherlands Institute for Sound and Vision, Sumatralaan 45, Hilversum, the Netherlands

Maarten de Rijke
Intelligent Systems Lab Amsterdam, University of Amsterdam, Science Park 107, Amsterdam, the Netherlands

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 professionals and a characterization of sessions, queries, and the content of terms recorded in the logs. One of our key findings is that there is a strong demand for short pieces of audiovisual material in the archive. In addition, while searchers are generally able to quickly navigate to a usable audiovisual broadcast, it takes them longer to place an order when purchasing a subsection of a broadcast than when purchasing an entire broadcast. Another key finding is that queries 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.

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, or a documentary maker describing the history of the Christmas tradition might desire footage from Christmas broadcasts in the 1930s. 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 (Edmondson, 2004). Although 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 (e.g., Oomen, Verwayen, Timmermans, & Heijmans, 2009; Wright, 2007). In such a digital environment, the media professional can search for and purchase multimedia material without leaving the comfort of his or her own office. 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 being digitally produced, little is known about the search behavior of media professionals locating material for production purposes.

In this study, we aim to characterize the behavior of users of an audiovisual archive and to give insight into the content of their searches. We work in the context of a large, national

Received September 30, 2009; revised January 25, 2010; accepted January 26, 2010

 DOI: 10.1002/asi.21327

audiovisual broadcast archive that is actively used by media professionals from a range of production studios. The archive presents a rich source of information because of its specialist nature. It serves a highly motivated group of professional users primarily searching for audiovisual material to reuse in new productions. The 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 a structured audiovisual thesaurus to investigate the content of query terms.

Our main contributions in this article 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 6 months) and 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 also may be used by audiovisual broadcast archives to better adjust their services to the user and by multimedia benchmarking activities to formulate new testing scenarios.

The remainder of the article is organized as follows. In the next section, we describe the research questions that will direct our transaction log analysis. We then review related studies of both transaction logs and audiovisual archives. This is followed by an outline of the experimental method used in the article, including a description of the audiovisual broadcast archive in which our study takes place and definitions of the primary units that will be used for analysis. We then describe in detail the sessions, queries, terms, facets, and orders contained in the query logs. We follow with conclusions, a discussion of implications of the study for both the archive and for research, and areas for future work.

Research Questions

In our transaction log analysis, we aim to shed light on the manner in which media professionals search through an audiovisual archive, and to gain insight into the types of content for which they are searching. We formulate the following research questions to direct our analysis:

RQ1: What constitutes a typical search session at the archive? *For example, how long does a session typically take? How many queries does a user issue in a session? How do queries develop within a session? Do sessions resulting in an order differ from sessions where no order is placed?*

RQ2: How are users issuing queries to the audiovisual archive? *For example, which search options are being used? What are commonly occurring queries?*

RQ3: What type of content is requested in the terms contained in queries issued to the archive? *For example, are terms often in the form of program titles? What is the distribution of terms across thesaurus facets? What are commonly occurring terms?*

RQ4: What are the characteristics of the audiovisual material that is finally ordered by the professional users? *For example, do users order whole programs or fragments of programs? What is the typical duration of ordered material? Is there a preference for recent material?*

When answering these questions, we will highlight the similarities and differences between the studied audiovisual archive and other audiovisual settings through a comparison to other transaction-log studies. By delving into the characteristics of the orders made in the archive, we exploit an unusual form of information that is present in the professional archive; namely, explicit indications of the desired audiovisual content given in the form of purchase orders. While we are limited in terms of interpretations of the log data due to the lack of direct interaction with users, we are able to generalize over the large amount of data that is gathered. The results of the analysis can serve as a baseline survey for system designers, quantifying the status quo in the studied archive. Thus, the effect of alterations to the search system can be compared to the baseline.

Related Work

Transaction Log Analysis

One way in which user search behaviors may be analyzed is through a transaction log analysis, which over the years has proved an apt method for the characterization of user behavior. Its strengths include its nonintrusive nature—the logs are collected without questioning or otherwise interacting with the user—and the large amounts of data that can be used to generalize over the cumulative actions taken by large numbers of users (Jansen, 2008). Note that transaction log analysis faces limitations: Not all aspects of the search can be monitored by this method, such as the underlying information need (Rice & Borgman, 1983). It also can be difficult to compare across transaction log studies of different systems due to system dependencies and varying implementations of analytical methods. Comparability can be improved to some extent by providing clear descriptions of the system under investigation and the variables used (Jansen & Pooch, 2001).

Information science has a long history of transaction-log analysis, from early studies of the logs created by users of library online public access catalog systems (Peters, 1993) to later studies of the logs of Web search engines (Jansen & Pooch, 2001). This was followed by the analysis of more specialized search engines and their transaction logs. For

DOI: 10.1002/asi

instance, Mishne and de Rijke (2006) studied the behavior of users of a blog search engine through a log file analysis, and Carman, Baillie, Gwadera, and Crestani (2009) examined the difference between the vocabularies of queries, social bookmarking tags, and online documents. Three frequently used units of analysis have emerged from the body of work: the *session*, the *query*, and the *term*, though the definition of each unit may vary across studies (Jansen & Pooch, 2001).

Let us now turn to studies of logs from searches for audiovisual material in particular. Here, Web-based search engines have formed the basis for a number of transaction log studies in multimedia retrieval. Such engines are closely related to the audiovisual archive in that they, too, provide access to audio and video material. There are a number of important differences between Web-based multimedia search engines and audiovisual archives: Web-based search engines serve the general online public rather than a group of specialist users; they offer access to heterogeneous audio and video of varying quality rather than broadcast-quality, professionally produced material; and as search data they generally rely on text obtained from Web pages or user tags rather than from professional archive catalog descriptions.

In one of the earliest studies of transaction logs for online multimedia search, Ozmutlu, Spink, and Ozmutlu (2003) compared multimedia queries from 2001 to multimedia queries from 1997 to 1999 and found that multimedia queries were changing rapidly as Web content and searching techniques evolved. In particular, search sessions were becoming shorter with fewer query modifications while queries themselves contained increasing numbers of terms. Jansen, Spink, and Pedersen (2004) found that multimedia Web searching was relatively complex as compared to general Web searching, with a longer average query length and higher use of Boolean operators, in a study of transaction logs gathered in 2002. In a later study of transaction logs gathered in 2006, Tjondronegoro, Spink, and Jansen (2009) found that multimedia searches used relatively few search terms. In addition, they found multimedia searches to be generally short in duration, with more than 50% of searching sessions being less than 1 minute in duration. The authors used an open-source classification tool to categorize approximately 2% of queries and found that many multimedia searches are for information about people. In a comparison with logs from earlier years, they found that online multimedia search had begun to shift from entertainment to other categories such as medical, sports, and technology.

Christel (2007) analyzed transaction logs and questionnaires to study the behavior of professional users—government intelligence analysts—who use an experimental, content-based video-retrieval system to answer a set of predefined information needs. He found that up to 57% of retrieved relevant results were obtained by utilizing detectors that automatically identify visual objects and events occurring in video. In our setting, detectors for content-based retrieval are not available; however, users are able to search on concepts in the form of thesaurus terms, as well as free text, manually entered in the audiovisual catalog.

Finally, Pu (2008) analyzed "failed" image queries obtained from the logs of three image search engines. She found that failed queries (i.e., those that resulted in zero hits) were much longer, more distinct, and more unique than were successful queries; 1,000 failed queries were manually categorized into 28 "refined types" and were found to be more conceptual than perceptual (i.e., more about abstract ideas than about things one can see). In this study, we do not examine failed queries, but we do examine the difference between sessions where an order was made and sessions where no order was made. From the perspective of the archive, a session without an order could be viewed as a "failed" session, as such a session does not lead to a purchase.

An important result from the increasing number of detailed studies of users' search behavior is multiple typologies of queries and searches. For example, Smeulders, Worring, Santini, Gupta, and Jain (2000) provided a categorization of content-based image-retrieval queries: target (or known-item) search (i.e., when the user has a specific image in mind), category search (i.e., retrieving an arbitrary image representative of a specific class), and search by association (i.e., search starts with no aim other than to find interesting things). Broder (2002) described three query types in the context of Web search: informational ("I need to know about a topic."), navigational ("Take me to a specific item or site."), and transactional ("I need to purchase or download a product or service."). This typology has served as the basis for a number of query-classification schemes, including those by Rose and Levinson (2004), Kellar, Watters, and Shepherd (2007), and Jansen, Booth, and Spink (2008). Early studies of audiovisual archives have indicated that searches for known items and specific people and events tend to be quite frequent, as discussed next.

Studies of Audiovisual Archives

Audiovisual archives have, in general, only recently started to digitize their content on a large scale (Wright, 2007), and we are not aware of any previous transaction log studies done within the context of the audiovisual archive. However, a number of specialized audiovisual archives have performed analyses of their users without the aid of transaction logs, instead analyzing by hand requests made to the archives. These requests are generally natural language texts in the form of e-mail, and so on, where the underlying information need is more explicitly identified than when issuing queries to a search engine.

Sandom and Enser (2001) analyzed 1,270 information requests to 11 different (mostly commercial) film archives. They found that approximately 10% of the information requests were for known items, specifying video titles, films by particular directors, or starring particular actors. These requests were excluded from the analysis. The remaining information requests specified the desired content or subject of the footage. These were categorized according to the Panofsky-Shatford matrix (Panofsky, 1962), which classifies different types of visual content. They found that, overall,

DOI: 10.1002/asi

68% of requests were specific, 56% were generic, and 2% were abstract. This is further broken down according to client type. Their final conclusion is that in the majority of cases, the moving image information seeker is looking for audiovisual material that illustrates specific events and shows named individuals or groups of people, in particular places or on unique dates.

