# The Use of Metadata for the Rendering of Personalized Video Delivery[1]

**W. Klippgen, T.D.C. Little, G. Ahanger, and D. Venkatesh**

Multimedia Communications Laboratory
Department of Electrical and Computer Engineering
Boston University, Boston, Massachusetts 02215, USA
(617) 353-9877, (617) 353-6440 fax
*tdcl@bu.edu*

MCL Technical Report 12-01-1996

**Abstract**–Information personalization is increasingly viewed as an essential component of any front-end to a large information space. Personalization can achieve both customization of the presentation of information as well as tailoring of the content itself.

In this paper we investigate techniques for personalizing information delivery based on metadata associated with diverse information units including video. We begin with a survey of approaches to information personalization and the requirements for this task. Subsequently we present a characterization of the use of metadata to facilitate video information personalization.

**Keywords:** personalization, information retrieval, multimedia authoring, video delivery, time-variant filtering, sequential filtering.

[1]In *Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Data,* Amit Sheth and Wolfgang Klas (eds.), McGraw Hill, 1998, pp. 287-318. This work is supported in part by the National Science Foundation under Grant No. IRI-9502702.

# 1 Introduction

The need for some form of information personalization is clear: it is unreasonable for individuals to traverse the vast amount of information currently available via electronic means. Efforts to deliver information in a broadcast mode consume enormous bandwidths with no guarantee of interest by the end recipient. Personalization attempts to bridge the gap between the users and the providers of information in a variety of application contexts.

The act of personalization can yield a variety of results. In a simple case, it can bring a region of interest to a user (e.g., a listing of football scores from last night's games) either by passive means (i.e., filtering) or active means. In a more complex scenario it might create a coherent composition of information that can be delivered to the user (e.g., a five minute synopsis of the nightly news).

Historically, most work on personalized content delivery has been in the text domain. Personalization for text-based information is commonly achieved through the use of keywords or text vector space analysis to compare a set of documents with a set of user profiles. Audio and video data have characteristics without parallels in text, most significantly in the way video sequences can be combined to create new meaning and by the ability of these media to directly represent real-world objects.

Consider a scenario for news video delivery consisting of an archive of thousands of hours of news broadcasts (e.g., the archive at Vanderbilt University). If suitably annotated (i.e., metadata have been collected identifying the contents of the scenes and their location), then it is feasible to use existing personalization techniques of the text domain to deliver personalized video-based news. Content can be indexed, segmented, and ultimately retrieved in a recorded sequence based on a viewer's needs. Fig. 1 illustrates such a scenario in which the viewer initiates the composition of a variety of news items into a video stream.

Supporting such scenarios requires the ability to create indices, match user characteristics with content, and locate video objects (news items). A similar scenario exists for the delivery of instructional video based on student needs.

Creation of video objects can be achieved by extracting video segments from live broadcasts, or archives of live broadcasts, or can be created and edited specifically for this format. In the former, tools are necessary to facilitate rapid conversion from live broadcasts to recorded and indexed topics, and the linking of related static materials (e.g., references to sources or text-based information). These same tools must also allow the elimination of

out-takes, or other errors that the editor deems unacceptable.

Once the video data are indexed, there is a data management and access problem. For example, if the news items of Fig. 1 consist of complex multimedia objects, satisfying a user request requires location of components, assembly, and timely delivery. Because video data are best served by a storage system supporting continuous media, it is desirable to separate the data searching and location functions from the raw storage functions. For these reasons a distinct metadata management scheme is appropriate for multimedia content. Appropriate search tools and engines are also needed and they can be conventional text-based engines. More sophisticated image-based or motion-based tools for content searching can be used, but their utility in this application domain is not clear at this time.

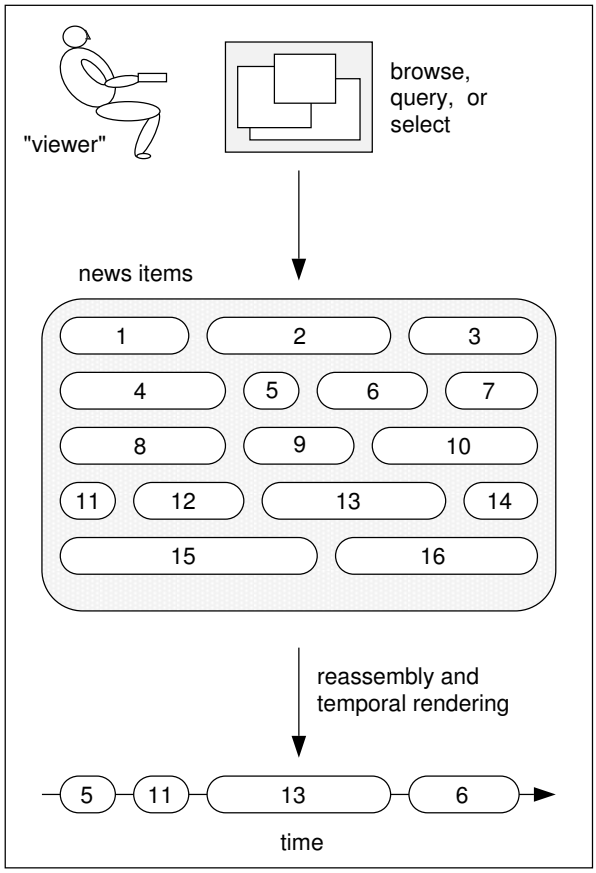


Figure 1: Explicit Selection and Composition of News Items

In the remainder of this chapter we consider personalization of video content for both filtering and composition. In Section 2 we consider existing text-based personalization schemes. Section 3 explores the unique characteristics of audio and video data. In Section 4 we dis-

cuss the metadata required to support personalization. Section 5 presents a video annotation tool for collecting video metadata necessary for personalization. In Section 6 we propose a personalized video environment based on the news domain. Section 7 concludes the paper.

# 2 Personalization of Content Delivery

Personalization deals with the tailoring of an application instance to the needs of individual users. For content delivery purposes, this personalization can affect the selection, scheduling and presentation of a set of documents. In this section we focus on how a set of user preferences can be applied to select and sort a collection of documents.

Typically, the current preferences can be expected to correlate with past and present user contexts. Context includes factors such as user knowledge, the tasks being performed, the problems being solved, and the user's system resources [20]. We define personalization as the process of adapting the selection, sequential sorting and presentation of the set of available documents to the user context. The aim of the adaption is to let the user complete a task in the shortest amount of time using the least amount of resources.

This selection, sorting, and presentation is represented as a sequence of operations performed on the combined attribute space of users and document profiles. A set of parameters is associated with each operation. Fig. 2 illustrates an example where metadata describing the user context combined with metadata describing the documents are used to generate a sorted view of the information space.

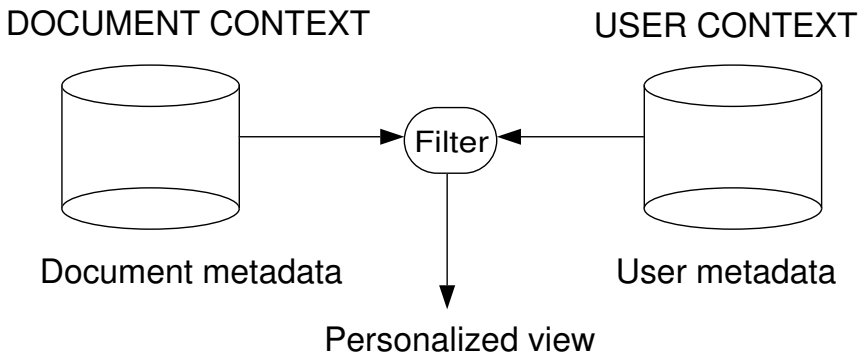


Figure 2: Personalization as the Filtering of Documents based on the Current User and Document Context Using Both Document and User Metadata

Traditional information retrieval techniques aim to find data in a large collection of

documents through a sequence of independent queries and query results. Information filtering can be defined as a reverse process of information retrieval where queries are stable and specific, reflecting long-term user interests.

