Journal of Documentation, vol. 59, no. 2 (2003), pp. 168-186
Author version. Published version at: https://doi.org/10.1108/00220410310463473

Requests for Information from a Film Archive:

A Case Study of Multimedia Retrieval

Morten Hertzum

Centre for Human-Machine Interaction, Risø National Laboratory, Denmark

morten.hertzum@risoe.dk

Contact details

Author: Morten Hertzum

Email: mhz@ruc.dk

Address:

November 19 - December 18, 2002:

Department of Computer and Information Sciences
University of Strathclyde
Livingstone Tower
26 Richmond Street
Glasgow G1 1XH
UK

As of January 1, 2003:

Computer Science
Roskilde University
Universitetsvej 1, bldg 42.1
P.O. Box 260
DK-4000 Roskilde
Denmark

Autobiographical note

Morten Hertzum has a Ph.D. in computer science and a commitment to empirically informed research. His research interests are within human-computer interaction (HCI) and studies of information-seeking behaviour in different contexts. After four years at Risø National Laboratory, Morten Hertzum is now associate professor in computer science at Roskilde University, Denmark; email: mhz@ruc.dk.

Word count: 8517 (excluding this page but including appendix, references, and tables)

Requests for Information from a Film Archive:

A Case Study of Multimedia Retrieval

Morten Hertzum

Centre for Human-Machine Interaction, Risø National Laboratory, Denmark

morten.hertzum@risoe.dk

Abstract. Multimedia retrieval is a complex and to some extent still unexplored area. Based on a full year of email requests addressed to a large film archive this study analyses what types of information needs real users have and how these needs are expressed. The findings include that the requesters make use of a broad range of need attributes in specifying their information needs. These attributes relate to the production, content, subject, context, and screening of films. However, a few attributes – especially title, production year, and director – account for the majority of the attribute instances. Further, as much as 43% of the requests contain no information about the context that gives rise to the request. The current indexing of the archived material is restricted to production-related attributes, and access to the material is, thus, frequently dependent on the archivists' extensive knowledge of the archived material and films in general.

Keywords: Information needs; Email requests; Multimedia retrieval; Need attributes; Film archives; Information-seeking behaviour

Introduction

Most work on information retrieval assumes that the stored material is textual. However, a substantial portion of the material held in archives and other collections is non-textual or multimedia. For example, newspapers have image databases, the police has text and photo archives of previously convicted persons, TV stations archive their programmes, manufacturing companies have databases with product documentation including CAD (Computer-Aided Design) drawings, and film archives store films, newsreels, and other film-related material to preserve cultural heritage and enable future research. Furthermore, increasing amounts of multimedia material is made available on the Web. To benefit from these masses of material we need techniques for retrieving the comparatively few items that are relevant to a specific person in a specific

situation. The present study seeks to inform the design of such techniques by analysing how users of a national film archive express their information needs: What is included in their requests and what is notably absent?

The data analysed in this study consist of information requests addressed to Deutsche Film Institut (DIF), an archive of twentieth century European films. The requests, submitted by email, are for texts (e.g., dialogue lists and censorship cards), images (e.g., photos of actors), sound (e.g., the music from a film), video (e.g., a video-copy of a film), as well as analyses (e.g., of the religious symbolism in the film "Metropolis"). Anecdotal evidence from a focus group interview gives the impression that there are many out-of-the-ordinary requests for things like "photos of women in a knight's costume". This study investigates a full year of email requests to determine what types of requests are the more frequent and to explore the diversity as well as the common characteristics of the ways in which the requesters express their information needs. The attributes used by the requesters in specifying their information needs are identified, discussed, and contrasted with the access points in the archive's film database and those which can be extracted by automatic indexing techniques. The International Federation of Film Archives (1991) has published cataloguing rules for films but they reflect a film archivist's view of films. The purpose of this study is to analyse what actual users of a comprehensive, multimedia film collection request and how they express their requests.

Related work

Vast amounts of multimedia material exist in physical, analog, and other non-digital formats and recently inexpensive storage media have also made it feasible to create large digital collections of multimedia material. The increased availability of multimedia material is accompanied by a need for indexing techniques that support effective retrieval. This calls for research on multimedia retrieval techniques and empirical studies of why, when, and how people search for multimedia material.

Multimedia retrieval

Irrespective of media, the traditional means of enabling retrieval from large repositories is *intellectual indexing*. In intellectual indexing human indexers use controlled vocabularies or natural language to express the content and subject of texts, images, sound, and video. This has led to the development and application of numerous indexing schemes, but it is at the same time well-documented that different indexers tend to index the same pieces of material rather differently (e.g., Markey, 1984; Sievert & Andrews, 1991). Nevertheless, the retrieval performance resulting from intellectual indexing has remained a challenging baseline for automatic indexing.

For non-text material, automatic indexing is a rather recent possibility and intellectual indexing is in general superior. For text, automatic and intellectual indexing produce different results but it is becoming an increasingly common view that the two types of indexing are, on balance, about equally effective (Anderson & Pérez-Carballo, 2001). It should, however, be noted that the more recent comparisons of intellectual and automatic text indexing are typically based on TREC (Text REtrieval Conference) data, in which news material is somewhat over-represented and, especially, technical material is poorly represented (Spark Jones, 2000).