Hertzum (2003) manually examined 275 e-mail requests to the Deutsche Film Institut, a noncommercial archive of European films. The text of the e-mail requests, which contained on average 111 words, was used to categorize them into search types: known-item retrieval, fact retrieval, subject retrieval, exploratory search, and other. He found that most requests could be classified as known-item retrieval (43%) or subject retrieval (32%) requests. A further analysis of the e-mail requests showed that 75% of the requests specified production-related attributes such as the title, director, or production year of a film. Only 38% of requests specified content, subject, or context attributes. This is in contrast to the study by Sandom and Enser (2001), where 90% of the requests studied desired content or subject matter. Hertzum speculated that such discrepancies reflect differences in the requester's tasks.

Jörgensen and Jörgensen (2005) analyzed 685 searches obtained from the logs of a commercial image archive. The archive does not include any audio media and therefore is not strictly an audiovisual archive. Nevertheless, we include this study in our overview, as it is one of very few studies of commercial archives of nontextual media. The broad goal of their research is to "fill the knowledge gap concerning 'typical' search sessions [. . .] and to provide a general picture of searching by a particular group of image professionals" (Jörgensen and Jörgensen, 2005, p. 1348). They found that unique term searches (searches for proper nouns, which refer to a specific concept; i.e., a type of navigational query) were less frequent than other studies in image archives (Armitage & Enser, 1997; Enser, 1993) had found. They also found a large amount of query modifications (62% of queries).

Method

Now that we have described related work in transaction log analysis and studies of users of audiovisual archives, we turn to a description of the methodology used in the current analysis. To facilitate comparison to other studies, we sketch the organizational setting in which we work, including what is known of the users and the retrieval system with which they work, as well as details of the transaction log data and the units of analysis. The archive gave us access to internal reports and unpublished technical documents, on which we base the organizational background described in this section.

Organizational Setting

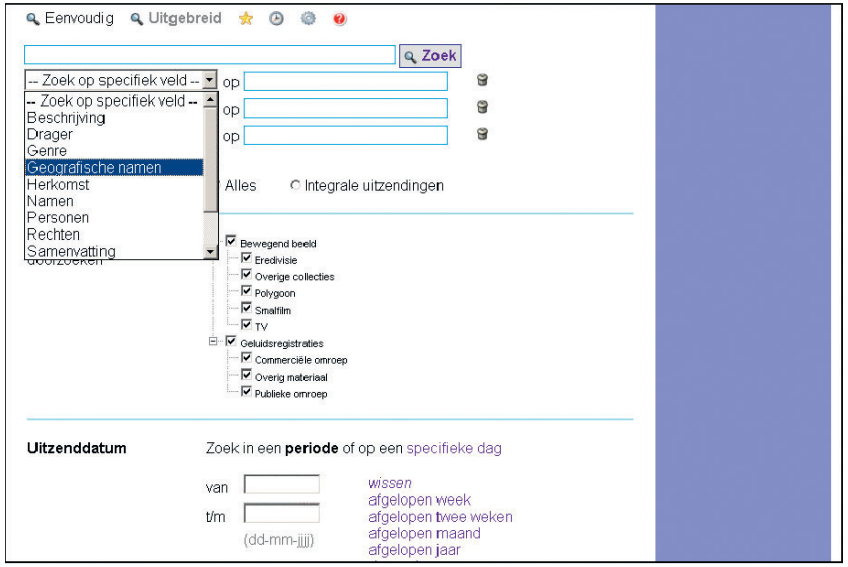
Our study takes place within the context of the *Nederlands Instituut voor Beeld en Geluid* (The Netherlands Institute for Sound and Vision), a large audiovisual archive, which we will refer to as "the archive." The archive functions as the main provider of archive material for broadcasting companies in the Netherlands. It manages over 70% of the Dutch audiovisual cultural heritage and is one of the largest audiovisual archives in Europe. The collection contains more than 700,000 hours of radio, television, movies, 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. A core task of the archive is to manage and preserve the Dutch audiovisual heritage; this includes contextualizing the material by manually adding metadata to catalog entries. Each catalog entry contains multiple fields that contain either freely entered text or structured terms.

Users of the archive. The archive is used primarily by media professionals, though there is some incidental use of the archive by members of the general public and others interested in the archive's content. The media professionals work for a variety of broadcasting companies, public and commercial, and are involved in the production of a range of programs. Not all media professionals interact directly with the search engine; they also may request material through the archive's internal customer service department, who then searches through the archive for them. Broadcast production companies typically employ one or more researchers whose sole task is to search for reusable audiovisual material. Production companies with a small team of editors usually let their editors perform the searches. In most cases, the decision whether to use a candidate piece of audiovisual material in the final cut of a program is made jointly by the researchers, editors, journalists, producers, and directors. It is not uncommon for researchers to transfer their search results to another staff member, who then does the final ordering for various programs. This transfer is made possible by the archive's online interface. Once purchased, ordered audiovisual material may be reused in many types of programs, especially news and current affairs programs, talk shows based on daily news, late night shows, and programs about history. In addition, it is sometimes used for other purposes, such as to populate online platforms and exhibitions.

Searching and ordering audiovisual material. To support users in sourcing audiovisual material, the archive has developed a retrieval system, iMMix. Some screen shots of the retrieval interface are shown in Figure 1. The iMMix interface contains a simple search page and an advanced search page; the latter is shown in Figure 1a. The search pages offer the following three search options: *keyword search* on freely entered text, returning results according to the textual content of all catalog fields; *date filtering*, which can be used to return

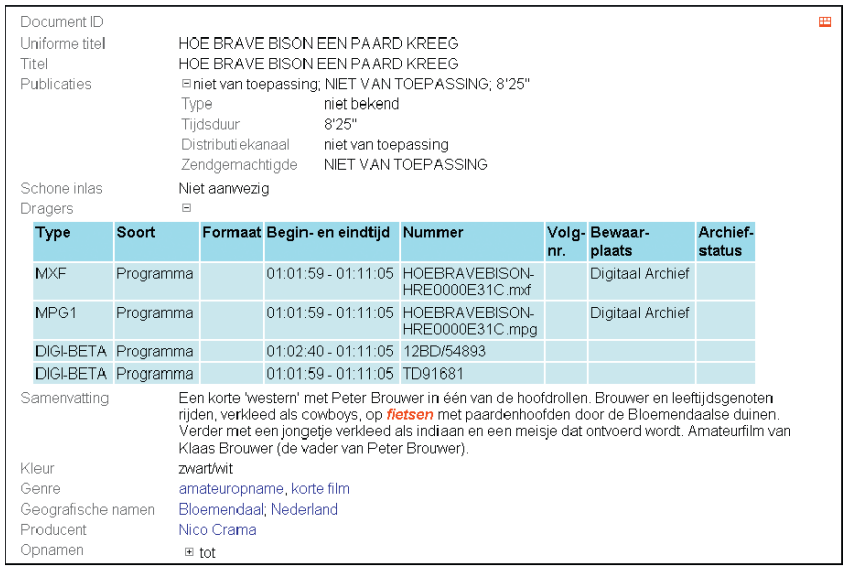
DOI: 10.1002/asi

DOI: 10.1002/asi

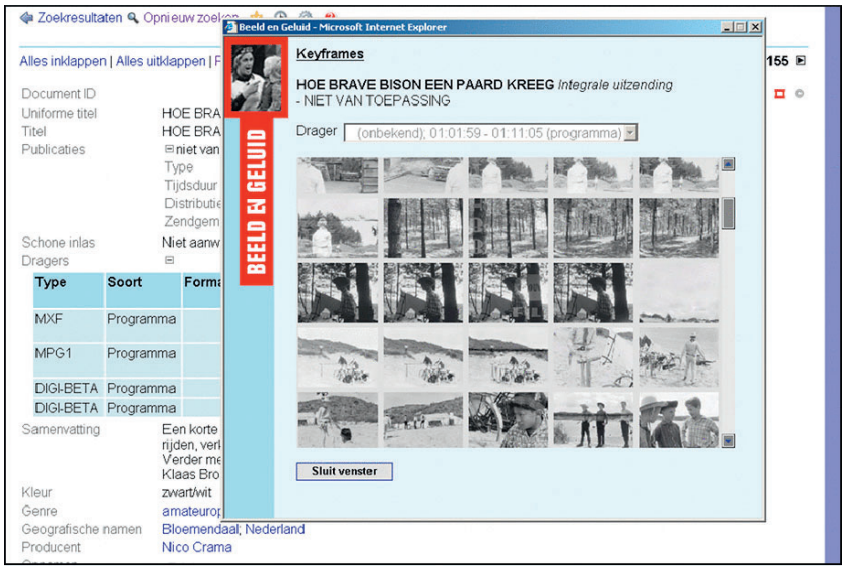


(a) advanced search form

(b) results delivered by search



(c) item details of clicked result



(d) keyframe preview of video content

FIG. 1. Screenshots of the interface of the archive's search engine.