A document now takes the role of a traditional query, while the stable collection of queries can be seen as documents representing user profiles. The time persistence of user profiles makes adaptation towards true user preferences possible via *learning*. Learning can be defined as updating the queries describing estimated user needs by minimizing the differences between predicted and observed user preferences. The user interaction with a set of documents must be mapped to the user query based on explicit and/or implicit preferences extracted from this interaction.

User profiles or metadata must be represented in a way that allows a proper mapping between the user and a universe of documents. The two main classes of user models are *canonical* and *descriptive*. The canonical model requires a formal encoding of a cognitive user model as in the BGP-MS user modeling system [12]. The models are hard to acquire and their complexity hides the represented semantics from the user [20]. Descriptive user models can be automatically created by observing user behavior. Their content is a mapping from previous document accesses and does not require any semantic processing. A large number of observations is needed to be able to draw high quality conclusions.

Recently, *agents* have been used to implement information filtering functionality along with other aspects related to the collection, selection and presentation of documents [17, 29, 30]. An agent is an autonomous program that has a set of goals it tries to fulfill in a given dynamic environment. The agents can operate individually or cooperate by sharing knowledge and work [15]. One important aspect of agents is their ability to travel across networks.

## 2.1 Information Filtering

Present strategies in information filtering are classified as *cognitive*, *social*, and *economic* [18, 30]. We also include *pattern based* filtering techniques, generalized from work in the field of *software agents* and adaptive hypermedia. We will discuss each of these approaches in the remainder of this section and how they can relate to the personalization of video delivery.

**Cognitive Filtering** Cognitive filtering is a technique in which the description of a document is matched against a user profile where descriptions relate to static autonomous properties. The document profile consists of descriptors associated with the actual contents (e.g., concept keywords, topics, or structural annotations) and with content creation (e.g., document author, creation date, or creation location). The user profile typically follows the structure of the document profile, but can contain other parameters describing the user.

Cognitive filtering draws on traditional textual information retrieval techniques that can be divided into *statistical*, *semantic* and *contextual/structural* [19]. While semantic and contextual/structural techniques try to extract meaning through natural language processing, the relatively simplistic statistical approach has been the most popular. In particular, the vector space model by Salton [27] introduces a *document vector* to represent documents and queries. The document vectors span an $n$-dimensional space consisting of a set of keywords or concepts extracted from the document set. There exist various methods to scale each component as well as to compute the similarity between two vectors. One method of comparison is simply to compute the vector angle. A smaller angle indicates higher similarity.

We briefly summarize Salton's ideas:

The document vector can simply be the terms as they appear in the text itself, or can be processed through one or more of the following three steps:

- Words that appear in a designated list of stop words that carry little or no meaning are removed from the text.
- A stemming function reduces words to their stems by using a list of common suffixes and prefixes or general stemming rules.
- A dictionary of synonyms maps each word stem to a concept class term reducing the dimensions of the space while improving the vector quality.

For a binary vector representation $W$, the presence of a term $i$ gives a binary weight $w_i = 1$; the absence gives $w_i = 0$. Salton's more expensive weighted document vector representation is described in the following. The occurrence of each term $i$ is counted to form a term frequency $TF_i$. By doing lookups in a larger collection of $N$ documents, the document frequency, $DF_i$, is found for each term $i$ (i.e., the number of documents in which the term appears). We can now calculate the inverse document frequency, $IDF_i$, often defined as:

$$IDF_i = log_2(\frac{N}{DF_i}). \tag{1}$$

Each vector component can be calculated as:

$$w_i = \frac{TF_i \times IDF_i}{\sqrt{\sum_{j=0}^{N} TF_j^2 \times IDF_j^2}} \tag{2}$$

This scheme weights rare terms more heavily, while frequent terms are weighted near zero. Finally, each vector is normalized to length 1 to make comparison scores uniform across documents.

To improve retrieval, content creation metadata about a document is often available. For example, a document represented in SGML can have portions tagged to describe titles, authors, structural labels (sections headings), abstracts, etc. On-line news articles frequently possess such fields. These fields can have variable importance when filtering documents for different users [16, 30]. Here, correlation of past documents that were chosen by the user has been used to assign weights to each field. The weights are used to scale the scoring of similarity between the user profile and each of the document fields. The user metadata contain corresponding sections to organize the user preferences as illustrated in Fig. 3. In this example, the news article *source* field is demonstrated to be more than four times more significant than the *location* field of the reported story.

**Social filtering** Cognitive filtering requires semantic descriptors to be attached to documents. Except for interpretation of closed-captioned text, cut-detection and extraction of basic camera operations like zoom, pan and tilt [1], automatic extraction of descriptors is difficult for video data. Even when automatic extraction is possible, cognitive filtering does not take assessment of quality, timeliness, composition or correctness into account when scoring documents [28]. To overcome these limitations and personalize delivery even when content descriptors are erroneous or missing, *collaborative* or *social filtering* has been applied in various formats [16, 18, 24, 26, 28].

Social filtering is based on the aggregate filtering of documents by a community of users. Documents are recommended to the user based on previous accesses by the community. For each access, the system estimates how a user liked the document, either implicitly or by letting the users explicitly rate the document. Based on the similarity measure between a user and individuals or subsets of the community, previous ratings by the community of a

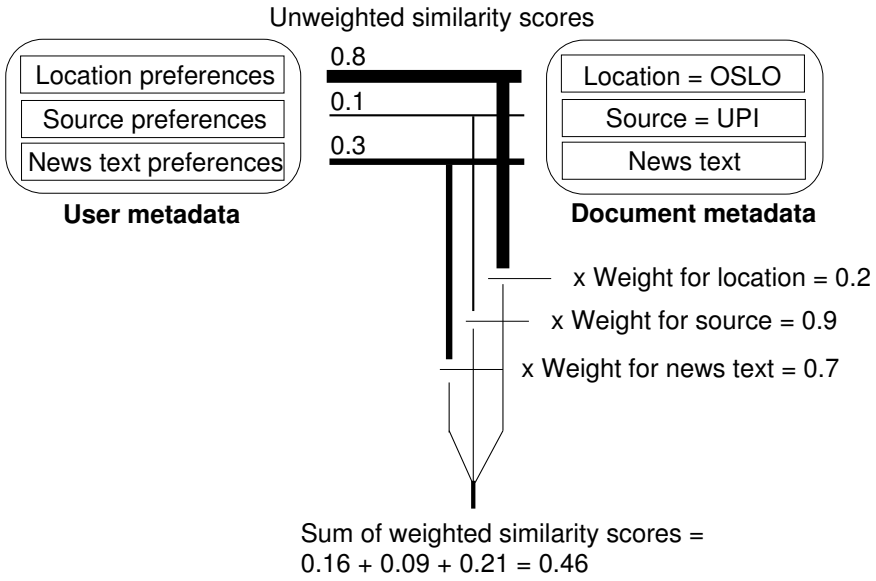


Figure 3: Different Fields of a News Message May Have Different Importance to Different Users

given document are mapped to a prediction of how the user will rate the same document. The Tapestry system [9] requires explicit specification of similarities by the user. The GroupLens system [26] uses previous ratings of USENET news articles to form similarity values between users of the system. GroupLens requires users to pass a threshold of similarity before using their ratings as advice for each other.