With respect to automatic indexing of *text*, the so-called partial-match techniques are founded on the idea that word-distribution statistics can guide the extraction of index terms (see Belkin & Croft, 1987). For example, the frequency with which a word appears in a text tells something about how central that word is to the subject of the text, and the number of texts in which a word appears indicates how well that word distinguishes texts in a collection from each other. Over the years automatic indexing techniques based on the extraction of individual words from texts have proved surprisingly effective. In fact, the only vocabulary control that has consistently yielded definite advantages is to reduce words to stems and incorporate simple synonym relations (Croft, 1987). The scalability of current partial-match techniques is evident from the search engines on the Web. The effectiveness of the techniques is, however, modest in that many relevant texts are typically not retrieved whilst a number of non-relevant ones are (Blair, 2002; Spark Jones, 2000). Furthermore, user studies provide plenty of evidence that text retrieval is often experienced as difficult and that many searches fail altogether (e.g., Borgman, 1996; Tonta, 1992).

In automatic indexing of *images* (often termed content-based image indexing) the basic approach is to count the number of pixels in different colours. On that basis information about, for example, texture can be extracted by clustering neighbouring pixels into small regions with a similar colour distribution. The techniques that are currently most refined are retrieval by colour, texture, shape, and overall image similarity (Idris & Panchanathan, 1997; Rasmussen, 1997). Automatic image indexing has proven effective in domains such as trademark retrieval and face finding and through the wide use of a few general-purpose systems, such as the QBIC system (Flickner et al., 1995). Further, many images have captions, which can be indexed by means of text-indexing techniques. Srihari and Zhang (1999) demonstrate that an image and its caption are often utilised to communicate different parts of the total message; for example, what a person looks like and what she has accomplished. This illustrates the potential gains of a multimedia approach, as opposed to indexing either the image or the caption.

With respect to *sound*, automatic speech indexing can be approached as speech recognition followed by the application of text-retrieval techniques (Renals & Robinson, 2000). While a number of speech recognisers give acceptable recognition accuracy for clean speech in a controlled environment, their performance degrades when they are applied to real situations, particularly in noisy environments (Gong, 1995). Further, speech recognition is not simply a matter of recognising the spoken words (or phonemes) since much information is communicated through prosody, for example emotion and rising tone for questions (Shneiderman, 2000). In addition to speech, work is also being done on music retrieval. However, little of this work involves multi-voice music, none of it tackles the problems of realistically large databases, and evaluation methods are currently at a crude stage (Byrd & Crawford, 2002).

For the purposes of automatic indexing, *video* can be considered as a sequence of images (or shots). This turns video retrieval into shot boundary detection followed by the application of image-retrieval techniques. It is, however, no trivial task to convert raw video into a sequence of shots. Shot boundaries are often not sharp cuts, and the types of shot transitions are very different in, for example, documentaries and music videos (O'Toole et al., 1999). Once detected, individual shots can be indexed based on colour, texture, and spatial relationships. Further, shape and shot sequences can reveal object and camera motion (Idris & Panchanathan, 1997). By viewing video as a soundtrack with associated moving images it becomes possible to utilise sound retrieval techniques for video retrieval. Further, some video material is subtitled and, thus, allows for the application of text-retrieval techniques.

User studies

While automatic text indexing is an established field with numerous studies of both algorithms and user behaviour, the automatic indexing of images, sound, and video has only recently been made practically relevant by the developments in digital storage media. Specifically, research is only beginning to accumulate on why people search for multimedia material, in what contexts they intend to use it, and how they pose their queries (Chen & Rasmussen, 1999). Existing studies of user behaviour in multimedia settings provide many valuable observations but few firm conclusions.

Jörgensen (1998) investigated the image attributes typically noted by people involved in viewing images, describing images to a retrieval system, and describing images from memory. She found that a wide range of attributes were mentioned but that the more frequent related to objects, people, social status, colour, body parts,

locations, and activities. While attributes such as objects and colour are factual, many images were described by means of more interpretive attributes (e.g., social status and activity), which concern the story of the image.

Enser and McGregor (1993) analysed 2722 requests addressed to the Hulton Deutsch Collection, which is an image archive covering news, historical landscapes, portraits, and other subcollections used by the press. The main finding was that the requests could be mapped onto four categories defined by two binary distinctions: unique or non-unique and refined or unrefined. Only unique unrefined requests (e.g., "Edward VIII" as opposed to "a king" and "Edward VIII looking stupid") were easily satisfied by the indexing scheme of the archive. Generally, the indexing scheme served as a pointer to regions of the collection, requiring further browsing to identify relevant images.

Goodrum and Spink (2001) analysed 33149 queries for image and video material on the Web and found that they contained an average of 3.74 terms. This is slightly more than the average number of terms reported in studies of text searching on the Web. The data also suggested that image queries were modified more than text queries. It is however unknown whether the modifications were query reformulations or reflected separate information needs expressed sequentially by the same user. The total set of terms appearing in the queries was very diverse. Over half of the terms appeared in only a single query while the most frequent terms appeared in less than 9% of the queries.

Markkula and Sormunen (2000) studied journalists' use of a digital photo archive. Almost half of the requests were for photos of people. The number of requests concerning themes was limited, and the journalists claimed they only considered such requests when they had time for a lengthy search process. Trivial access points were sufficient for many requests (e.g., requests for named persons) and these requests generally gave satisfying results. The journalists' main search strategy was browsing, though it was not well supported by the retrieval system. Among the motivations for browsing was the ease with which it accommodates supplementary requirements such as currency; for example, photos of other than the current season were hardly ever selected.