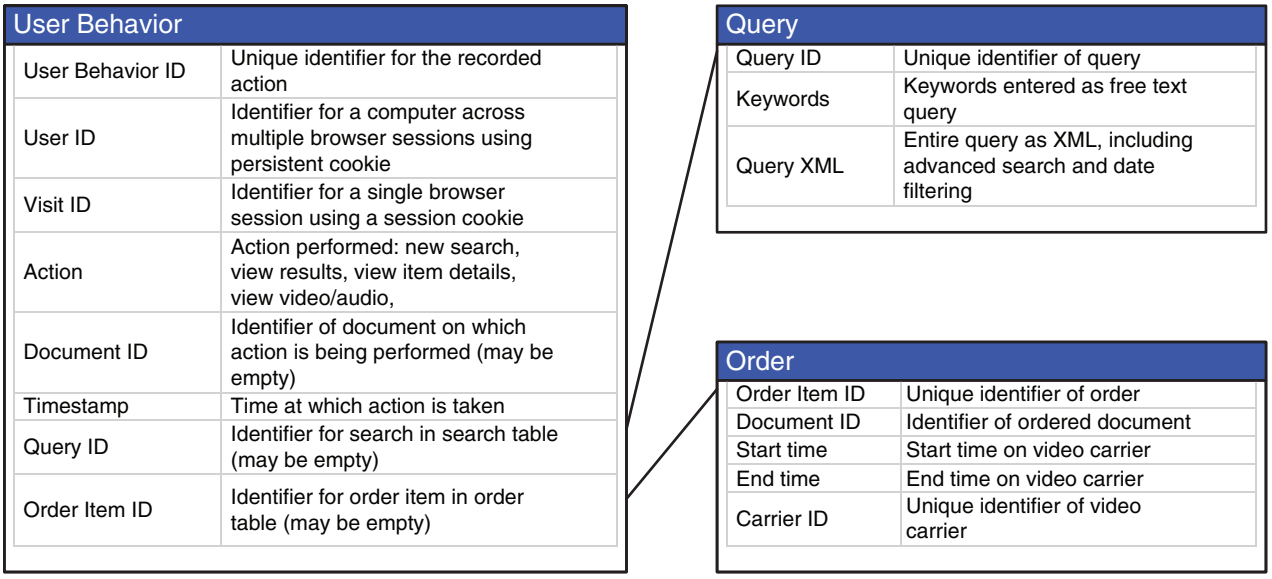


FIG. 2. Overview of the utilized tables and fields from the transaction logs. All actions are recorded in the *User Behavior* table. Additional information about queries and orders issued to the system is recorded in the *Query* and *Order* tables.

results broadcast on a specific date or within a set period; and an *advanced search* option that allows users to search on specific fields within the audiovisual catalog, such as the media format, the copyright holder, or thesaurus terms contained in the catalog annotations. After the search results have been displayed, they can be refined by using a list of suggested filtering terms displayed on the right-hand side of the result page (see Figure 1b). These suggested terms are based on the content of the different fields of returned results, allowing the user to refine the search without manually typing in values in the advanced search field. The user can navigate to several types of detailed result views: the catalog entry for the result as shown in Figure 1c, keyframes from the result as shown in Figure 1d, or a video or audio preview of the result in a media player. The latter two views are available only if a result is digitally available. Once a user has decided to purchase a result, or a fragment of footage from the results, the material can be ordered using an online form. Ordered video items are paid for on a per-second basis. Radio can be ordered only in chunks of at least 1 hour. Rates vary per copyright owner and per user subscription type, and some items can be purchased free of charge. Through iMMix, video material can be ordered as a fragment or as an entire program. Reasons to order a fragment rather than an entire program include: to increase efficiency during the editing process; budget restrictions; and copyright restrictions, which mean that only some parts of a program may be reused.

Audiovisual thesaurus. 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." The GTAA thesaurus is continually evolving, with new terms being added subject to an internal review process. At the time of this study, it contained approximately 160,000 terms, organized in six facets: *Location*, *Person*, *Name*, *Program Maker*, *Genre*, and *Subject*. *Location* describes either the place(s) where audiovisual material was recorded or the place(s) mentioned or seen in a broadcast. *Person* is used for proper names of people who are either seen or heard in a broadcast, or who are the subject of a broadcast. *Name* has the same function for named organizations, groups, bands, periods, events, and so on. *Program Maker* and *Genre* are used to describe a creator and a genre of a broadcast, respectively. The *Subject* facet is used to describe what a broadcast is about and what can be seen in the broadcast, and aims to contain terms for all topics that could appear in the audiovisual material in the archive, which makes its scope quite broad. Each term in the *Subject* facet is assigned to 1 or more of 15 categories, which we will specify in our analysis (discussed later). In addition, *Subject* terms may be associated with multiple synonyms.

Experimental Design

Data collection. Transaction logs from the archive's online search and purchasing system were collected between November 18, 2008 and May 15, 2009. The logs were recorded using an inhouse system tailored to the archive's online interface. The logs were stored in a database with multiple tables. For this study, we utilized three main tables, illustrated in Figure 2: the *User Behavior* table, which recorded information allowing users to be connected to actions performed on the Web site during a visit, the *Query* table, recording information about the queries issued to the system, and the *Order* table, recording information about

DOI: 10.1002/asi

orders placed through the system. As is the case with many online systems, the archive's search interface is periodically accessed by automated scripts. This phenomenon is typified by a large number of repeats of the same query. To minimize the impact of these nonhuman interactions on the query-log analysis, traces left by users issuing over 80 identical queries on a single day were removed from the logs. In addition, log entries generated by users known to belong to the archive's search system development team also were removed. In total, the logs contained 290,429 queries after cleaning. The transaction logs often reference documents contained in the archive's collection. 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.

Definitions. We define five key units that will play a role in our analysis: 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.

Session: A portion of a user's visit spent searching for items addressing the same subject or topic, as identified by overlapping words between queries (Jansen, Spink, Blakely, & Koshman, 2007). Session boundaries within a single visit are assigned when a query has no words in common with the previous query. Some studies have included a session timeout to identify session boundaries. We do not follow their example as in this study, users routinely view audiovisual material during the search process; this can lead to periods of inactivity even though the user is still fully engaged in the search process.

Query: A single information request issued to the archive, using any of the three previously described search options available in the interface: (a) *keyword search*, (b) *date filtering*, and (c) *advanced search*. A query also may be empty, such as when a user clicks the search button without entering information in any of the search boxes.

Term: A single *thesaurus term* or *title term* contained in a query, identified by matching phrases and words from the query to thesaurus entries and titles from clicked results. For example, if a user issues a query for "obama white house" and clicks a result containing the thesaurus entry "barack obama," then the query is considered to contain the thesaurus term "barack obama." This content-based coding approach is similar to that adopted by Jörgensen and Jörgensen (2005), in that content-bearing concepts are identified as terms, but differs in that the coding process is done automatically. As a consequence, term identification can be done over large numbers of queries, but not all terms can be identified. A detailed description of the term-matching approach and its evaluation is given in the Appendix.

Facet: The predefined category for a thesaurus term, as assigned within the GTAA inhouse audiovisual thesaurus described earlier. This is either *Location*, *Person*, *Name*, *Program Maker*, *Genre*, or *Subject*.

TABLE 1. Overall transaction log statistics.

Description	*n*
Search sessions	139,139
Queries	
Empty	15,675
Nonempty	274,754
Matched terms	
Title terms	72,620
Thesaurus terms	83,232
Orders	27,246

Order: A purchase request for audiovisual material issued to the archive. An order may be for an entire broadcast or a subsection of a broadcast. This is an unusual form of information that is available due to the dedicated professional nature of the archive.

Transaction Log Analysis

Now that we have described the methods behind our study, we move on to the transaction log analysis itself. Table 1 gives general statistics about the data collected in terms of the number of queries issued, query terms matched to the audiovisual thesaurus, search sessions, and the number of orders made. We will examine each of these statistics in detail.

Session-Level Analysis

We start by investigating the characteristics of sessions at the archive in terms of queries, result clicks, duration, and distribution over days of the week.

Distribution Over the Week. Figure 3 illustrates the distribution of sessions across the days of the week. As in Jörgensen and Jörgensen (2005), we find that the majority of sessions occur on Monday through Friday, reflecting the professional nature of the archive use. Moreover, fewer sessions occur on Fridays, reflecting an ongoing societal development in the Netherlands: Increasing numbers of people have 4-day work weeks (with the weekend starting on Friday).

Session Duration. The session statistics shown in Table 2 provide us with other insights into the characteristic behavior of users of the archive. There is considerable variation in session duration, with the longest search session lasting over 800 hours—approximately 1 month. The data include a number of search sessions that last more than 24 hours, caused by users remaining logged into their browser, leaving their workstation for 1 or more night, and returning to complete their session. The 2008 Christmas holidays seem to have been a contributing factor here, with users continuing to be logged on over the holidays. Consequently, the average session duration of approximately 18 minutes has a very high standard deviation of 4 hours. This corresponds to the

DOI: 10.1002/asi

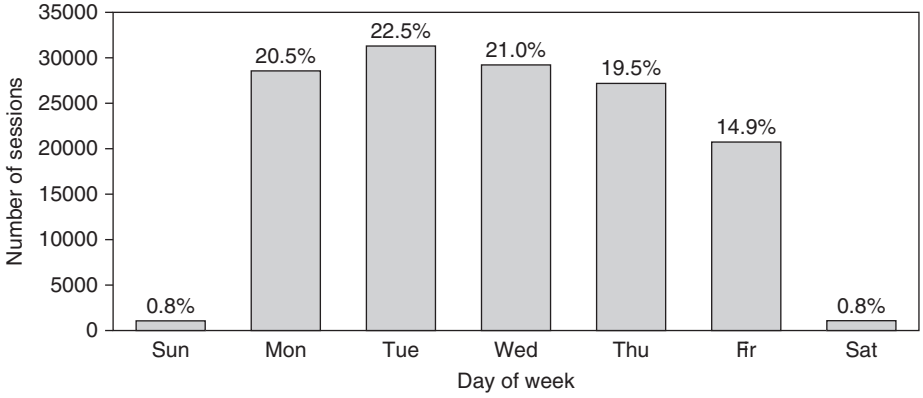


FIG. 3. Distribution of sessions across days of the week. Most activity takes place on weekdays.