A problem with social filtering is the need for a large community. A document in general must be exposed to a high number of users before reliable advice can be given. This suggests that cognitive or economic filtering techniques be used in conjunction or that other methods ensure unfiltered access to documents until a level of confidence in the system prediction has been reached.

We expect social filtering for video to prove an interesting research area as the infrastructure for interactive video delivery improves.

**Economic filtering** Economic filtering techniques base their selection and ordering on the costs and benefits associated with production and consumption of a document [18]. This type of filtering can be applied to both the document provider and receiver. In the video arena, the content provider wants to sell a video "stream" for the highest price while minimizing costs (e.g., bandwidth), while the receiver would like to minimize the price and maximize

the quality.

For this video delivery perspective, cost factors can include the cost of the intellectual property (movie licensing), bandwidth, transmission time, image resolution, required screen size, required buffer space, and playback time including client processing time.

A video news story can be priced according to its age from the time of creation. An economic filtering approach might attempt to find the most recent story with the lowest price. Using additional cognitive filtering techniques, the user's willingness to pay can be made a function of expected interest in the news topic.

**Pattern-Based Filtering** Pattern-based filtering can be defined as the selection and sorting of documents based on non-cognitive analysis of previous access patterns.

Supposing that merely updating a set of user preferences is insufficient, work by Cypher demonstrates how repeated patterns of document accesses can be generalized [5]. These generalized patterns can be used by the system to present documents during subsequent accesses to the same documents. We view this as a filtering of the set of possible sequences.

As indicated by Orwant [24], sequences of user actions can be modeled as discrete Markov chains. For example, it is possible to use higher-order Markov models to capture World Wide Web access patterns and to group users based on similarities in this access using clustering.

For use in personalized content delivery systems, pattern-based filtering extends previous methods by considering the sequence of document access. In a hypermedia network or in a temporal medium such as video, the sequence of access is significant and pattern-based filtering can select and order a personalized sequence chosen from a set of valid sequences. In a hypermedia presentation, hyperlinks can be automatically generated based on previous access patterns. The pattern based approach presented by Yan et al. [31] also makes use of social filtering to let users with similar access patterns share information.

## 2.2 Content, Format, and User-Driven Presentation

Previous work in multivariant movies by Davenport and Murtaugh [6] and Evans [8] has focused on automated sequencing of story elements using mainly visual story telling techniques.

As the video granules are played back, objects in the video stream appear and disappear,

concepts are introduced, treated in more detail and replaced. By using metadata associated with each scene, we can model the current sequence context by mapping the metadata attributes to a movie context model [6]. We can use this movie context model with the composition conventions to generate a pure *movie driven* composition. Instead we could keep a static or slowly changing user context model when choosing granules, making it a fully *user driven* composition.

As previously shown, such a model requires the use of metadata for both the past and present user actions. While present session metadata can be constructed by implicit or explicit expression of user preferences during playback, the past session metadata can represent a summarized and processed version of a past session's user interaction. The present user metadata can consist of two parts: A user query to establish the overall objective and format of the presentation and mapping from a user's subsequent interactions as the sequence is constructed and played back.

*Format driven* composition is guided by a representation of cinematographic knowledge trying to preserve rules such as continuity and the use of an establishing shot when changing topics in the presentation. Appropriate metadata are required as input for the rules. Later we will identify classes of these metadata necessary to obtain different properties for personalization.

We believe that a system for presenting news video should be driven by a combination of the three composition strategies. User preferences, available content, and cinematographic rules form a set of constraints that in combination make it possible to automatically choose an satisfactory sequence (composition) at any given time. We call this personalization of temporal composition *sequential information filtering*.

## 2.3 Existing Systems

A number of systems for personalized delivery of information exist today. Most process information available via the Internet such as electronic mail, USENET articles, and World Wide Web content. A few commercial services offer access to more general news material from wire services and other commercial news providers. The only video delivery system that can be said to be partly personalized is the ConText system [6].

ConText demonstrates how cognitive annotations of video material can be used to individualize a viewing session by creating an entirely new version through context-driven

concatenation. This dynamic reconstruction can include video material made in a totally different context, thus performing a *repurposing* of the material. The system requires a uniform set of metadata which poses a challenge to metadata normalization when repurposing a distributed population of video material.

Evans defines a framework in which a limited set of metadata are assigned to a collection of shots using an application called *LogBoy* [8]. This video database supports a companion access application called *FilterGirl* that relying on a hierarchy of filters to generate queries to the database. The queries produce a sequence from a subset of the available shots in the database as output.

*MovieSelect* from Paramount Interactive Inc. uses a variation of social filtering also proposed by Shardanand and Maes [28]. By measuring similarity between documents instead of users, new documents are suggested by evaluating the previous ratings of other documents weighted by the estimated similarity between the rated and unrated document. MovieSelect suggest movies to users based on previous ratings but does not carry ratings to the next session. Other video-on-demand systems surveyed only consider delivery of entire linear movies, that is, any content-based selection yields large fragments of the original video.

By observing user actions while using a USENET news reader and a mail application, agent software developed by Maes gradually learns user preferences [16]. This work also introduces the concept of *trust* in agent decisions. The agents are reluctant to perform operations when the amount of past observations in similar situations is small. Each action taken by the user is associated with the current situation into so-called *situation-action pairs* and is used to find patterns. This system makes use of both cognitive and social filtering and is interesting for a wide range of applications since it present a generalized method of personalization using pattern-based filtering.

# 3 Characteristics of the Video Medium

Computer technology has finally evolved to a point where Nelson's browsable, vari-sequenced hyperfilm [23] is a practical reality. Until now, access to the video medium, even via computer, has been accomplished only using edited linear sequences. Computer assisted access to the video medium provides opportunities for at least four conceptually different methods of navigating video sequences [11] with potential for supporting personalization:

- Navigation in the representation of the original storage medium, (e.g., by moving back and forth along the original linear timeline or between various tracks).
- Navigation in the recording structure (e.g., by jumping from shot to shot or from scene to scene).
- Navigation in the three-dimensional reality represented in the recording (e.g., a virtual environment).
- Navigation in a *semantic space* derived from the video contents. The semantic space can be defined as the sum of meaning decoded from the visual and aural contents of the video stream.

Personalization of video delivery can adopt aspects of all of these methods. We focus on the semantic space. In the remainder of this section we describe the characteristics of the video medium that are relevant to video personalization. This includes an overview of basic techniques for a visual narration, a description of relevant metadata, and a synopsis of the television news format.

## 3.1 Structures of Video Narration

The language of film consists of conventions for spatial and temporal composition [21]. We choose to consider personalization a function of only the temporal composition and treat spatial composition as fixed. Since our focus is on techniques for the personalization of navigation in the semantic space, we want to look at how the information is conveyed to the user, (i.e., how the story is told). Existing story structures for providing nonlinear video access can be divided into the following categories [8]:

- **Hypermedia networks using hardwired links:** The links have simple rules based on previous access patterns.
- **World models:** As in the Oz project [2], viewers are immersed in a simulated world and become characters. The video presentation depends on how the user interacts with objects and other characters in this world.
- **Description-based structures:** These structures focus on filters that use content description and user preferences to create a customized version of the contents.