Sandom and Enser (2002) surveyed eleven British film archives with respect to their cataloguing practice and analysed a sample of 1270 requests from the clients of these archives. The requests were predominantly about retrieving footage that featured specifically named persons, places, or events. Such information was, however, not systematically recorded in the catalogues, which displayed a marked lack of consistency across archives. The archives also faced a large backlog of uncatalogued footage. Further, the film archivists in the survey all

expressed the view that automatic indexing does not at present offer a generally effective alternative to the intellectual indexing currently performed by the archives.

Turner (1990) described how the National Film Board of Canada indexes its large collection of stock footage, which mostly consists of outtakes from the Board's own productions, newsreels, and war footage. Access to the footage is provided by means of standard attributes such as title and production year as well as by subject and more special attributes such as camera angles and time periods (winter, dawn etc.). Part of the indexing is done up front by the selectors who capture footage for inclusion in the collection, the rest is done later by indexers. Main challenges faced by the Board are the continuous reviewing/reworking of the subject indexing as new material comes into the collection, and the provision of some support for browsing.

The film archive

Deutsche Film Institut (DIF) is one of the largest cinematic institutions in Germany. DIF is a non-profit organisation sponsored by, among others, the German Federal Ministry of the Interior, the German broadcasting authorities, and leading organisations in the German film industry. Established in 1949, DIF has built a collection of more than 11000 classic German silent and early sound films, German versions of foreign films, short films, film clippings, and documentaries. In addition, DIF has an extensive collection of film-related materials such as newspaper cuttings, censorship cards, film programs, photos, and posters. DIF is open to the public as well as to organisations and industry, it engages in and encourages film-related research, and it promotes film culture nationally and internationally.

The foundation of DIF is its collection and the activities that enter into maintaining and managing it. Together with preservation of the archived materials these activities create the conditions for the user services, research projects, and other activities. The about 23 staff members spend a large part of their working hours searching the collection for information requested by users and directing users to other information sources when the resources at DIF have been exhausted. Research projects may be initiated by user needs, archivists' personal interests, and the institutional goals of the archive. As an example of research initiated and driven by the archive's status as a national institute, DIF has taken on the responsibility of creating a filmography of all German films since 1946.

The data set

The data collected for this study were the information requests sent to DIF by email during the year 2000. Analysis of emails is a completely unobtrusive method, as opposed to for example observation of user-archivist conversations. Thus, the ecological validity of the data is very high. Regarding the reliability of the data, the DIF archivists consider the email requests representative of the total body of information requests received by phone, face-to-face contacts, and postal letter in addition to email. It should be noted that the analysis involves the requests only, the answers to the requests are not included in the analysis.

The data analysis involved two passes. First, 128 requests were examined, annotated, and categorised according to a coding scheme that was developed in parallel with the examination of the requests. The outcome of this bottom-up analysis was a coding scheme founded on the actual data. Second, the full set of 275 requests were examined and categorised according to the final coding scheme. The coding scheme covers the type of request, the context giving rise to the request, the overall topic of the request, the types of material requested, and the attributes used in specifying the information need. The full coding scheme is not included in this paper; rather, selected categories are defined as they become relevant to the analysis.

The analysis of the emails is supplemented with a focus group interview of DIF archivists (see Pejtersen et al., 2001). The interview provides background information about the archive, its staff, their primary work tasks, and the tools used in performing these tasks.

Results and discussion

The 275 email requests range from very specific requests that could be answered right away to complex requests where the requested information was provided in a stepwise manner as the request was clarified and the material collected. To give an initial feel for the requests it can be mentioned that the contexts from which the requests arise comprise student work/theses (22%), festivals and exhibitions (10%), family-related research and events (9%), academic research and teaching (8%), commercial activities (7%), and other (1%). However, 43% of the requests contain no contextual information.

Context and focus

The requests suggest a distinction between context and focus. Contextual information describes the context in which the outcome of a request is to be used (i.e., *why* the information is needed). Focus information describes

the focus of a request in terms of the specified need attributes (i.e., *what* is specifically requested). The value of the distinction between contextual information and focus information is that both are of key importance for most types of searches but in the studied email requests one of them is often absent. Table I combines the context/focus distinction with a classification of the requests according to the four types of searches identified by Meadow (1992):

- *Known-item retrieval*. The requester knows what records are wanted and can specify them by means of searchable attributes. Example: "The Austrian movie 'Der Feldherrnhügel' by E. Marischka, 1953". The requester will recognise the desired records, if seen.
- *Fact retrieval*. The requester is looking for specific information, but without necessarily knowing where to look for it. Example: "Where and when did the silent-movie actress Lya Mara die?" It is not certain what terms to use for searching, but some initial candidates are readily available.
- *Subject retrieval*. The requester is looking for information on a subject in general. Example: "What are the reasons why 80% of the people who went to the cinema in Germany in 1998 saw an American, rather than a European, film?" There is no one way to describe the subject and no one way the desired information will be represented.
- *Exploratory retrieval*. The requester intends to find out what kinds of information are available, not to answer a specific question. Example: "If you are selling video-copies of silent movies, please send me a catalogue with the titles of the films you have available".