TABLE 2. Session-level statistics over the entire transaction-log collection. Statistics are given over all sessions, and also according to whether an order was placed during a session.

			Session resulted in order?	
Variable	Measure	All sessions	Yes	No
Total sessions	Count	139,139	15,403 (11%)	123,736 (89%)
Queries	Median	1.0	1.0	1.0
	Mean	2.0	2.6	1.9
	σ	2.4	3.3	2.3
	Maximum	95.0	94.0	95.0
Result views	Median	1.0	1.0	1.0
	Mean	2.2	2.1	2.2
	σ	4.5	4.4	4.5
	Maximum	165.0	132.0	165.0
Orders	Median	0.0	1.0	0.0
	Mean	0.2	1.8	0.0
	σ	1.0	2.5	0.0
	Maximum	85.0	85.0	0.0
Duration (hh:mm:ss)	Median	00:00:59	00:07:47	00:00:43
	Mean	00:17:42	00:45:31	00:14:09
	σ	03:59:52	05:19:04	03:47:32
	Maximum	816:24:10	358:10:10	816:24:10

findings of Jansen et al. (2007), who found average session-duration lengths to be unstable while the large percentages of short sessions remained more stable. This suggests that the median duration of 59 seconds over all search sessions is a more representative indicator of session length. These results are similar to the median session durations of online multimedia search engines reported by Tjondronegoro et al. (2009) study, who found that 50% of the searches took less than 1 minute. Session durations reported at a professional image archive, on the other hand, were longer, with a median session time of 5 to 6 minutes (Jörgensen & Jörgensen, 2005).

Sessions Resulting in Orders. The discrepancy between session duration in the audiovisual archive and the session duration in the professional image archive can perhaps be explained by considering the sessions that resulted in an order separately from those where no order for audiovisual material was placed. The 11.1% of sessions that resulted in an order exhibit a much longer median duration of over 7 minutes, and are more consistent with the results of the professional image archive (Jörgensen & Jörgensen, 2005). The increased duration for sessions resulting in an order is not accompanied by a similarly large increase in the number of queries issued and results viewed. This suggests that the extra time spent is primarily on other tasks such as reviewing audiovisual content or placing the order. We will examine this in more detail, later in the analysis of orders placed to the system.

Number of Queries per Session. Figure 4 gives more insight into the numbers of queries issued per session. As implied by the median statistics in the previous paragraph, the majority of sessions contain one query. A number of sessions contain no query at all; these are sessions where, for example, a user navigated to the search page, but did not issue a query.

Query Modification. We also look at the development of queries within a session, examining the number of queries used in each session and how successive queries differed from previous queries in the same session. The results are shown in Table 3. Similarly to Jansen, Spink, and Saracevic (2000), we classify our queries as *initial*, *modified*, or *repeat*. An initial query is the first query issued in a session, and therefore represents the number of sessions where at least one nonempty query was issued. A modified query is a query that is different than the previous query. A repeat query is a query that is identical to the previous query. This may be due to the user retyping the same query words, but also can be caused by the user reloading the search page. The transaction logs do not allow us to distinguish between these underlying causes. The proportion of query modifications is relatively small as compared to Jörgensen and Jörgensen (2005), where 61% of queries were found to be modified.

Lessons Learned. The transaction logs contain a large number of short sessions as well as a small number of very long

DOI: 10.1002/asi

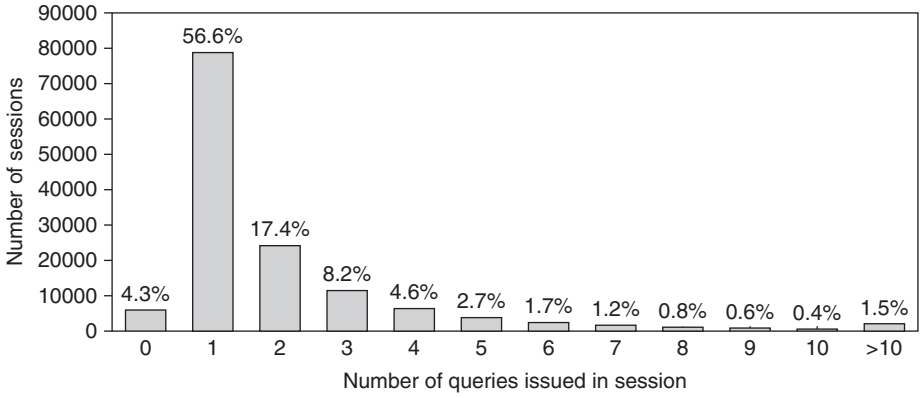


FIG. 4. Overview of the numbers of queries issued per session.

TABLE 3. Number of initial, modified, and repeat queries recorded in the search logs.

Modification type	No.	%
Initial	132,289	48
Modify	95,274	35
Repeat	47,191	17

TABLE 4. Breakdown of different search options used in the logged nonempty queries. Multiple search options may be employed in a single query.

Functionality	No.	%
Keyword search	264,270	96
Date filtering	62,647	23
Advanced search	23,332	9

sessions. These can sometimes last for weeks, as users may not log out of their browser while away from their workstation. Though most sessions are very short, sessions resulting in an order are generally much longer than are sessions that do not result in an order. The longer session duration is not accompanied by a corresponding increase in queries or result views. Most sessions consist of a single query.

Query-Level Analysis

Now we turn to an examination of the queries issued to the archive, studying in particular the utilization of the different search options afforded by the archive's interface and summarizing the most frequently issued queries.

Search Option Utilization. As described previously, the search interface provides users with multiple options for search: keyword search, date filtering, and advanced search on specific catalog metadata fields. A query may be specified using any or all of these search options. Table 4 gives a breakdown of the utilization of the available search options over the nonempty queries in the transaction logs. The clear majority of queries contain a keyword search. Interestingly, almost one fourth of the queries in the logs incorporate a date filter. The advanced search option, which allows users to search the catalog for thesaurus terms used to annotate broadcasts as well as for specific technical properties (e.g., copyright holder, physical format), was the least used of the three search options.

Frequent Queries. Table 5 separately lists the most frequently occurring queries for each of the three search options offered by the archive, with English language descriptions where appropriate. Of the top-20 most frequent keyword searches, 17 consist of the title or partial title of an audiovisual broadcast. Furthermore, the titles are mostly of Dutch news programs or Dutch current affairs shows. Advanced searches are commonly used to specify the media format of the results to be returned, with the large majority of common advanced queries in Table 5b specifying that only audiovisual documents available in the digital *Material eXchange Format* (MXF) should be returned. The date filter is commonly used to specify date ranges of a decade, which account for the five most frequent date ranges in Table 5c, or otherwise a single day.

Lessons Learned. Almost all queries issued to the archive contain keyword search terms, and an analysis of the most frequent searches indicates that common keyword searches tend to be derived from broadcast titles, especially news and current affairs broadcasts. The date filter is an often-used function, with results being filtered on date in almost one fourth of all queries. Advanced search appears to be used primarily for specifying technical characteristics of the desired results in terms of media format and digital rights. These determine, among other things, the availability and the cost of reusing the audiovisual material returned by the system. While advanced search offers the opportunity to specify terms contained in the audiovisual thesaurus, this option is not frequently used. Instead, these types of terms are often entered in the keyword search field, as we will see in the following section.

DOI: 10.1002/asi

TABLE 5. Most common searches, split according to search option into keyword search, advanced search, and date filter

(a) Most frequently occurring keyword searches accompanied by an English language description. Percentages are with respect to the total of 264,270 keyword searches.

Keyword search	*n*	%	Description
journaal	3,933	1.49	Title of news series
nova	1,741	0.66	Title of current affairs series
eenvandaag	1,340	0.51	Title of current affairs series
netwerk	1,103	0.42	Title of current affairs series
journaal 20	1,052	0.40	Title of news series
brandpunt	807	0.31	Title of current affairs series
jeugdjournaal	778	0.29	Title of news series
nos journaal	674	0.26	Title of news series
het klokhuis	517	0.20	Title of educational series
de wereld draait door	439	0.17	Title of talkshow
nos	416	0.16	Name of broadcasting company
polygoon	396	0.15	Title of news series
videotheek	393	0.15	Physical media location
andere tijden	369	0.14	Title of documentary series
twee vandaag	339	0.13	Title of current affairs series
kruispunt	337	0.13	Title of current affairs series
obama	317	0.12	Person name
klokhuis	284	0.11	Title of educational series
20 uur journaal	284	0.11	Title of news series
20	265	0.10	Title of news series

(b) Most frequently occurring advanced queries, given in the format *field name:value*. Percentages are with respect to the total of 23,332 advanced searches.

Advanced search	*n*	%
media format:mxf	5,337	22.87
copyright holder:nos	800	3.43
genre:documentaire	772	3.31
media format:digi-beta	728	3.12
copyright holder:vara	451	1.93
copyright holder:ncrv	447	1.92
copyright holder:vpro	398	1.71
copyright holder:tros	383	1.64
media format:mpg1	380	1.63
copyright holder:niet van toepassing	378	1.62
media format:vhs	354	1.52
copyright holder:teleac	331	1.42
copyright holder:eo	241	1.03
copyright holder:kro	229	0.98
copyright holder:nps	222	0.95
copyright holder:avro	205	0.88
media format:mxf, copyright holder:ncrv	188	0.81
genre:journaal	147	0.63
media format:wmv	140	0.60
media format:mpg1, media format:mxf	127	0.54

(c) Most frequently occurring date filters used in queries. Percentages are with respect to the total of 62,647 date-filtering searches.