As pointed out by Evans [8] and Davenport [6], description-based structures are suitable for automated presentation of content without constant user interaction. This achieves a level of user immersion in the story that is lost by traditional browsing in hypermedia networks. Most interactive movies have a looser conversational user control in which the user chooses among several options at specific points in the storyline to create a personalized story. The ConText system described earlier [6] provides the user with the option to interfere with the selection of video sequences, making it a promising hybrid method where the user interacts only when dissatisfied with the current presentation.

## 3.2 Video Metadata

We classify video metadata as describing structure or content. Video structure includes media-specific attributes such as recording rate, compression format, and resolution; and cinematographic structure such as frames, shots, sequences, and the spatio-temporal characterization of represented objects.

Video content metadata deal with the remaining universe of semantic information. We further decompose this universe into the set of tangible objects, and the set of conceptual entities including events, actions, abstract objects, and concepts appearing in or resulting from the media stream. This classification of video content metadata is not intended to yield disjoint sets.

In the following, we concentrate on metadata for news video; however, these aspects can be generalized to other video domains.

**Structure Metadata** A personalized presentation of video will typically range from small modifications on a composed linear sequence through complete re-composition of a set of individual linear units as illustrated in Fig. 4.

Video structure includes media-specific attributes such as recording rate, compression format, and resolution; and cinematographic structure such as frames, shots, sequences, and the spatio-temporal characterization of represented objects. In summary, video structural metadata:

- **Media-specific metadata:** Describe implementation-specific information (e.g., video compression format, playout rate, resolution).

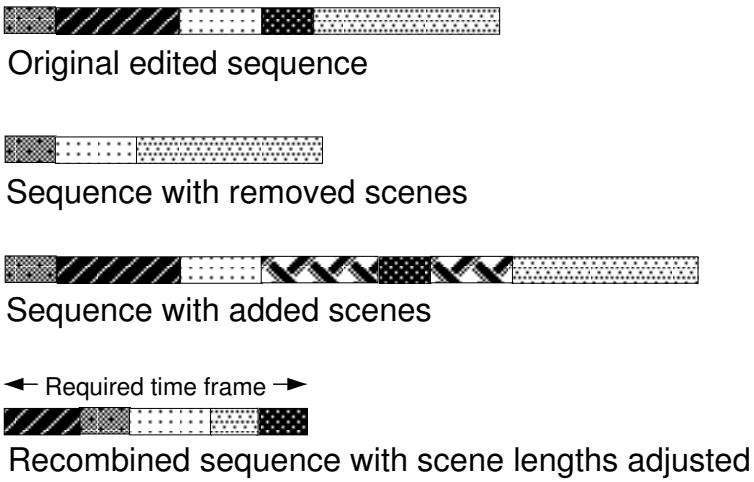


Figure 4: Personalization Ranging from Simple Filtering of a Linear Sequence to Recomposition Using One or More Shots

- **Cinematographic structure metadata:** Describe creation-specific information (e.g., title, date recorded, video format, camera motion, lighting conditions, weather; shots, scenes, sequences; object spatio-temporal information).

Structural annotations represent linear video sequences as a hierarchy of frames [7]. Frames recorded continuously in time are called *shots* and represent the smallest structural unit. A set of shots presented continuously along a time line combine to form a higher level structural granule called a *scene*. *Sequences* are again made up of scenes. Thus, hierarchy of frames, shots, scenes, and sequences constitutes the simplest video structural model.

**Content Metadata** Video content metadata are concerned with objects and meaning in the video stream that appear within or across structural elements. Content metadata are further decomposed as:

- **Tangible objects:** Describe objects that are appear as physical entities in the media stream (e.g., a dog, a disc).
- **Conceptual entities:** Describe events, actions, abstract objects, context, and concepts appearing in or resulting from the media stream (e.g., running, catching, tired, master).

Both content and structure metadata are required for personalization. Structure metadata (e.g., shots) can sometimes be automatically extracted from raw video data. Content metadata, necessary for personalization usually must be obtained manually. Closed captioning text contains both structural (time offsets) and content (text) information.

## 3.3 Characteristics of News Video

Most television news is composed of aural story telling techniques delivered by an anchor person, reporter, or people being interviewed. The common news story formats are differentiated by how the anchor or reporter interacts with the visual and aural based footage. Often a television news item consists of an introduction by the studio anchor followed by field footage or images and graphics illustrating the story. A reporter usually mediates the material either on location or as voice-over. The anchor person contributes to the credibility of the presentation. In a personalized multi-source presentation an anchor person can have an even more important role, but instead of hosting a broadcast, the anchor would host individual segments.

We can classify news stories by the following five distinct story styles [22]:

- **Spot news or actuality:** A breaking news story that is covered live or quickly after it has happened. Scenes follow each other in a linear fashion and voice over from the studio and live sound including commentaries from the scene are used.
- **Stand-upper:** This second most common format is prepared by a reporter gathering information and footage recorded to match the story. The reporter is reading the story into the camera. Often, the reporter is seen at the end of the story for the last lines including the sign-off.
- **Wraparound:** The wraparound consists of two parts, the opening and closing sequences and the middle sequence. Typically, the opening and closing might be done by the anchor and the reporter will do a stand-upper middle part intermixed with interviews or other on-location footage.
- **Voice-over:** A less important story is made by the anchor or reporter reading the whole story while footage matching the text is shown either as motion pictures or as stills.

- **Interview:** The interview is either taken right on the site of the story or pre-arranged to find optimal backgrounds and camera positions. It might be part of a stand-upper or wraparound news story.

In addition to these styles, Musburger [22] identifies two more extensive categories of *feature story* and *sports story*.

# 4 Metadata for System Resource Management

Ultimately the task of delivering personalized information to the user is the burden of the host computer, storage, and network distribution system in cooperation with client computers. Collectively, this system must achieve the goals of personalization. Because personalization introduces diversification (e.g., narrowcasting) rather than generalization (e.g., broadcasting) , it increases the burden on all aspects of the system's resources. In this section we describe the tradeoffs in personalization versus generalization and the metadata required for both.

Personalization, from a system's viewpoint can be approached from two perspectives:

- From a system performance standpoint, where the aggregate behavior of a user population is used to predict usage patterns and optimize resource usage.
- From a user performance viewpoint, where the attributes specific to a user's preference (e.g., content preferences, interface requirements, reneging behavior, connectivity, cost) are used to customize the presentation of information to a user.

Unfortunately, these objectives are orthogonal to one another. Aggregating users (e.g., broadcasting) reduces the ability to personalize information delivery while personalization increases the cost of providing the service. It is clear that a balance must be achieved for optimal system performance and user acceptance.

Aggregated user behavior can be used by the system in prefetching of data on a client or server-initiated basis. In client initiated personalization, metadata that are processed remotely by the server are made available to the client which subsequently decides whether or not to retrieve the information. In server-initiated personalization, the metadata are processed at the server which then aggressively uses this information to push the data to its client population.

Metadata are commonly cited for use in indexing data to simplify the search process that a user must undertake in extracting relevant information from a data set. As a result, the metadata typically contain key attributes extracted from the data and organized in a form optimized to minimize database search times. This view is consistent with traditional text-only databases where searches typically involve searching for keywords in a document set.

For these systems, there is little need to embed additional information about the "documents" into their metadata. Accesses to these databases are characterized by repetitive requests for a small subset of the entire data space and traditional cache management techniques suffice to improve the performance of the system.