The archive receives requests of all four types and, in addition, a small number of 'Other' requests, which are more about establishing collaboration than about retrieving information. Table I shows that 117 (43%) of the analysed requests provide a focus only and that these requests are distributed across all four search types. These 117 requests provide no information about the reasons why the requested information is needed and, consequently, it is very difficult for the archivists to form an interpretation of these requests and assess the relevance of the archived material. Half of the known-item and fact requests provide a focus only. If the information provided in these requests is sufficient to unambiguously identify the desired material there is, however, no need for contextual information in these cases. It is more notable that the focus-only requests also make up 25 (28%) of the subject requests. An example of these requests is one asking for all information about a specific film:

> You are my last option. I am looking for material about the film "Schindler's List": reviews, studies of specific scenes, scene analyses; in short, everything there is to be found about this film. I would be very grateful for your help.

In some cases expressions like "everything there is to be found" are to be taken literally. In other cases they are not (rather, the requester is not aware of the masses of material available). To clarify such requests the archivists enter into a dialogue with the requesters about the types and amounts of material that are available and the context in which the request is being made (see Abels, 1996). The archivists use this information to form an impression of the amount of time it is reasonable to spend on a request. This way contextual information has a large impact on the kind of search that will be carried out and, thus, on the results that will be obtained.

The context that gives rise to an information need is an inherent part of the requester's understanding of his information need. When a requester externalises an information need by putting it into writing he also creates a representation of the need that is divorced from this understanding. In Taylor's terminology this compromises the information need – whereas some aspects of the need are retained others are inevitably left out (Taylor, 1968). The email requests are posed to a human so there is no retrieval system that forces the requesters to specify what they need but leave out information about why they need it. Especially for subject requests the situation is rather that the archivists must have at least some contextual information before they can respond. Further, the requests are posed in writing so we must presume that the requesters try to provide all the information they consider necessary for others to understand and respond to their information need. This can be contrasted with oral communication where the first turns are often merely intended to open the dialogue. It seems as if the requesters are insufficiently aware of how important contextual information is to others who want to understand the requests. The requesters frequently provide a context-free specification of their information need and, thereby, erroneously assume that their information need can be communicated and understood in isolation from its context.

A substantial number of the requests are about retrieval of known items (see Table I). The prototypical known-item request concerns a person who wants to purchase a video-copy of a film that is specified by its title, production year, and often also its director. In fact, subject retrieval, which is the focus of most information-retrieval research, accounts for only 32% of the requests. To the extent that the distribution of the email requests onto search types can be generalised to other real-world settings it identifies a need for a more unified

approach to retrieval than the current, and quite sharp, distinction between information retrieval which focuses on subject retrieval and data retrieval which focuses on known-item retrieval.

Take in Table I

Categories of need attributes

The attributes used in specifying the information needs can be grouped into six categories (a more detailed listing of the need attributes and their frequencies appears in the Appendix):

- *Production-related attributes*, which include title, production year, director, actors, film music, book on which film is based, production country, film company, and type.
- *Screening-related attributes*, which include cinema, TV channel, exhibition/festival, date or period where the film was shown, programs, and film listings of contemporary newspapers.
- *Content-related attributes*, which concern the identifiable entities appearing in a film. These attributes include location, time, persons, events, and objects.
- *Subject-related attributes*, which concern the message or meaning of a film discerned by interpretation. These attributes include theme, genre, author intentions, and emotional experience.
- *Context-related attributes*, which include reviews, censorship material, film magazines, film sections of newspapers, film festivals, film societies, film industry, the public, and society.
- *Other*, which includes any other attribute used in specifying an information need.

Table II shows the number of requests containing need attributes from these categories. For example, the upper left cell shows that 81 of the known-item requests contain production-related attributes only and that an additional 28 known-item requests contain production-related attributes in combination with attributes from other categories. Across all 275 requests a total of 196 (71%) requests contain attributes from only a single attribute category.

It is evident that production-related attributes – such as title and production year – are central to the requesters' specification of their information needs. As much as 139 (51%) of the requests are specified by means of production-related attributes only, and this category is also the one most frequently appearing in combination

with other categories. In fact, only 25% of the requests contain no production-related attributes. Screening-related attributes always appear in combination with other categories and can thus be classified as supplementary. This may reflect that the requesters recognise the difficulties involved in maintaining updated information about when films appear in cinemas and on the various TV channels. Attributes concerning content, subject, and context appear about equally often and more often than not in requests that combine attributes from several categories. The attributes concerning content, subject, and context capture how the requests express what films are of, what films are about, and how films are inscribed in the surrounding society (see Shatford, 1986). In spite of the vast span of issues covered by these attributes, only 38% of the 275 requests contain attributes from one or more of these three attribute categories. This is in contrast to Sandom and Enser (2002) who report that especially content-related attributes were very frequent in the requests received by the surveyed film archives. Such discrepancies probably reflect differences in, among other things, the requesters' tasks. Most requests in Sandom and Enser (2002) were commercial in nature, whereas requests from students and academic researchers were rare. In contrast, the 275 requests received by DIF contained few commercial requests and comparatively more requests from students and academic researchers. It should, however, be noted that the limited number of requests containing attributes relating to the content, subject, and context of films could also be an attempt by the requesters to phrase their requests in terms of attributes they perceive as suited for the currently available retrieval tools.