Start date	End date	*n*	%
2000-01-01	2009-12-31	1,083	1.73
1980-01-01	1989-12-31	410	0.65
1990-01-01	1999-12-31	309	0.49
1970-01-01	1979-12-31	308	0.49
1960-01-01	1969-12-31	276	0.44
2009-03-12	2009-03-12	201	0.32
2009-02-25	2009-02-25	187	0.30
2009-01-26	2009-01-26	174	0.28
2009-01-20	2009-01-20	152	0.24
2008-11-05	2008-11-05	147	0.23
2009-01-09	2009-01-09	147	0.23
2008-01-01	2008-12-31	142	0.23
1950-01-01	1959-12-31	141	0.23
2008-11-26	2008-11-26	138	0.22
2009-03-09	2009-03-09	138	0.22
2008-11-18	2008-11-18	132	0.21
2009-01-18	2009-01-18	132	0.21
2009-02-17	2009-02-17	129	0.21
2008-11-25	2008-11-25	125	0.20
2008-12-31	2008-12-31	124	0.20

Term-Level Analysis

We now turn to an analysis of the query terms contained in the free text searches. We add structure to the query terms as previously described, by matching terms from the user query to titles and thesaurus entries in the results clicked by the user. The matched terms will be referred to as *title term* and *thesaurus term*, respectively. Of the 203,685 queries where users clicked a result during the session, 68,520 (34%) contained at least one title term, and 72,110 (35%) contained at least one thesaurus term. A total of 71,252 (35%) queries contained no title term or thesaurus term. The manual analysis of a subsample of queries showed that users sometimes type dates and program codes directly into the keyword search field; further detail is given in the Appendix.

Title Terms. The majority of the 20 most frequent title terms (see Table 6a) are the titles of news and current affairs programs. Many of the title terms in the top-20 are identical to frequent keyword searches in Table 5a. In addition to news

DOI: 10.1002/asi

TABLE 6. Most common terms occurring in keyword search, with title terms and thesaurus terms specified separately. An English language description is provided for each term.

(a) Most frequently matched title terms. Percentages are with respect to the total of 72,620 matched title terms.

Title term	Description	*n*	%
journaal	News series	9,087	12.5
nova	Current affairs series	2,568	3.5
netwerk	Current affairs series	1,597	2.2
eenvandaag	Current affairs series	1,440	2.0
het klokhuis	Educational series	1,318	1.8
jeugdjournaal	News series	951	1.3
brandpunt	Current affairs series	834	1.1
pauw & witteman	Talk show	820	1.1
twee vandaag	Current affairs series	779	1.1
de wereld draait door	Talk show	728	1.0
andere tijden	Documentary series	593	0.8
zembla	Documentary series	444	0.6
het zandkasteel	Children's series	409	0.6
studio sport	Sports news series	397	0.5
kruispunt	Current affairs series	373	0.5
pinkpop	Documentary	367	0.5
goedemorgen nederland	Morning show	326	0.4
man bijt hond	Magazine show	325	0.4
het uur van de wolf	Documentary series	294	0.4
tv show	Variety show	292	0.4

(b) Most frequently matched thesaurus terms. Percentages are with respect to the total of 83,232 matched thesaurus terms. The facet in which each term appears is also indicated.

Thesaurus term	Description	Facet	*n*	%
vs	U.S.	Location	1,002	1.2
nederland	The Netherlands	Location	941	1.1
nieuws	News	Genre	695	0.8
obama, barack	U.S. President	Person	663	0.8
amsterdam	Dutch capital	Location	610	0.7
wilders, geert	Dutch politician	Person	531	0.6
irak	Iraq	Location	515	0.6
voetbal	Soccer	Subject	460	0.6
beatrix (koningin nederland)	Queen of Netherlands	Person	414	0.5
bos, wouter	Dutch politician	Person	407	0.5
bush, george (jr.)	U.S. President	Person	371	0.4
willem-alexander (prins nederland)	Prince of Netherlands	Person	368	0.4
kinderen	Children	Subject	342	0.4
afghanistan	Afghanistan	Location	330	0.4
fortuyn, pim	Dutch politician	Person	279	0.3
ajax	Dutch soccer team	Name	272	0.3
rotterdam	Dutch city	Location	261	0.3
vrouwen	Women	Subject	260	0.3
balkenende, jan peter	Dutch prime minister	Person	257	0.3
israël	Israel	Location	251	0.3

and current affairs series, the list also includes television programs such as talk shows, educational programs, and documentaries. The list gives an indication of the types of programs that are often searched for by media professionals at the archive, but gives few clues as to what kind of content they desire. News broadcasts, documentaries, talk shows, and current affairs may cover a wide range of subjects and entities.

Thesaurus Terms. Table 6b shows the 20 most frequent thesaurus terms. The list contains a variety of locations, people, and subjects. Popular locations include the United States, the Netherlands, and Dutch cities, and popular people include political figures from the Netherlands and the United States, as well as Dutch royalty. The most frequently searched-for general subject is *soccer*, a popular sport in the Netherlands. Correspondingly, *Ajax*, the name of one of the dominant Dutch soccer clubs, is the most frequent nonperson and nonlocation name. Similar lists of frequently used search terms have been presented by Jansen et al. (2000, Jörgensen and Jörgensen (2005), and Tjondronegoro et al. (2009). However, a direct comparison is complicated by the fact that we include multiword terms whereas Jansen et al. (2000, p. 219) defined a term as "any series of characters bounded by white space." Both Jansen et al. (2000) and Tjondronegoro et al. (2009) examined queries on the Web; they contain mostly adult terms. Tjondronegoro et al. (2009) did observe that next to the adult material, names of people and song titles are frequently sought, which is in agreement with our findings. Jörgensen and Jörgensen (2005) found frequent search terms such as *woman*, *beach*, *computer*, and *man* in their study of search in a professional image archive. It is striking that these terms are very general as compared to the majority of frequent terms that occur in this study. Though the list does include general terms such as *women* and *children*, these are not in the majority.

Lessons Learned. The frequent use of broadcast titles as search terms implies that users are performing a large number of navigational, or known-item, searches. If we take the use of title terms in a query to be indicative of known-item search, the results are consistent with the findings of the early manual examination of e-mail requests to a film archive by Hertzum (2003), who found that 43% of the requests submitted to the archive were for known items. Many user queries contain terms that can be matched to audiovisual thesaurus entries, allowing us to investigate what types of content users are searching for within broadcasts. Popular thesaurus terms include countries and cities, the news genre, and the names of political figures.

Facet-Level Analysis

The frequently occurring terms in Table 6 provide us with an impression of the types of audiovisual terms for which users frequently search; however, the majority of thesaurus terms occur only a few times. This "long tail" phenomenon has been noted by many (e.g., Jansen et al., 2000). The semantic structure of the thesaurus can be used to abstract over the thesaurus terms and thereby also include infrequently used terms in the analysis. Two types of semantic structure were used: (a) the organization of all terms into facets and (b) the organization of terms in the *Subject* facet into categories.

DOI: 10.1002/asi

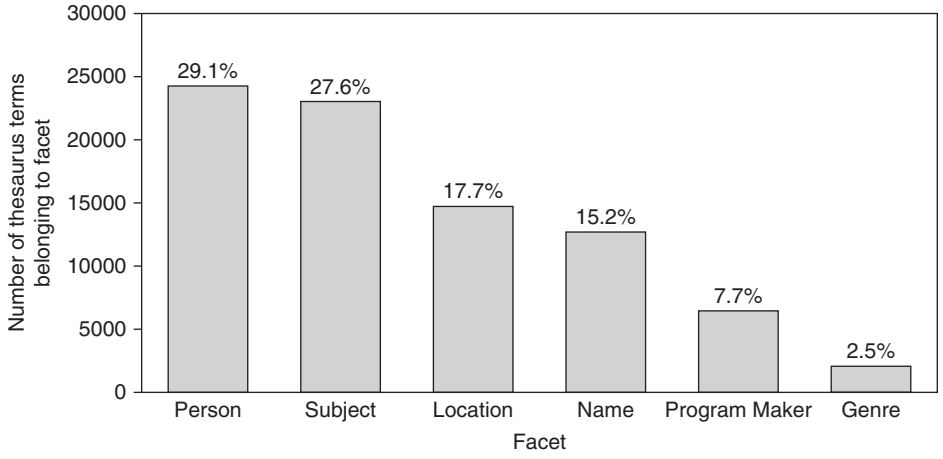


FIG. 5. Distribution of query terms matched to the GTAA thesaurus across the six different facets contained in the thesaurus. Percentages are with respect to the total of 83,232 thesaurus terms.

Frequencies of Facets. Figure 5 shows the cumulative frequency of each of the six thesaurus facets over all of the thesaurus terms identified in user queries. The *Person Name* and *Subject* facets were most frequently identified, together accounting for more than half of the thesaurus terms. They are followed by the *Name*, *Location*, and *Program Maker* facets Overall, *Genre* was the least frequently identified facet. The manual evaluation of the term-identification procedure in the Appendix indicates that locations and genres are easier to match than are terms from other facets, and that these therefore may be somewhat overrepresented in the analysis. Nevertheless, it is apparent that the proper names of people, places, program makers, and other proper names account for a large proportion of the thesaurus terms identified in the queries.

Facet Combination. User queries may contain multiple matched thesaurus terms, and in total, 13,532 queries contained more than one thesaurus term. Figure 6 itemizes the most frequently co-occurring facets in queries. The *Subject* facet appears frequently in combination with other facets—49% of the combinations consist of a *Subject* and either a proper name (*Location*, *Person*, or *Name*) or another *Subject*). An example of such a combination is a query consisting of the keywords "forest fire california," which is matched to the *Subject* "forest fires" and the *Location* "california."