Multimedia databases change this requirement. Multimedia data types such as audio and video also have the additional attributes of playout times and required bandwidth associated with them. It thus becomes necessary for the system to be aware of the timing and bandwidth characterizations of a multimedia object when requested, so that appropriate resources can be reserved to ensure the timely delivery of data. The use of metadata to support database browsing or personalization functions is significant due to the potential bandwidth used in video delivery. By using a metadata scheme, the data delivery process can be decoupled from the database management functions.

Personalization of information delivery adds a new dimension to this perspective. Data must now not only be characterized by their physical attributes, but in addition, a correlation between the data and their users must be established. Once such a relationship has been established it must be used effectively to enhance the performance of the system at a maximum benefit to the end-user. We therefore see that the previous mentioned constraints must also include system performance constraints when finding an optimal playout sequence.

The following scenario illustrates the use of metadata in resource management for the system:

- A user initiates a session. The user is presented a menu based on predicted user preferences. As the user navigates through the sessions, user preferences are mapped onto existing usage patterns to tailor information presented (aggregate metadata plus user metadata).
- The metadata about the actual data are combined with the user preferences to ensure that appropriate resources are reserved ahead of time (e.g., bandwidth for live video

playback).

Tailoring individual delivery to a large number of users is extremely expensive without some form of simplification or aggregation of behaviors. In the extreme case, all information about an individual can be recorded and maintained by the system. For each user a database and database analysis would be required to understand each user's behavior and personalization requirements. Such analysis, if based on text vector-space techniques is computationally expensive with current processing and storage technologies. Therefore, we seek techniques that simplify the data set characterizing individuals and permit a simpler representation of the same information.

**User Clustering** One method of behavior aggregation is by clustering of users and assigning user profiles. Clustering can be achieved by observing interface usage behavioral patterns [24, 25] or by analyzing accesses to the set of documents previously visited by the user. The result is a set of profiles characterizing the user population, but pre-presented by a subset of the original data describing the great detail about individual users.

Of interest here as with any clustering technique is the effectiveness in reducing the amount of information and information processing required to achieve the personalization goals. To this end, we introduce the concept of *preference entropy* to measure the relative success of a set of parameters in characterizing a personal profile. Essentially we can quantify the relationships among the metadata categories, the user behavior, and the deviation from the ideal case with this concept.

A preference entropy close to zero indicates that the user has no particular preference among the options, while a large entropy indicates that the preferences are non-uniform.

Once we have established a need for a particular metadata component for a single user, it must be considered with respect to the user population. The individual's set of option preferences might coincide with the average option preferences in a user population, it might coincide with the average option preferences in a subset of the population or it might be clearly distinct from any other user. To decide whether the population can be divided into clusters of users with similar preferences we can use the simple technique of finding the covariance of the difference between a user's option preferences and the average preferences in the population at large. Or we can apply more advanced calculations on the set of users for the same goal.

By only allowing a subset of the available parameters to be personalized for any given user or for a given user cluster, we can reduce computation while minimizing loss of adaptation by choosing parameters with minimum entropy.

# 5 Collection of Metadata for Video: Vane

As one might imagine, the ability to provide personalization on video metadata depends on the quality and quantity of information extractable from the video data. Here, we overview a system designed to capture metadata from video content to support video personalization.

## 5.1 An Overview of Vane

Vane is a video annotation tool developed at the Multimedia Communications Laboratory at Boston University [4]. Video annotation with the tool involves automatic segmentation of video data followed by manual content annotation (Fig. 5). The tool automatically detects camera breaks and displays the detected shots along a timeline (structural metadata). The information model is defined by SGML document type definitions (DTD) for the desired video domain. After annotations (structural and content metadata) have been collected as SGML files, they can be converted to any database format suitable for personalization functions (we currently use a relational DBMS). Specifics of the Vane tool's construction for metadata collection are described next.

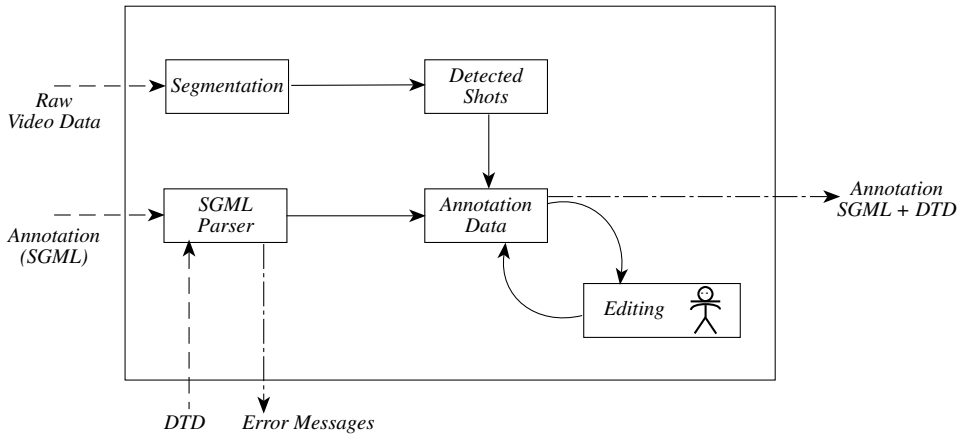


Figure 5: Data Flow for Vane

## 5.2 The DTD and the Dynamic Annotation Interface

Vane, at load-time, reads in the domain DTD and brings up appropriate fields to record metadata. Metadata are stored in SGML files according to the format specified in the DTD. If an existing annotation file is loaded into Vane, it first checks the file format for errors with respect to the DTD. One of the advantages of having a information model defined by a DTD is its adaptability. One need only change the DTD rather than the tool to accommodate a new video domain. With a new DTD, the interface gets automatically updated.