Attributes specifying the content, subject, and context of films are more common in the 88 subject requests than in the full set of 275 requests. It is, however, notable that only slightly more than half (56%) of the subject requests contain need attributes relating to content, subject, and context. Subject retrieval is conventionally held to require that archived material is indexed with terms describing its subject matter. Such indexing is however vulnerable to change (see, e.g., Turner, 1990) and this may make it attractive to replace subject retrieval with known-item retrieval (if the searcher knows the collection or subject well) or with the browsing activities characteristic of exploratory retrieval (if the searcher is new to the collection and subject). The email requests include 33 instances of subject requests that are expressed solely in terms of production-related attributes. In these cases the subject is typically 'converted' into known-item searches for a couple of films that are known or believed to be pivotal to the subject. The requester can either believe that these prime examples are sufficient or they are used as a label (Ingwersen, 1982) that signifies the full subject. In the former case the requester is probably overestimating the extent to which information relevant to her information need clusters in a few records. Thus, the request is likely to be too narrow to capture all material relevant to the subject. In the latter

case the requester uses the prime examples to concretely specify what relevant material looks like and to provide the archivists with a starting point. Thus, the request is set up for similarity retrieval and will fail unless the archivists realise this and are able to correctly identify how the examples relate to the full subject. This is prohibitively difficult in the situations where the request contains no contextual information.

Take in Table II

Need attributes versus access points

Moving from categories of need attributes to individual need attributes we find that a total of 33 different attributes appear in the requests (see Appendix). The five most frequent attributes are title (appearing in 53% of the requests), director (35%), production year (34%), actor (16%), and production country (11%). These are all production-related attributes and thus reflect the dominance of this attribute category.

As much as 57 (21%) of the requests contain no other need attributes than the title, production year, and/or director of a film, and an additional 123 (45%) of the requests contain one or more of these top-three attributes in combination with other attributes. The previously mentioned subject request for all material about "Schindler's List" shows that title, year, and/or director are not only used in known-item requests. A number of requesters expect to be able to use title, year, and/or director as an access point to all material about a film. This expectation is, however, not as straightforward as it may seem. Quite a few of the archived films exist in several versions with different titles and production years. Some of these versions are simply foreign language replicates of the original film, others are brought about by censorship decisions that banned different scenes at different times or in different countries, and still others are complete remakes by a different director. In many situations it is no simple matter to determine whether two films should be considered independent films or versions of the same film. Further, considerable amounts of film footage are missing or in bad condition due to old age. Many versions exist only in incomplete copies and it may be a genuine research task to determine whether individual scenes from different pieces of footage were included in a film or cut out. Consequently, standard bibliographic access points such as title and production year cannot be assigned unambiguously to all the archived material.

Even if title, year, and director did unambiguously identify all the material about a film then the amount of material available about some films would still necessitate additional access points to allow for more specific requests. Additional access points would also be needed to handle requests that cut across film boundaries. The email requests show that people make use of a diverse range of attributes in specifying their information needs. These attributes, of which only some relate to the subject of the films, could form the basis for the development of an indexing scheme based on the contents of actual requests. Compared to Jörgensen (1998) it is noteworthy that subject-related attributes account for a relatively small subset of the attributes used in specifying the email requests. This may be a consequence of the number of requesters with considerable film knowledge – that is, a characteristic of the film-archive domain – or it may reflect an attempt by the requesters to phrase their requests in terms of attributes they perceive as suited for searching.

At present the archivists are registering the films in a database they have designed themselves because they find that current standards for film registration are too coarse-grained. The film database is evidently biased toward careful registration of the facts and figures regarding the production of films. The database supplies, for example, 169 types of ways in which a person may have contributed to a film and 20 types of titles for a single film. Conversely, the database contains no information about the content, context, and screening of films, and apart from what can be deduced from the film title(s) there is no subject information either. The only information remotely related to content and subject is that a film can be assigned one of six genres and it can be registered whether the film is fiction or non-fiction. The film database may be immensely valuable to people with an interest in the production of European films but the archived material contains information about numerous additional issues. Whereas the film database reflects a rather narrow view of the material, the email requests comprise a broader spectrum of views on the types of issues to which the archive may contribute valuable information. For example, some requests treat films as a source of data about political and societal issues. Currently, the film archive can only be utilised for these additional purposes if the archivists possess the knowledge necessary to translate these requests into searchable attributes. The less the film database contains in addition to factual fields about the production of films, the more the treatment of requests that transcend this focus becomes dependent on the archivists' domain expertise and knowledge of the contents of the archive.

Implications: from email requests toward multimedia retrieval systems

The people who submit the email requests know they are communicating with a human. This affects the way in which they formulate their requests. After stripping header and signature information, the average length of an

email request is 111 words. Thus, the number of words used in specifying the actual contents of the email requests is much larger than the number of query terms typically specified during the use of retrieval systems. Jansen and Pooch (2001) report that searches on OPACs (Online Public-Access Catalogues) typically contain 1-2 query terms, Web searches typically 2 query terms, and traditional information-retrieval searches typically 6-9 query terms. This shows that people find it natural to write much more in communicating with a human. It may also indicate that people have a better intuition about what a human will need to know in order to help them than about what kinds of information a search engine needs to perform effectively. Comparing the length of the email requests with the spread of need attributes they contain it is, however, apparent that many of the words in the requests are spent on formulating full sentences rather than providing mere search terms. As previously mentioned, 57 (21%) of the requests contain no other need attributes than one or more of title, production year, and director. This amounts to much less than 111 words, although a number of the requests contain the title, production year, and/or director of more than one film and the titles usually consist of more than one word each. The length of the email requests also provides some basis for believing that people will be willing to formulate longer queries to retrieval systems if they experience that this leads to improved performance. It is generally assumed that information-retrieval algorithms benefit from longer, more refined queries (e.g., Salton & McGill, 1983; Sparck Jones, 2000; but see also Blair, 1980).