Subject Categories. The organization of thesaurus terms belonging to the *Subject* facet into categories enables us to examine at a more detailed level what users search for in this facet. Table 7 shows how frequently each category occurs across the matched thesaurus terms. Each of the categories covers a broad range of subjects, as might be expected when attempting to cover the spectrum of audiovisual subjects in only 15 categories. In addition, some terms are assigned to multiple categories. Nevertheless, *governance and history*, *social issues*, and *economy* appear to be the most popular themes for *Subject* searches while *ethnology and geography*, *education*, and *science* each account for less than 2% of the *Subject* matches.

Lessons Learned. Proper names, of people, places, events, and more, are the most common type of thesaurus term identified in queries. Together, these account for 70% of the identified thesaurus terms issued to the archive. Approximately 28% of the identified terms are for more general subject-related thesaurus terms.

Order-Level Analysis

We turn to the orders issued to the archive by media professionals, examining their age, duration, and the amount of time typically taken to place an order.

Order Age. As is clear from the analysis so far, a primary function of the archive is to supply media professionals with reusable audiovisual material for new productions. Users with an account can order and purchase material from the archive. The transaction logs of the archive in this study are exceptional in that they contain detailed information about orders of audiovisual material placed through the online interface. In addition, the orders can be connected to catalog information about the associated audiovisual broadcasts. This allows us to determine, for example, the age of the audiovisual material relative to the date of purchase. Figure 7 shows that the majority of ordered items were broadcast more than 1 month before the date of purchase; 46% of ordered items were broadcast more than 1 year before the date of purchase.

Order Units. To provide a more detailed level of insight into individual orders, we categorize them according to the unit

DOI: 10.1002/asi

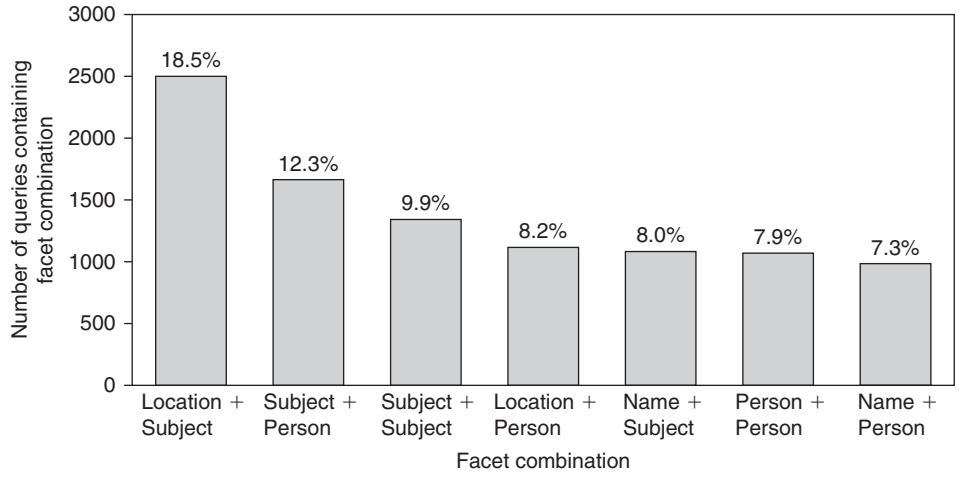


FIG. 6. The seven most frequently co-occurring combinations of facets when considering queries containing multiple thesaurus terms. Percentages are with respect to the 13,532 queries that contain multiple matched thesaurus terms. The most common combination of facets in a query is that of *Location* and *Subject*.

TABLE 7. Thesaurus-based categorization of matched thesaurus terms belonging to the *Subject* facet. The last column gives the three most frequently occurring thesaurus terms, with frequency, per category.

Subject category	*n*	%	Top thesaurus terms
Governance and history	5,313	17.1	election (212), police (146), terrorist attack (141)
Social issues	4,474	14.4	children (342), women (260), marriage (150)
Economy	3,782	12.2	factory (98), fashion (96), economic crisis (88)
Health and wel-lbeing	2,468	7.9	humor (118), obesity (74), babies (70)
Nature and environment	2,467	7.9	fire (92), evolution (79), natural gas (71)
Sport and recreation	2,422	7.8	soccer (460), skating (197), sport (82)
Art and culture	1,992	6.4	humor (118), cabaret (109), fashion (96)
Spatial environment	1,624	5.2	school (143), factory (98), opening (61)
Traffic and transport	1,616	5.2	car (156), train (76), bike (63)
Technology	1,567	5.0	car (156), natural gas (71), bike (63)
Communication and media	1,311	4.2	curios (104), aerial recordings (57), film (50)
Philosophy and religion	681	2.2	islam (67), church (67), muslim (63)
Ethnology and geography	602	1.9	commemoration (64), aerial recording (57), birthday (49)
Education	481	1.5	school (143), education (45), students (43)
Science	296	1.0	history (127), tests (20), scientific research (20)
All categories	31,096	100.0	

of purchase. The unit of material ordered by the professional falls into one of four categories:

Broadcast: The user orders an entire broadcast.
Story: The user orders a subsection of the broadcast that has been predefined by the archive catalogers as a cohesive unit.
Fragment: The user orders a subsection of the broadcast that has been dynamically defined by the user, typically by using the preview or keyframe view function.
Unknown: The unit of purchase cannot be determined due to missing or incomplete data.

Table 8 gives an overview of order statistics in terms of the number of orders, the time taken to order an item, the length of the final order, and the age of the ordered item, broken down according to the unit of purchase.

Time to Order. When considering the time taken to order an item, we see that some orders take an exceptionally long time to place (with a maximum time to order of over 300 hours). As observed in the session-analysis section, these unusually long times to order are due to users leaving the office without logging out of the session and returning days or weeks later to place the order. Therefore, median values are a more appropriate measure for comparing different units of order. At the median point, it takes users over twice as long to place an order for a fragment as it does for them to place an order for an entire broadcast. We attribute this to the need for a user placing a fragment order to manually review the video content to determine which portion to order. This is a time-consuming process.

Order Length. There is a strong demand for short pieces of audiovisual material, with 66.7% of the material ordered being for either predefined stories or manually defined fragments, both of which tend to be quite short (approximately 2 minutes). The duration of ordered items is further broken

DOI: 10.1002/asi

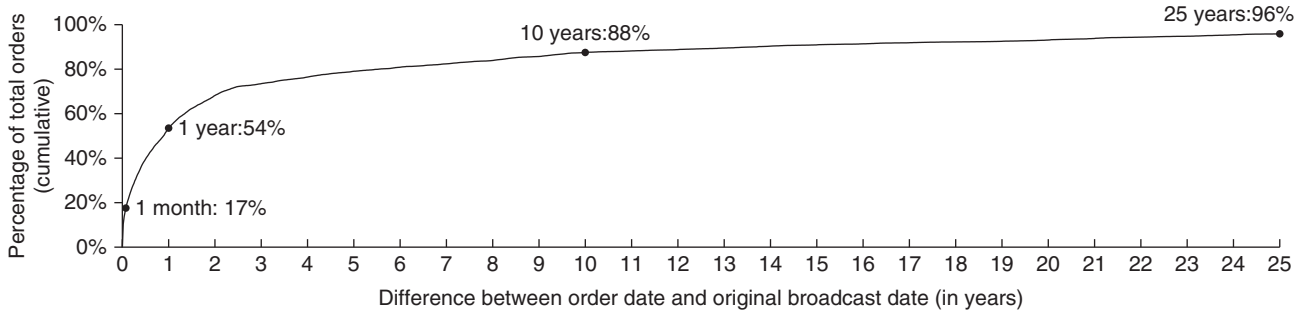


FIG. 7. Cumulative graph showing the elapsed time since the original broadcast date, at time of ordering. The majority of purchased multimedia items was ordered less than 1 year after the original broadcast.

TABLE 8. Order statistics, over all orders, and separately over individual units of orders.

			Order unit			
Variable	Measure	All orders	Broadcast	Story	Fragment	Unknown
Total orders	Count	27,246	8,953 (33%)	4,611 (17%)	13,329 (49%)	353 (1%)
Time to order	Median	00:06:38	00:03:44	00:05:51	00:09:27	00:05:51
(hh:mm:ss)	Mean	00:36:34	00:50:13	00:22:58	00:32:36	00:18:02
	σ	04:17:49	06:56:42	00:55:24	02:13:40	01:00:10
	Maximum	332:43:18	332:43:18	24:31:00	144:57:59	17:13:40
Order length	Median	00:03:45	00:25:03	00:02:04	00:02:00	00:04:22
(hh:mm:ss)	Mean	00:12:51	00:29:41	00:03:57	00:04:40	00:11:45
	σ	00:29:33	00:22:41	00:07:57	00:33:27	00:22:37
	Maximum	23:57:52	08:15:17	04:45:56	23:57:52	04:28:07
Order age	Median	00-10-18	01-01-25	00-05-02	00-11-28	00-01-16
(yy-MM-dd)	Mean	04-01-21	04-02-05	01-05-05	04-11-28	01-03-01
	σ	07-10-29	07-11-01	03-11-13	08-07-18	05-06-29
	Maximum	53-03-21	52-07-02	50-11-23	53-03-21	49-08-22