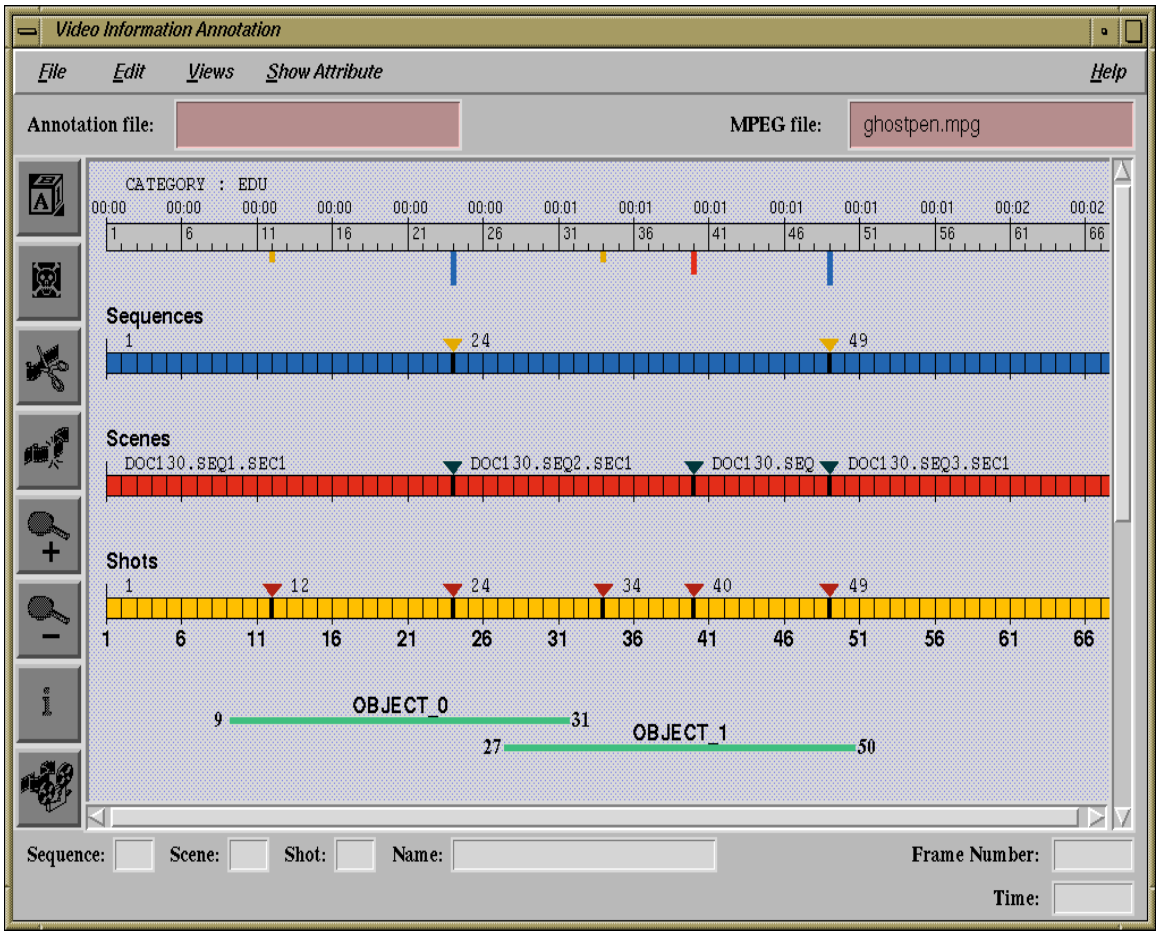


Figure 6: Vane: Video Annotation Engine

Scenes and sequences are identified manually. Metadata about each shot, scene, sequence, and complete document are captured in the forms indicated by the DTD. Metadata are stored at various levels of granularity, e.g., keywords, headlines, abstract, transcripts, etc. to cater to different requirements of the application domain. Links are provided within text (abstracts and transcript), linking text to images, text, or video segments. Shots, scenes,

and sequences can also be linked to images, text, and video segments, creating information paths for user traversal and information presentation.

The ability of SGML to nest several elements inside one another allows Vane to easily define a structured view of the video content. For example, a shot description can be nested inside a scene description. We applied these concepts in the creation of a baseline DTD with the following syntax:

```
<!ELEMENT FULLDOC  - - (ABSTRACT?, CATEGORY?, REF*, SEQUENCE*, OBJECT*)>
<!ELEMENT SEQUENCE - - (ABSTRACT?, REF*, SCENE*)                        >
<!ELEMENT SCENE    - - (ABSTRACT?, REF*, SHOT*)                         >
<!ELEMENT SHOT     - - (ABSTRACT?, REF*, TRANSCR?)                      >
<!ELEMENT OBJECT   - -   (OBJECT*)                                      >
<!ELEMENT ABSTRACT - - (#PCDATA & REF*)*                                >
<!ELEMENT TRANSCR  - - (#PCDATA)                                        >
<!ELEMENT REF      - -  O EMPTY                                         >
<!ELEMENT CATEGORY - - (EDU | NEWS | MOVIE | DOC | SPORT)               >
```

The salient features of this description are the ability to characterize video structure and content attributes as metadata, the ability to extent the DTD (and interface) at run-time, nesting of concepts and objects., and the automatic generation of cross-references. Details of the tool and its construction can be found elsewhere [4].

## 5.3 Translation to Alternative Representations

Once the metadata are collected, it is desirable to analyze, condense and reformat collected metadata into a representation suitable for serving the user population. This includes generation of indices to information, cross-references, and organization of data for fast access and delivery.

In Fig. 7 we provide an example of a metadata representation constructed for simple access to news video. Here a single news broadcast is considered as a document. The structure of this news document is comprised of elements from different categories (sports, politics) where each category consists of multiple news items. We collect cinematic metadata such as the source, time, date, and location for a complete document. We can then extract the content information from each news item such as keywords, transcript, representative still image, and objects. Objects can belong to different categories such as people, location, field footage, shots, events, texture, color, etc. An object can be composed of multiple

objects. We can also store metadata about the associations between the objects, (e.g., a dog playing with a disc).

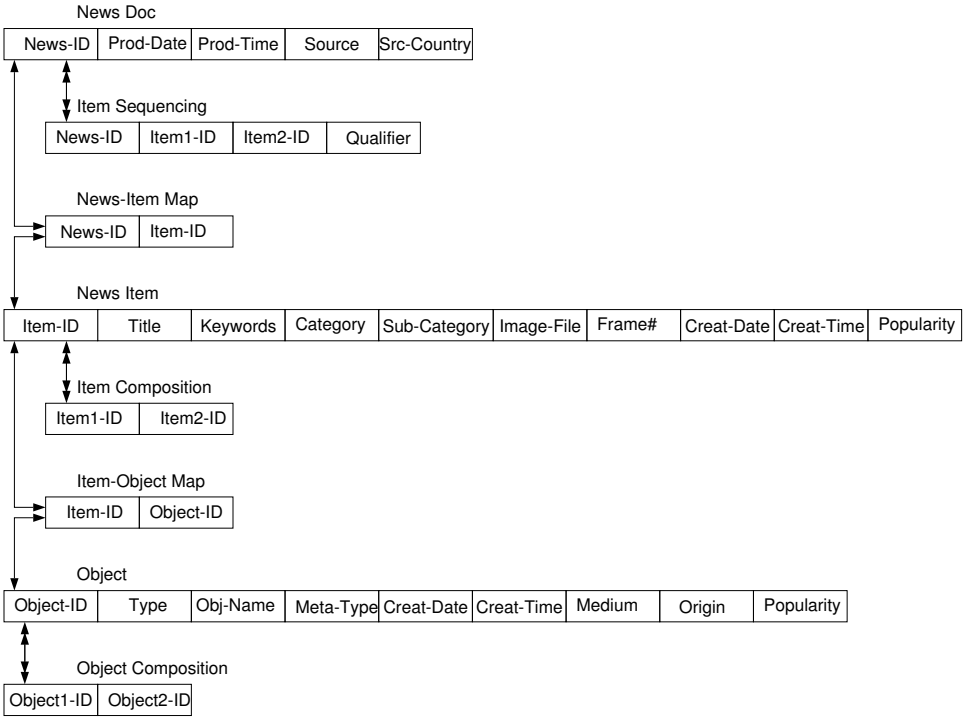


Figure 7: Newscast Schema

In this example a relational database model is used; however, any suitable representation can be applied.

# 6 A Framework For Composing Personalized Video

In the previous sections, we investigated a variety of techniques for filtering, collecting, and describing metadata for personalized video delivery. In this section we focus on a scenario in the news video domain and present techniques for composing news "documents" with sequential elements, as illustrated in Fig. 1. Here our objective is to achieve a methodology for creating cohesive, time-constrained, and personalized video delivery when content originates from many sources while preserving a format close to present television news stories. The characteristics of our personalized news delivery service are:

- A composition of multiple sources make up each news item and priority is given to

sections deemed relevant to both the flow of the story and the user.

- Duplication from multiple sources producing almost equivalent accounts of news events must be suppressed.
- The duration and order of each news story sequence reflects its estimated interest to the user. High interest items should be played out early and be given more time than less interesting items.
- The presentation accepts, but does not rely on interactive participation by the user. Hyperlinks providing more information on individual news stories are one way of giving control to the user.
- Learning user preferences for content, structure and presentation is achieved through implicit and explicit feedback.
- Notification of major news events is provided.