While requesters may be willing to provide information retrieval systems with lengthy and refined queries, the collections of DIF and many other film archives are currently not available in digital format. Hence, intellectual indexing is often the only option available. This could be considered a temporary phenomenon, but it is more likely to remain a permanent condition because resources are scarce and digitisation is a laborious process. Even if full digitisation is assumed, it is still beyond the current capabilities of multimedia retrieval to respond satisfactorily to the email requests on the basis of automatic indexing, at least as regards the non-text materials. One of the requests for image and video material provides an illustrative example of this:

> I am looking for pictorial material about East Germany in the period 1955-1959. My interests concern everyday-life in the countryside, particularly the agricultural production cooperatives, the youth, and 'ordinary' folks. […] I need this information for my thesis (diploma) which I began in September. I am a costume design student and will sketch the costumes for a film. For that purpose, I utterly need contemporary pictorial material.

This request is clear and rich in concrete details but it is difficult to imagine how it can be transformed to a query consisting of low-level image attributes such as textures and shapes. Rather, the attributes in the request are examples of the types of information one could hope to find in the caption of an image or the soundtrack of a video (If present, such captions and soundtracks can be thought of as a special instance of intellectual indexing). Actually, nearly all the need attributes that appear in the email requests have this 'textual orientation' (see Appendix). The most notable exceptions are requests that specify *events* or *objects* in films, for example the longest film kiss and a (named) Viking ship. These exceptions relate directly to the visual appearance of objects in the image or video and can, in principle, be described in terms of low-level image attributes. It is, however, an open question whether techniques for automatic image and video indexing will reach the sophistication necessary to process such queries effectively. In accordance with the limited number of requests that specify objects and events, Greisdorf and O'Connor (2002) find, in an analysis of image perception, that the majority of the terms people used in describing the contents of images were not visibly present in the images. This included some of the references to colours. The word "green" was, for example, used to describe several black and white images of trees. Most of the terms that were not visibly present were, however, emotive terms and other semantic-level attributes, such as "'ordinary' folks" in the request above. This suggests that interpretation and other information beyond the pixel-distribution content of still and video images will often be crucial to effective multimedia retrieval.

The film-archive domain is complex because many of its constituent elements are fluid. While the actual film footage and other multimedia materials are given, analyses of films involve interpretations that must embrace a web of cultural, political, and societal issues. These analyses can be challenged by other interpretations that give priority to a different set of issues. Another source of fluidity is that a central unit of analysis – the film – is often not readily available but has to be constructed from a number of incomplete versions, individual scenes, censorship documents, and other materials. Disagreements and ambiguity may therefore arise over what is actually referred to by standard access points such as title and production year. As a result film archivists are involved in creating new knowledge – about films, what they tell about real-world issues, and how they should be indexed. In such an environment the archivists are very valuable as knowledgeable information sources and as intermediaries between requesters and the archive collection with its materials and indexes (Hertzum et al., 2002). If requesters are to be able to search the collection themselves it will be necessary to substantially extend the indexing of the films. It is, however, questionable whether such extended indexing can in practice be made sufficiently elaborate and self-contained to yield the same quality of search results when requesters

search themselves as when the archivists act as expert intermediaries (Hertzum et al., 2002; Pejtersen et al., 2001).

Conclusion

Multimedia indexing and retrieval is a complex and to some extent still unexplored area. This study has analysed a full year of email requests addressed to a large film archive and looked at what types of information needs real users have and how these real information needs are expressed in terms of need attributes. It is found that:

- 75% of the requests concern subjects and known items, with known-item requests being slightly more frequent than subject requests. The email requests contain an average of 111 words, which is many times more than the number of words in typical queries to retrieval systems. This may indicate that people have a better intuition about what humans need to know in order to help than about what kinds of information a retrieval system needs to perform effectively.
- 43% of the requests contain no information about the context that gives rise to the request. It is impossible for the archivists to judge whether the actual information needs are captured well by these requests, which include 28% of the subject requests. Thus, unless the archivists initiate a dialogue these requesters do not exploit that they are interacting with a human rather than a retrieval system.
- 38% of the subject requests are specified by means of production-related attributes only. This suggests a label effect where many requests are expressed in terms of an example of a film that is pivotal to the subject. While these requests call for similarity retrieval there is a considerable risk that they will be treated as known-item requests and thus fail to retrieve most of the relevant material.
- 21% of the requests contain no other need attributes than the title, production year, and/or director of a film (and many more requests contain these three attributes in combination with others). This indicates the importance of these three attributes but also reflects that many requesters – unrealistically – assume that these attributes are unambiguous and can be used as access points to all information about a film.

A broad range of 33 different need attributes is used in specifying the information needs. Collectively these attributes relate to the production, content, subject, context, and screening of films. The current indexing of the archived material is, however, restricted to production-related attributes, and the gap between the requests and these attributes is bridged by the archivists on the basis of their extensive knowledge of the archived material

and films in general. If the archive is to become less dependent on the knowledge of individual archivists it is necessary to extend the indexing of the films. Given the complexity of the film domain and the fluidity of several of its constituent elements, it is however questionable whether such extended indexing can in practice be made sufficiently elaborate and self-contained to enable requesters to perform effective searches themselves, unless the requesters are film specialists *and* know the archive well. It seems as if considerable judgement and expertise is necessary in interpreting many of the requests and in bridging the gap between the requests and the materials in the archive. At least as regards the non-text materials, it is also beyond the capabilities of contemporary multimedia retrieval to automatically index the materials in ways that match the need attributes contained in the email requests. The most viable way to improve retrieval from the archive is probably to acknowledge the archivists' capabilities as expert intermediaries and direct retrieval systems at supporting the archivists in their work with the collection and with requests from the users of the archive.