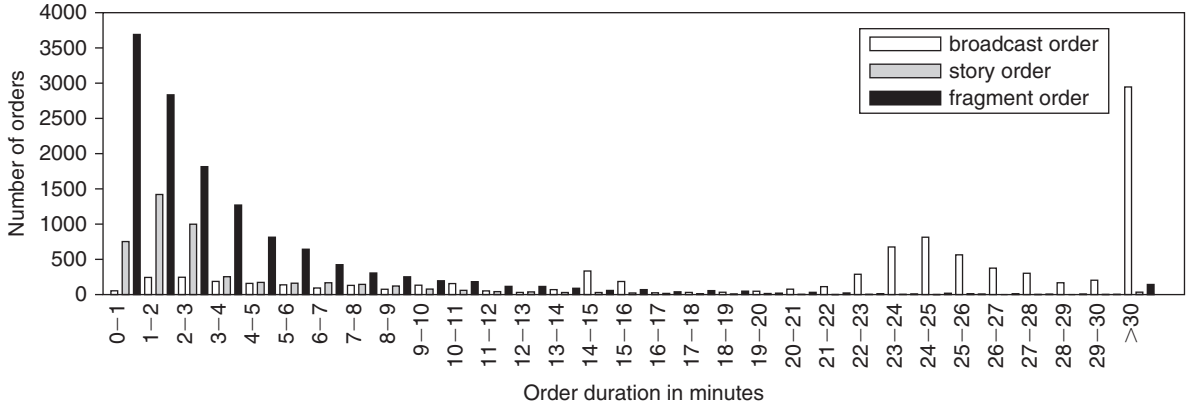


FIG. 8. Distribution of duration of fragment orders, in minutes. Note that the last column combines all orders with a duration of 30 minutes or more.

down in Figure 8, which shows fragment orders peaking at between 0 and 1 minute long, story orders peaking at between 1 and 2 minutes long, and broadcast orders peaking at 24 to 25 minutes in length—a common duration for a Dutch television broadcast.

Video Versus Audio. As shown in Table 9, nearly all orders were for video material rather than audio material. Audio results are more frequently clicked than purchased.

Lessons Learned. While one might assume that the logs of a professional broadcast archive are dominated by orders for very recent material to be used in news and current affairs programs, the order analysis shows that this is not the case. A little under half of the orders are for footage broadcast more than 1 year before the order date, and 12% of orders are for material that is over 10 years old. Users who manually specify the length of the audiovisual fragment they want to order take considerably longer than those who

DOI: 10.1002/asi

TABLE 9. Results clicked and ordered items divided into multimedia type: video and radio.

	Result clicks		Ordered items	
Multimedia type	*n*	%	*n*	%
video	287,291	91	26,618	98
audio	24,646	8	262	1
unknown	4,409	1	356	1

simply purchase an entire broadcast. In some cases, users purchase a selection from a broadcast that has been predefined by the archive's catalogers as a cohesive story. In these cases, users are able to place an order more quickly than when they themselves define the fragment. Manually defined fragments can be very short, often under 60 s in length. Internal interviews by the archive have indicated that this is often due to budgetary reasons: Users pay by the second for purchased material, depending on the copyright owner. Though the archive contains both video and audio material, almost all orders placed in the archive are for video material.

Discussion and Conclusions

We have presented a descriptive analysis of transaction logs from an audiovisual broadcast archive. The analysis was structured around four research questions, the answers to which we summarize here and follow with a discussion.

With respect to the question "What characterizes a typical session at the archive?," we found the typical session to be short, with over half of the sessions under 1 minute in duration. In general, there also were 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 7 minutes, but no increase in terms of the number of queries issued and results viewed.

To answer the question "What kinds of queries are users issuing to the audiovisual archive?," nearly all of the queries contained a keyword search in the form of free text while almost one fourth 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 (e.g., 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 some limitations, as terms can be identified only in sessions where users click at least one result, and even in these cases, a term can be identified only if it is present as a title or a 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. In a further analysis of the thesaurus terms, we used structured information from the thesaurus to identify different types of terms. Approximately one fourth of thesaurus terms consisted of general subjects such as *soccer*, *election*, and *child*. Another fourth consisted of the names of people, especially the names 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 both recent and historical material, with 46% of orders being for items that were broadcast more than 1 year before the order date. We identified three different units of ordering: *programs*, *stories*, and *fragments*. We saw that less than one 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. Just under half of the orders were for audiovisual fragments with a start and end time specified by the users themselves. The fragments were typically on the order of a few minutes in duration, with 28% of fragments being 1 minute or less. When users manually specified fragment boundaries, sessions typically took more than two and half times as long as when ordering an entire broadcast.

As might be expected from a group of professionals, it is apparent that the users of the audiovisual archive demonstrate a high level of expertise. In sessions where an order is made, users often issue only one query and view only one result before obtaining the audiovisual item. This, in combination with the high proportion of searches on program titles, and on specific dates, implies that the users often perform known-item (or target) search, knowing exactly which audiovisual item it is that they wish to obtain when initiating a session. A potential method by which the archive could assist users is to add auto-completion and spelling-correction functions tuned to titles and named entities to the search interface, as it is apparent that users often search on these types of information. Similarly, the users of the archive could be helped by online query-analysis tools that can parse dates and program codes in keyword searches. While almost one fourth of all queries used an explicit date filter, we observed in a sample of 200 sessions (see the Appendix) that users also tend to type dates and codes directly in the keyword search box; 21 dates and 38 program codes were entered in a total of 356 keyword searches.

Users often wish to order small fragments of broadcasts, but the audiovisual result pages show only short textual summaries of the entire audiovisual item that is returned for a search. There seems to be room for improvement here, as it currently takes users a relatively long time to place an order

DOI: 10.1002/asi

for a fragment. In some cases, predefined story segments are available in the archive; in these cases, it seems to take users considerably less time to place an order than when they themselves must define the fragment. We conjecture that the archive could assist users in decreasing their time to order by enabling fine-grained search within individual broadcasts. This could be done, for example, by using automatic annotation methods such as speech recognition (de Jong, Westerveld, & de Vries, 2007) and detection of semantic concepts that occur in a video (Snoek & Worring, 2009). Fine-grained search might allow professionals to jump more quickly to the appropriate section of the audiovisual material than they can by scanning through an entire broadcast using a media player or keyframe viewer, as is currently the case.

In future work, we plan more detailed examination of the transaction logs, for example, by studying the search behavior of individual users, and by investigating sessions where the user appears to have difficulty obtaining relevant results, typified by a large number of reformulations. The ordering information is an unusually strong indication of the relevance of video fragments to a search session, and it could serve as an indicator of ground truth in large-scale video retrieval and personalization experiments. Related to this is the use of order information to aid automatic annotation of video content. For example, click-through data have been used with some success to help train models for automatically annotating images (Tsikrika, Diou, de Vries, & Delopoulos, 2009). Result click data in the audiovisual archive encompass entire broadcasts instead of individual images and therefore seem unsuited for providing training data, but short fragment orders might be exploited to provide more specific training information. Finally, we plan to contrast this study with a qualitative examination of the underlying motivations and information needs of users, using research methods such as interviews and observation.

Acknowledgments

We are extremely grateful to the Netherlands Institute for Sound and Vision for making available the transaction logs used in this study and for providing access to internal reports and unpublished technical documents describing the Institute's organizational background. Without their openness and generosity, this study would not have been possible. We also thank the anonymous reviewers for their helpful comments on earlier versions of this article. This research was supported by the Netherlands Organization for Scientific Research (NWO) under Project Numbers 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802, and by the Dutch–Flemish research program STEVIN under Projects DAESO and DuOMAn (STE-05-24 and STE-09-12).

References

Armitage, L.H., & Enser, P.G.B. (1997). Analysis of user need in image archives. Journal of Information Science, 23(4), 287–299.

Broder, A. (2002). A taxonomy of Web search. SIGIR Forum, 36(2), 3–10.

Carman, M., Baillie, M., Gwadera, R., & Crestani, F. (2009). A statistical comparison of tag and query logs. In Proceedings of the 32nd Annual ACM SIGIR Conference (pp. 123–130). New York: ACM Press.

Christel, M.G. (2007). Establishing the utility of non-text search for news video retrieval with real world users. In Proceedings of the 15th International Conference on Multimedia (pp. 707–716). New York: ACM Press.

de Jong, F.M.G., Westerveld, T., & de Vries, A.P. (2007). Multimedia search without visual analysis: The value of linguistic and contextual information. IEEE Transactions on Circuits and Systems for Video Technology, 17(3), 365–371.

Edmondson, R. (2004). Audiovisual archiving: Philosophy and principles. Paris: UNESCO.

Enser, P. (1993). Query analysis in a visual information retrieval context. Journal of Document and Text Management, 1(1), 25–52.

Hertzum, M. (2003). Requests for information from a film archive: A case study of multimedia retrieval. Journal of Documentation, 59(2), 168–186.

Jansen, B.J. (2008). The methodology of search log analysis. In B.J. Jansen, A. Spink, & I. Taksa (Eds.), Handbook of research on Web log analysis (pp. 99–121). Hershey, PA: Idea Group Inc.

Jansen, B.J., Booth, D.L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. Information Processing & Management, 44(3), 1251–1266.

Jansen, B.J., & Pooch, U. (2001). A review of web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology, 52(3), 235–246.

Jansen, B.J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on web search engines. Journal of the American Society for Information Science and Technology, 58(6), 862–871.

Jansen, B.J., Spink, A., & Pedersen, J.O. (2004). The effect of specialized multimedia collections on web searching. Journal of Web Engineering, 3(3–4),182–199.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing & Management, 36(2), 207–227.

Jörgensen, C., & Jörgensen, P. (2005). Image querying by image professionals. Journal of the American Society for Information Science and Technology, 56(12), 1346–1359.