The remainder of this section describes an approach to achieving this vision.

## 6.1 Metadata for a Personalized News Service

To be able to filter the news segments using user preferences, we have to require identification of attributes that must be present in the news segment and user metadata.

**News Segment Metadata** Table 6.1 shows the news segment metadata used in our trials. For the personalization scheme to function, we require acquisition of structural and content metadata describing the news segments. In our trials, this is captured by reading the close-captioned text or by the use of Vane. A more suitable approach of obtaining this metadata is directly from the news providers.

**User Metadata** Since we use cognitive filtering, our user profile corresponds closely to metadata available for each news segment (i.e., user metadata are nearly identical to content metadata). We thereby choose a descriptive user profile constructed by observing user manipulations of previously constructed news story sequences.

Table 1: Metadata Fields for User Profiles with Sample Data from an Evening News Segment

| **Field** | **Description** | **Sample Entry** | **Vector** |
|---|---|---|---|
| .time | Time of broadcast | 96062518 66.38 | |
| .source | News Source | CBS | B |
| .semantics | Audio, video or audio-video | A | |
| .tracks | Audio, video or audio-video | AV | |
| .origtime | Time of content recording | n/a | |
| .question | Commentator question | | D |
| .duration | Duration of recording | 32.40 | |
| .structure | Structure of composition | Anchor | |
| .synchronous | True or false (audio) | True | |
| .concept country | Country under consideration | Saudi Arabia | B |
| .concept state | State under consideration | | B |
| .concept city | City under consideration | Dhahran | B |
| .country | Country of origin | USA | B |
| .state | State of origin | DC | B |
| .city | City of origin | Washington | B |
| .building | Building of origin | White House | B |
| .concept person | Person under consideration | Clinton, Bill | B |
| .freetext | Free-Form text | In Saudi Arabia... | D |

Not all fields represented in Table 6.1 are interesting for representing long-term user preferences. Time of recording and fields specifying what track is the most important for the segment semantics are parameters anticipated to be of little value for expressing long-term user preferences.

**Mapping Content to User Metadata** Using Salton's techniques [27], mapping content metadata to user metadata can be achieved with field vectors. The user metadata can be stored in vector format and the news segment metadata must be converted to vector format, either at the news provider or locally on the client. Since some parameters needed for calculating vector entries are only available at the provider side, we suggest that the news provider construct the vectors.

By using binary vectors for the subset of fields (labeled 'B' in the table) and weighted

vectors for another subset of fields (labeled 'D' in the table), we now have a representation that leads to simple relevance comparison between segments or between segments and user profiles. The vectors span spaces made up of all terms encountered in the available collection of similar fields.

To find the anticipated user preference for a new news story, we calculate the similarity between the user profile and the corresponding metadata fields associated with the news story, both represented as vectors. The similarity measure can be estimated by computing the vector angle. We expect varying user interest in matches between different field pairs as well as varying weight distributions among news subjects. The distribution is likely to be a function of both the field values and the field type; however, we currently pursue the one-profile-per-user simplification.

## 6.2 Composing News Video On-the-Fly

To generate a personalized news presentation, a typical client system would first process all metadata for a given day's news from a number of sources. Queries to find additional material could also be issued at this time. The material necessary for presentation would be down loaded from the various sources, stored in a cache and eventually played back to the user. Personalized presentations could be provided on-demand, or scheduled for particular times to meet system resource constraints. They could also be triggered by the availability of user-defined material or by breaking news exceeding a certain interest threshold defined by the user.

In the remainder of this section, we will concentrate on the composition of a single news story sequence from a number of sources.

**News Segments** News is a narrative with a combination of voice and images. News is rarely conveyed with images only, but often is represented with voice and a picture of the speaker. Sometimes it is rendered with the speaker plus supporting images and original sound is of little importance. Here the speaker interprets the images and forms a view. Thus, current newscasting norms indicate correlation of spoken words and images.

News stories are comprised of fragmented pieces of events, commentaries and interviews that are well suited for automated concatenation. Today, the editor and narrator ensure that a news item is coherent. In our personalized system want this to be automated.

To personalize the presentation of an individual's news story, each segment of audio and video must be given a relevance score as a *member* of a set of related content and as a *part* of a particular composition.

We define a *news segment* as our basic structural unit. It is a sequence of related shots that are available from a given provider. The segment is expected to have a meaningful linear structure and it may include shots of varying lengths and quality. Available news segments about a particular news story are expected to be ordered in a meaningful way even if this is not essential for our purposes.

Attributes within a segment can be relevant to the beginning, the middle, the end, or any combination. The sample entries of Table 6.1 give an example of a news segment and associated attributes about a truck bombing in Saudi Arabia from a CBS newscast.

To simplify the discussion, a valid sequence of news segments consists of a path of non-overlapping segments. To further refine the presentation, we want to fill-in missing audio or video with data from other segments, but we defer this task. We also note that many news segments suffer from lack of moving images due to the absence of live or recent feeds. We therefore greatly improve the quality of the composition by allowing alternative video segments to augment concatenations in which audio is the main carrier of the storyline.

We expect metadata for each news segment to be available for retrieval, independent from the access to the aural and visual data (e.g., as collected by the Vane tool). The metadata for a segment must indicate whether the audio, the video or both audio and video are necessary to convey the content. (It becomes important to distinguish audio and video as separate media.)

**Inter- and Intra-story Clustering of News Segments** Our initial task to group related news segments into sets of material addressing a particular news story is pre-supposed. While related segments from the same source will carry a common thread identifier, we cannot expect a uniform identifier across all sources. Simple clustering methods like the Leader algorithm can be used [3, 10] to divide the segments into dissimilar groups.

The semantics of a segment group are represented by creating a *super segment vector* set. This super vector set is a set of vector sums created by adding the vectors in each metadata field and normalizing them. Our observation of closed caption data for news stories indicates that relevance decreases with time. Using clustering techniques, one can also find clusters within each news story cluster. In this way, duplicate news items are grouped to optimize

for spread when picking and scheduling representative segments later.

**Assigning Durations to News Story Clusters** After the clustering step, we expect all available news story material to belong to distinct news story groups. Comparing all super segments with the relevant user profile, we end up with a set of normalized relevance scores for each cluster.

Unlike traditional approaches, we also include the production time and amount of available material (cluster size) in the relevance judgment. This facilitates playout time constraints (e.g., a news synopsis in 10 minutes). Borrowing from concepts used for automated layout in the Krakatoa Chronicle [13], we can use *density* as a measure for how many news stories a user wants to see per unit of relevance, and *sensitivity*, as a measure for how much the news item duration is allowed to vary as a function of its perceived interest. A high sensitivity would let a news item capture almost the entire time of the personalized newscast if it had a much better match with the user profile than any other item.

**Scheduling a Sequential News Presentation** To perform set and sequential filtering on a particular cluster of news segments we use a set of filters that each are associated with a weight that is a function of playout time within the news story. This approach is an application of Evans' ideas but with support for a time-dependent filter hierarchy. We call this *timed-based filtering*. The following filters would be applied:

- **Relevance to user profile:** This filter generates the relevance between a segment and the appropriate user profile.
- **Relevance to center of story:** This filter generates the relevance between a news segment vector set and the super segment vector set for the segment cluster.
- **Relevance to the introduction, middle, and end of the story:** These filters are based on the values in the *structure* field that let news providers define where the segment will be appropriate (e.g., introduction, background, or coverage of the current event). The middle of the story might be about the current event while the end of the story might contain background coverage.
- **Timeliness:** This filter returns a normalized value that reflects the relative age of an input news segment.