Acknowledgements

This study was supported by the Danish National Research Foundation through its funding of the Centre for Human-Machine Interaction and by the European Commission's Information Society Technologies Programme through its funding of the Collate project on collaboratories for annotation, indexing, and retrieval of digitised archive material (IST-1999-20882). I wish to thank my colleagues at Risø – Hanne Albrechtsen, Hans Andersen, Verner Andersen, Bryan Cleal, and Annelise Mark Pejtersen – and the people at Deutsche Film Institut for their contributions to this study.

Appendix: Need attributes and their frequencies

Take in Table III

References

Abels, E. G. (1996), "The e-mail reference interview", *RQ*, Vol. 35 No. 2, pp. 345-358.

Anderson, J. D. & Pérez-Carballo, J. (2001), "The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing", *Information Processing & Management*, Vol. 37 No. 2, pp. 231-254.

Belkin, N. J. & Croft, W. B. (1987), "Retrieval techniques", in *Annual Review of Information Science and Technology*, Vol. 22, ed. M. E. Williams, Elsevier, Amsterdam, pp. 109-145.

Blair, D. C. (1980), "Searching biases in large interactive document retrieval systems", *Journal of the American Society for Information Science*, Vol. 31 No. 4, pp. 271-277.

Blair, D. C. (2002), "Some thoughts on the reported results of TREC", *Information Processing & Management*, Vol. 38 No. 3, pp. 445-451.

Borgman, C. L. (1996), "Why are online catalogs still hard to use?", *Journal of the American Society for Information Science*, Vol. 47 No. 7, pp. 493-503.

Byrd, D. & Crawford, T. (2002), "Problems of music retrieval in the real world", *Information Processing & Management*, Vol. 38 No. 2, pp. 249-272.

Chen, H.-L. & Rasmussen, E. M. (1999), "Intellectual access to images", *Library Trends*, Vol. 48 No. 2, pp. 291-302.

Croft, W. B. (1987), "Approaches to intelligent information retrieval", *Information Processing & Management*, Vol. 23 No. 4, pp. 249-254.

Enser, P. G. B. & McGregor, C. G. (1993), *Analysis of Visual Information Retrieval Queries*, British Library R&D Report No. 6104, British Library Board, London.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. & Yanker, P. (1995), "Query by image and video content: The QBIC system", *IEEE Computer*, Vol. 28 No. 9, pp. 23-31.

Gong, Y. (1995), "Speech recognition in noisy environments: A survey", *Speech Recognition*, Vol. 16 No. 3, pp. 261-291.

Goodrum, A. & Spink, A. (2001), "Image searching on the Excite web search engine", *Information Processing & Management*, Vol. 37 No. 2, pp. 295-311.

Greisdorf, H. & O'Connor, B. (2002), "Modelling what users see when they look at images: A cognitive viewpoint", *Journal of Documentation*, Vol. 58 No. 1, pp. 6-29.

Hertzum, M., Pejtersen, A. M., Cleal, B. & Albrechtsen, H. (2002), "An analysis of collaboration in three film archives: A case for collaboratories", in *CoLIS4: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science*, eds. H. Bruce, R. Fidel, P. Ingwersen & P. Vakkari, Libraries Unlimited, Greenwood Village, CO, pp. 69-83.

Idris, F. & Panchanathan, S. (1997), "Review of image and video indexing techniques", *Journal of Visual Communication and Image Representation*, Vol. 8 No. 2, pp. 146-166.

Ingwersen, P. (1982), "Search procedures in the library – analysed from the cognitive point of view", *Journal of Documentation*, Vol. 38 No. 3, pp. 165-191.

International Federation of Film Archives (1991), *The FIAF Cataloguing Rules for Film Archives* (compiled and edited by H. W. Harrison), Saur, Munich, Germany.

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*, Vol. 52 No. 3, pp. 235-246.

Jörgensen, C. (1998), "Attributes of images in describing tasks", *Information Processing & Management*, Vol. 34 Nos. 2&3, pp. 161-174.

Markey, K. (1984), "Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials", *Library & Information Science Research*, Vol. 6 No. 2, pp. 155-177.

Markkula, M. & Sormunen, E. (2000), "End-user searching challenges indexing practices in the digital newspaper photo archive", *Information Retrieval*, Vol. 1 No.4, pp. 259-285.

Meadow, C. T. (1992), *Text Information Retrieval Systems*, Academic Press, San Diego, CA.

O'Toole, C., Smeaton, A., Murphy, N. & Marlow, S. (1999), "Evaluation of automatic shot boundary detection on a large video test suite", in *CIR 99: The Challenge of Image Retrieval. Proceedings of the 2nd UK Conference on Image Retrieval*, University of Northumbria at Newcastle, Newcastle, UK. Available at http://www.compapp.dcu.ie/~asmeaton/pubs-list.html (consulted July 31, 2002).