Kellar, M., Watters, C.R., & Shepherd, M.A. (2007). A field study characterizing web-based information-seeking tasks. Journal of the American Society for Information Science and Technology, 58(7), 999–1018.

Krippendorff, K. (1980). Content analysis: An introduction to its methodology. Beverly Hills, CA: Sage.

Mishne, G., & de Rijke, M. (2006). A study of blog search. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), Proceedings of the 28th European Conference on Information Retrieval (Vol. 3936 of LNCS, pp. 289–301). London, United Kingdom: Springer.

Monz, C., & de Rijke, M. (2002). Shallow morphological analysis in monolingual information retrieval for Dutch, German, and Italian. In Revised Papers from the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems (pp. 262–277). London: Springer-Verlag.

Oomen, J., Verwayen, H., Timmermans, N., & Heijmans, L. (2009). Images for the future: Unlocking value of audiovisual heritage. In Museums and the Web 2009: Proceedings. Toronto, Canada: Archives & Museum Informatics. Retreived February 24, 2010, from http://www.archimuse.com/mw2009/papers/oomen/oomen.html

Ozmutlu, S., Spink, A., & Ozmutlu, H.C. (2003). Multimedia web searching trends: 1997–2001. Information Processing & Management, 39(4), 611–621.

Panofsky, E. (1962). Studies in iconology. New York: Harper & Row.

Peters, T.A. (1993). The history and development of transaction log analysis. Library Hi Tech, 11(2), 41–66.

Porter, M.F. (2009). Dutch stemming algorithm. Retrieved January 25, 2010, from http://snowball.tartarus.org/algorithms/dutch/stemmer.html

Pu, H.-T. (2008). An analysis of failed queries for web image retrieval. Journal of Information Science, 34(3), 275–289.

Rice, R.E., & Borgman, C.L. (1983). The use of computer-monitored data in information science and communication research. Journal of the American Society for Information Science and Technology, 34(4), 247–256.

Rose, D.E., & Levinson, D. (2004). Understanding user goals in web search. In Proceedings of the 13th International Conference on the World Wide Web (pp. 13–19). New York: ACM Press.

Sandom, C.J., & Enser, P.G.B. (2001). Virami—Visual information retrieval for archival moving imagery. In Proceedings of the International Cultural Heritage Meeting (pp. 141–152). Milan, Italy: Archives & Museum Informatics.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1349–1380.

Snoek, C.G.M., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., & Worring, M. (2007). Adding semantics to detectors for video retrieval. IEEE Transactions on Multimedia, 9(5), 975–986.

Snoek, C.G.M., & Worring, M. (2009). Concept-based video retrieval. Foundations and Trends in Information Retrieval, 4(2), 215–322.

Tjondronegoro, D., Spink, A., & Jansen, B.J. (2009). A study and comparison of multimedia web searching: 1997–2006. Journal of the American Society for Information Science and Technology.

Tsikrika, T., Diou, C., de Vries, A.P., & Delopoulos, A. (2009). Image annotation using clickthrough data. In Proceedings of the Eighth International Conference on Content-Based Image and Video Retrieval (pp. 1–8). New York: ACM Press.

Wright, R. (2007). Annual report on preservation issues for European audiovisual collections. Retrieved Febraury 24, 2010, from http://www.prestospace.org/project/deliverables/D22-8.pdf.

Appendix

Identifying Title and Thesaurus Terms in Queries

In the Term-Level Analysis section, terms are identified by matching phrases and words within queries to metadata in clicked results. We include titles and thesaurus terms in the matching process, but exclude free-text descriptions in the clicked results as these cannot be linked to any knowledge source. The resulting matched titles and entries are referred to as *title terms* and *thesaurus terms*, respectively. The algorithm used to accomplish this is partially based on Snoek et al. (2007) and is outlined in Figure A1. To summarize, a term is identified if: (a) at least one result is clicked during the search session in which the query is contained, (b) a candidate query phrase or word is contained in a program title or a thesaurus entry from a clicked result, and (c) there are not multiple candidate titles and/or entries conforming to the previous two conditions. As a result, not all terms can be identified, but coverage is extensive: Of the 274, 754 nonempty queries, 203,685 were associated with at least one result click, and in turn, 132,433 queries were associated with at least one title term or thesaurus term.

Manual Identification of Query Terms

To evaluate the quality and coverage of the automatically matched terms, an evaluation set was created. Here, three annotators manually identified the title terms and thesaurus terms contained in the queries from a randomly selected sample of 200 search sessions. The annotation procedure was set up as follows: The annotator was presented with the queries and clicked results from a search session, as well as a copy of the audiovisual thesaurus. The annotator was asked to identify the different title and thesaurus terms contained in a query using the clicked results and the audiovisual thesaurus. When a term could not be identified in the thesaurus or the titles of the clicked results, the annotator was asked to mark this separately as an *unknown* term. Our sample of 200 search sessions contained a total of 356 queries; our annotators identified a total of 157 title terms, 249 thesaurus terms, and 109 unknown terms. The unknown terms were not contained in the audiovisual thesaurus or the titles, and included 21 dates and 38 codes that were entered by the user into the keyword search. The remaining unknown terms consisted of concepts not present in the thesaurus, such as *night recording*. We investigated the agreement between the three annotators on a hold-out set of 30 sessions using Krippendorf's alpha (Krippendorff, 1980); we found the average pairwise agreement to be 0.81.

Evaluation

The performance of our automatic term-identification method on the manually annotated sample of sessions is shown in Table A1. Performance is measured in terms of precision (i.e., the number of automatically identified terms that were labeled as correct, divided by the total number of automatically identified terms), and recall (i.e., the number of automatically identified terms that were labeled as correct, divided by the total number of terms that were labeled as correct). Overall, the automatically identified terms are accurate, with on average 9 of 10 of the identified terms being correct. Some types of terms are easier to identify correctly than are others; title terms and thesaurus terms from the *Subject*, *Location*, and *Person* facets all have a precision of over 0.95 whereas thesaurus terms from the *Genre* and *Program Maker* facets have a precision of less than 0.50. Recall is similarly variable, with over a relatively large proportion of the manually identified *Location* and *Genre* terms being returned by the automatic method. Less than one in two of the manually identified terms are identified for the *Person*, *Program Maker*, and *Subject* facets.

DOI: 10.1002/asi

Input
Queries (all keyword searches)
Thesaurus entries (all thesaurus entries in the GTAA, including synonyms and other morphological variants)
Titles (all program titles contained in the archive catalog)
Result click data (mapping from each query to the results clicked in the query session, and from each clicked result to the thesaurus entries and titles contained in that result)
Output
Set of query terms (titles and thesaurus entries from the clicked results that match or contain phrases within the queries)
Step 0: Preprocessing
- associate each thesaurus entry in the collection with any synonyms that may be contained in the thesaurus
- *for each* query, title, and thesaurus entry in the collection
  - strip punctuation, diacritics
  - remove frequently occurring stop words
  - stem words using the Porter stemming algorithm for Dutch (Porter, 2009)
  - split compound words using a compound splitter adapted from (Monz & de Rijke, 2002)

Step 1: Selection of Candidate Terms
- *for each* query
  - associate the query with clicked terms in the form of titles and thesaurus entries contained in the clicked session results
  - *for each* clicked term
    - count how many times the clicked term appears in the clicked results

Step 2: Processing of Query Phrases
- *for each* query
- create a set of all possible phrases within the query that maintain the sequence ordering (The longest phrase will be the entire query, the shortest phrases will be the individual words contained within the query.)
- order phrases by length, so that the phrase with the most words comes first
- *for each* query phrase
  - initialize empty set of matched terms
  - *if* the query phrase is identical to (exactly matches) at least one clicked term
    - add all identical clicked terms to the set of matched terms
  - *else, if* the query phrase is contained in (phrase matches) at least one clicked term
    - add all container clicked terms to the set of matched terms
  - *if* the set of matched terms contains exactly one term
    - **add the matched term to set of query terms**
    - remove all query phrases with words overlapping current query phrase from further processing
    - *go to* next query phrase
  - *if* the set of matched terms contains more than one term
    - select the matched terms that occur the most frequently in clicked results
    - *if* there is a single matched term that occurs most frequently in clicked results
      - **add the single most matched term to set of query terms**
      - remove all query phrases with words overlapping current query phrase from further processing
      - *go to* next query phrase
    - *if* multiple terms occur most frequently in clicked results, the query term is ambiguous
      - remove all query phrases with words overlapping current query phrase from further processing
      - *go to* next query phrase
  - *go to* next query phrase
- *go to* next query

FIG. A1. Process to identify title terms and thesaurus terms contained in user queries. Phrases within a query are matched to candidate titles and thesaurus entries from clicked results. When a match is found, all words from the query phrase are removed from the term-identification process.

TABLE A1. Evaluation of automatic term matching using precision and recall, based on a sample of 356 queries from 200 sessions. *No. correctly matched* indicates the number of automatically matched terms that also were manually identified, *No. matched* indicates the total number of automatically matched terms, and *No. correct* indicates the total number of manually identified terms.

Term type	Facet	No. correctly matched	No. matched	No. correct	Precision	Recall
Title	n/a	108	114	157	0.95	0.69
Thesaurus	Genre	1	13	1	0.08	1.00
	Location	34	35	40	0.97	0.85
	Name	31	33	59	0.94	0.53
	Person	22	23	56	0.96	0.39
	Program Maker	3	7	7	0.43	0.43
	Subject	42	42	86	1.00	0.49
All terms	n/a	241	267	406	0.90	0.59

DOI: 10.1002/asi