- **Continuity:** This filter ensures continuity. One possible continuity filter uses a division of the freetext field into two equal parts, each represented by a document vector. This filter returns the expected normalized relevance between the second part of the present segment and the first part of the next segment in the composition.

- **Spread of representative segments:** This filter prevents related segments originating from the same cluster from appearing in the same composition. Clusters contributing segments to a composition will be given lower scores than unvisited clusters.

We propose to make these filters operate in parallel except for the repetition filter that makes sure no segment is scheduled for playback twice. Scheduling of segments will continue until the allowed time slot for the news story has been filled and scheduling of the next news story will begin. Thus, the operation of the filters on the metadata from the segments and from the user yields personalized delivery of video content.

# 7 Summary and Conclusion

The existence of solutions for information personalization using text-based methods affirms the viability of creating personalized video delivery using metadata. In this chapter we have described techniques to facilitate this personalization, overviewed the unique characteristics of the video medium, and proposed a framework for personalized delivery of video information in the news domain.

Although unrealized, our proposal is based upon the application and integration of existing techniques of video content and structure modeling, metadata collection, vector space analysis, personalization and filtering, metadata management, and video and audio composition. The core of the concept is the use of a set of vector-based and time-dependent filters for audio and video segment selection to generate formatted video-based compositions on-the-fly. We focused primarily on the news domain, but we expect the concepts to be appropriate for other areas of information composition using motion video.

# References

[1] G. Ahanger, D. Benson, and T.D.C. Little, "Video Query Formulation," *Proc. IS&T/SPIE Conf. on Storage and Retrieval for Image and Video Databases*, Vol. 2420,

February 1995, pp. 280-291.

[2] J. Bates, "The Nature of Character in Interactive Worlds and The Oz Project," in *Virtual Realities: Anthology of Industry and Culture*, C.E. Loeffler (ed.), 1993.

[3] J. Bezdek and S. Pal, *Fuzzy Models for Pattern Recognition*, IEEE Press, 1992.

[4] M. Carrer, L. Ligresti, G. Ahanger, and T.D.C. Little, "An Annotation Engine for Supporting Video Database Population," ta appear in the *Journal of Multimedia Tools and Applications*, 1997.

[5] A. Cypher (ed.), *Watch What I Do: Programming by Demonstration*, MIT Press, Cambridge MA, 1993.

[6] G. Davenport and M. Murtaugh, "ConText Towards the Evolving documentary," *Proc. ACM Multimedia '95*, San Francisco, November 1995.

[7] G. Davenport, T.G. Aguierre Smith, and N. Pincever, "Cinematic Primitives for Multimedia," *IEEE Computer Graphics & Applications*, July 1991, pp. 67-74.

[8] R. Evans, "Log Boy Meets Filter Girl: A Tool Kit for Personalizable Movies," *MS Thesis*, MIT, Cambridge MA, 1994.

[9] D. Goldberg, D. Nicholas, B.M. Oki and D. Terry, "Using Collaborative Filtering To Weave an Information Tapestry," *Communications of the ACM*, Vol. 35, No. 12, December 1992, pp. 61-70.

[10] J. Hartigan, *Clustering Algorithms,* John Wiley, 1975.

[11] W. Klippgen, "Navigation in Digital Video Archives," (in Norwegian), *Diploma Thesis,* Norwegian Institute of Technology, March, 1995.

[12] A. Kobsa and W. Pohl, "The User Modeling Shell System BGP-MS," *User Modeling and User-Adapted Interaction*, Vol. 4, No. 2, 1995, pp. 59-106.

[13] T. Kamba, K. Bharat, and M.C. Albers, "An Interactive, Personalized, Newspaper on the WWW," *Proc. 4th Intl. World Wide Web Conf.*, 1995.

[14] T.D.C. Little, G. Ahanger, H.-J. Chen, R.J. Folz, J.F. Gibbon, A. Krishnamurthy, P. Lumba, M. Ramanathan, and D. Venkatesh, "Selection and Dissemination of Digital Video via the Virtual Video Browser", *Journal of Multimedia Tools and Applications*, Vol. 1, No. 2, June, 1995, pp. 149-172.

[15] Maes, P.,"How to do the right thing," AI-Laboratory, Vrije Universiteit Brussel and AI-Laboratory, MIT, Cambridge MA, 1989.

[16] Maes, P., "Agents that Reduce Work and Information Overload," *Communications of the ACM*, Vol. 37, No. 7, July 1994, pp. 31-40.

[17] Maes, P. and Shardanand, U., "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," *Proc. ACM CHI '95*, Denver CO, May 1995.

[18] T.W. Malone, K.R. Grant, F.A. Turbak, S.A. Brobst, and M.D. Cohen,"Intelligent information sharing systems," *Communications of the ACM*, Vol. 30, No. 5, May 1987, pp. 390-402.

[19] R.S. Marcus, "Computer and Human Understanding in Intelligent Retrieval Assistance," *Proc. 54th American Society for Information Science Meeting*, Vol. 28, October 1991, pp. 49-59.

[20] N. Mathe and J. Chen,"A User-Centered Approach to Adaptive Hypertext based on an Information Relevance Model," *Proc. 4th Intl. Conf. on User Modeling (UM'94)*, Hyannis MA, August 1994, pp. 107-114.

[21] J. Monaco, *How to Read a Film. The Art, Technology, Language, History and Theory of Film and Media*, Oxford University Press, New York, 1981.

[22] R.B. Musburger, *Electronic News Gathering*, Focal Press, Boston, 1991.

[23] T. Holm Nelson,"A File Structure for the Complex, the Changing and the Indeterminate," *Proc. 20th National ACM Conf.*, New York, 1965, pp. 84-100.

[24] J. Orwant, "Heterogeneous Learning in the Doppelgänger user Modeling System," *Journal of User Modeling and User-Adapted Interaction*, 1995.

[25] A. Perrig and A. Ballim, "The Design of a User Classification System," *http://diwww.epfl.ch/~aperrig/userap/userap.html*, Project of 8th Semester LITH, EPFL Lausanne, Switzerland, 1996.

[26] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proc. CSCW '94*, Chapel Hill, NC, October 1994.

[27] G. Salton, *Automatic Text Processing – The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, 1989.

[28] U. Upendra Shardanand and P. Maes,"Social Information Filtering: Algorithms for Automating Word of Mouth," *Proc. ACM CHI '95*, 1995.

[29] B. Sheth and P. Maes,"Evolving Agents for Personalized Information Filtering," *Proc. 9th IEEE Conf. on Artificial Intelligence for Applications*, 1993.

[30] B.D. Sheth, "A Learning Approach to Personalized Information Filtering," *Master's Thesis*, MIT, Cambridge MA, February 1994.

[31] T.W. Yan, H. Jacobsen, H. Garcia-Molina, and U. Dayal, "From User Access Patterns to Dynamic Hypertext Linking," *Proc. 5th Intl. World Wide Web Conf.*, May 1996, Paris, France.

[32] Yeung, M. M., Yeo, Boon-Lock, Wolf, W. and Liu, B., "Video Browsing using Clustering and Scene Transitions on Compressed Sequences," *Proc. Multimedia Computing and Networking 1005*, SPIE, San Jose California, February 1995.