Pejtersen, A. M., Albrechtsen, H., Cleal, B., Hansen, C. B. & Hertzum, M. (2001), *A Web-based – Multimedia – Collaboratory: Empirical Work Studies in Film Archives*, Risø Report No. Risø-R-1284(EN), Risø National Laboratory, Roskilde, Denmark. Available at http://www.risoe.dk/rispubl/SYS/ris-r-1284.htm.

Rasmussen, E. M. (1997), "Indexing images", in *Annual Review of Information Science and Technology*, Vol. 32, ed. M. E. Williams, Information Today, Medford, NJ, pp. 169-196.

Renals, S. & Robinson, T. (eds.) (2000), "Accessing information in spoken audio" (Special issue), *Speech Communication*, Vol. 32 Nos. 1&2.

Salton, G. & McGill, M. J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Sandom, C. J. & Enser, P. G. B. (2002), *VIRAMI: Visual Information Retrieval for Archival Moving Imagery*, Library and Information Commission Research Report No. 129, Re:source: The Council for Museums, Archives and Libraries, London.

Shatford, S. (1986), "Analyzing the subject of a picture: A theoretical approach", *Cataloging & Classification Quarterly*, Vol. 6 No. 3, pp. 39-62.

Shneiderman, B. (2000), "The limits of speech recognition", *Communications of the ACM*, Vol. 43 No. 9, pp. 63-65.

Sievert, M. C. & Andrews, M. J. (1991), "Indexing consistency in Information Science Abstracts", *Journal of the American Society for Information Science*, Vol. 42 No. 1, pp. 1-6.

Spark Jones, K. (2000), "Further reflections on TREC", *Information Processing & Management*, Vol. 36 No. 1, pp. 37-85.

Srihari, R. K. & Zhang, Z. (1999), "Exploiting multimodal context in image retrieval", *Library Trends*, Vol. 48 No. 2, pp. 496-520.

Taylor, R. S. (1968), "Question-negotiation and information seeking in libraries", *College and Research Libraries*, Vol. 29 No. 3, pp. 178-194.

Tonta, Y. (1992), "Analysis of search failures in document retrieval systems: A review", *Public-Access Computer Systems Review*, Vol. 3 No. 1, pp. 4-53.

Turner, J. (1990), "Representing and accessing information in the stockshot database at the National Film Board of Canada", *Canadian Journal of Information Science*, Vol. 15 No. 4, pp. 1-22.

Table I. Distribution of requests onto search types

Search type	**Focus only**	**Focus and context**	**Total frequency**
Known-item retrieval	66	51	117 (43%)
Fact retrieval	16	19	35 (13%)
Subject retrieval	25	63	88 (32%)
Exploratory retrieval	9	18	27 (10%)
Other	1	7	8 (3%)
Total	117	158	275 (100%)

Table II. Categories of need attributes. For each category (production-related, screening-related, etc.) the table gives the number of requests containing attributes from this category only and, in parentheses, the additional number of requests in which attributes from this category appear in combination with attributes from other categories.

Search type	**Production-related**	**Screening-related**	**Content-related**	**Subject-related**	**Context-related**	**Other**
Known-item retrieval	81 (+28)	0 (+10)	1 (+10)	0 (+11)	4 (+4)	2 (+1)
Fact retrieval	16 (+4)	0 (+0)	4 (+0)	0 (+1)	9 (+2)	2 (+1)
Subject retrieval	33 (+30)	0 (+5)	4 (+13)	3 (+19)	7 (+12)	4 (+1)
Exploratory retrieval	9 (+2)	0 (+1)	0 (+4)	2 (+6)	4 (+2)	5 (+0)
Other	0 (+2)	0 (+0)	1 (+0)	0 (+0)	2 (+0)	3 (+2)
Total	139 (+66)	0 (+16)	10 (+27)	5 (+37)	26 (+20)	16 (+5)

Table III

[No caption; the heading of the Appendix, which consists of nothing but this table, provides the caption. In the paper all references are to the Appendix, not to Table III.]

Need attributes	Number (and percentage) of requests containing attribute
Production-related	**205 (75%)**
Titles (German, English, etc.)	146 (53%)
Nickname (not the title but a broadly understandable 'name')	3 (1%)
Production year	94 (34%)
Director	96 (35%)
Actor	43 (16%)
Other person on cast list	2 (1%)
Composer of film music	7 (3%)
Music used in film	4 (1%)
Author of book on which film is based	23 (8%)
Title of book	13 (5%)
Production country	30 (11%)
Film company	7 (3%)
Type (silent movie, black/white, etc.)	28 (10%)
Screening-related	**16 (6%)**
Cinema	3 (1%)
TV channel	7 (3%)
Date or period where the film was shown	9 (3%)
Programs and film listings of contemporary newspapers	2 (1%)
Exhibition or festival	1 (0%)
Content-related	**37 (13%)**
Location (e.g., urban/countryside, country, etc.)	19 (7%)
Time (when the film takes place)	13 (5%)
Person	11 (4%)
Event (e.g., the longest film kiss)	9 (3%)
Object	6 (2%)
Subject-related	**42 (15%)**
Theme	25 (9%)
Genre (detective story, 'Berg film', etc.)	12 (4%)
Author intentions	4 (1%)
Emotional experience	4 (1%)
Context-related	**46 (17%)**
Reviews and censorship material	5 (2%)
Film magazines / film sections of newspapers	14 (5%)
Film festivals, film societies, and their members	3 (1%)
Film industry – individual companies and film industry as a whole	24 (9%)
The public, society	7 (3%)
Other	**21 (8%)**