Video Content Personalization for IPTV Services

David C. Gibbon, Zhu Liu, Harris Drucker[+], Bernard Renger, Lee Begeja, and Behzad Shahraray
AT&T Labs – Research
200 Laurel Ave. South
Middletown, NJ 07748, USA
{dcg, zliu, renger, lee, behzad}@research.att.com
[+]drucker@monmouth.edu

Abstract

IPTV customers will have access to thousands of video content sources and will require powerful yet intuitive tools to locate desired content. We propose a solution based on stored user interest profiles and multimodal processing for content segmentation to produce manageable content subsets for users. Segmentation involves part-of-speech tagging on extracted closed-caption text combined with video shot boundary detection. Query relevance ranking with temporal and other metadata constraints is used to form timely, focused content sets for users. We present a fully automated end-to-end system including standard definition video acquisition, media processing for segmentation, a content storage and retrieval subsystem, and content presentation via a user interface designed for set-tops requiring only a simple remote control for browsing.

Introduction

IPTV offers the ability to go beyond the current bandwidth limitations of broadcast or cable by effectively moving the tuning function upstream. While arguably more complex than legacy systems, IPTV systems support not only the existing TV channel model, but also pave the way for a consumption model that is more internet-like, where the number of broadcasters could be as large as the number of internet sites. RSS technologies such as MediaRSS and Video Podcasts offer an alternative content syndication and delivery mechanism and support a feed-based consumption model. In the shorter term, IPTV offers the ability to cost-effectively deliver international channels and niche channels where viewership does not support the financial returns required for traditional TV distributions systems.

All these trends yield more content options for the viewer and yet present a challenge. Users already have trouble locating programs on their 100 cable channel lineups; surfing sequentially to locate content becomes impractical as the number of channels moves into the 1000s. Since most media business cases rely on advertising revenue, the system as a whole will not be economically viable if the users cannot locate the content that they desire or that is relevant to their particular interests. The desired content must be presented to viewers through intuitive, simple interfaces designed to work using only an inexpensive infrared remote control device for input and the low spatial resolution of the standard definition television screen for output.

Automated media processing for content segmentation and indexing combined with user interest specifications enables fine grained filtering of large volumes of content to produce manageable content subsets. User interests can be captured on a PC as part of a self-provisioning process, and can be augmented over time based on viewing habits. Viewers can effectively define their own channels, where video clips from a wide range of sources can be clustered based on a topic or on a highly specific genre and presented in sequence to generate a customized programming experience. In this way, the consumption model resembles RSS feed reading more than a broadcast channel paradigm.

While designed primarily for an IPTV environment, the personalized content subsets concisely represent timely relevant content and therefore enable effective synchronization to portable mobile devices such as Apple iPods, Sony PSP, Microsoft Zune, etc. These devices have similar constraints on the user input (e.g.,

text entry is difficult at best) and the screen resolutions are lower than for desktop PCs.

Other solutions for managing large volumes of content include the use of relevance feedback, recommendation engines, filtering based on global genre tags from the Electronic Program Guide (EPG) or social tagging. Additionally, text entry on the set-top has been enabled using T9 or soft keyboards for search. These schemes are cumbersome and do not yield a positive user experience. Also, in many cases only a segment of a long form program is of interest and most systems identify content on a per-program basis.

Content personalization has been an active research area for a number of years. Dimitrova et al. have developed systems for personalizing DVR content with similar design goals in mind [1]. We have reported on content personalization and repurposing for mobile devices [2] and this paper extends that work to support a set-top box usage scenario. We also have developed an alternative segmentation methodology that improves accuracy and scales well as the number of users increases. Methods for determining content hierarchies automatically [4] can be applied to this domain as well. Our focus here is on segmentation which is a sub-component of hierarchy determination.

Content Personalization

Multimodal topic or story segmentation techniques can be employed to identify semantically consistent segments from long-form video programs. In our earlier work [2], we reported on a multimodal topic segmentation technique in which the textual component was seeded by user query terms. While this was effective, we describe a different approach here which involves content segmentation at ingest time (in a query-independent manner), followed by clip relevance ranking and filtering to form personalized clip sets for each user. These constitute clip play lists, or if we think in terms of the long-form source content, the data structures more closely resemble edit decision lists (EDL).

Content Segmentation

For each long-form media asset (typically 30 to 120 minutes in length,) a two phase scheme is used. In the first phase, a rough segmentation is created using commercial boundary detection. This is motivated by the observation that the topical units that we consider do not span commercial boundaries. This assumption is true for typical broadcast news programs, but may not hold for other genres. In the second phase, these segments may be further divided into topical units. For real-time captioned content, the transitions from roll-up caption mode to pop-up mode can be used to identify commercials. Many other media cues can be used, such as detecting black frames, or identifying transitional phrases in the CC text [6]. We do not use the ">>>" markers sometimes found in the closed caption to indicate topic change since these are not used consistently, and are often completely absent. However, hybrid approaches which use these cues when present may improve the segmentation accuracy.

Text Mode Topic Segmentation

After segmenting long form content at commercial boundaries, the segments range from 2 to 10 minutes in length. A topic segmentation algorithm operating solely on the text component attempts to divide these segments further on topic boundaries. Tentative topic segmentation consists of the following steps and is entirely based on the closed captioned text or automatic speech recognition of the audio component of the video programs:

Procedure:

Input: a set of sentences corresponding to the program dialog transcription for a program unit between commercial breaks, typically from the processed closed caption. For non-commercial content, the entire program text may be used, with slightly lower accuracy.

1. Use a part-of-speech tagger to mark all nouns.
2. Stem all nouns to their roots.
3. Define an upper diagonal matrix S such the element $S(i,j) = 1$ if sentences i and j $(j>i)$ have at least one noun in common, otherwise zero.
4. Define a Density $D(i,j)$ between sentences i and j:

$$D(i,j) = \frac{\sum_{m=i}^{j-1} \sum_{n=m+1}^{j} S(m,n)}{(j-i+1)^r},$$

where the exponent r is obtained by cross-validation. The numerator is a count of the number of 1's in the upper triangular matrix between sentences i and j and therefore bounded by elements $(i,i+1)$, (i,j) and $(j-1,j)$

5. Find the set of sentences $(i_1, j_1), (i_2, j_2), \ldots (i_k, j_k) \ , \ldots, (i_K, j_K)$ with $j_k > i_k$ and $j_k + 1 = i_{k+1}$ such that the following is maximized:

$$J = \sum_{k=1}^{K} D(i_k, j_k)$$

J can be found by dynamic programming and basically finds the K sets of sentence intervals $(i_k \rightarrow j_k)$ such that the sum of the densities over these K intervals is maximized.

The algorithm has been developed and tested using the LCD TDT-3 dataset which includes closed caption data from CNN, NBC, ABC, and PRI and includes topic boundary indications.

This process assumes that a good algorithm exists to determine sentence boundaries and we do so using the closed captions in the programs which typically include end of sentence punctuation. We also constrain the dynamic programming algorithm so that topic segments are at least three sentences long.

This procedure is based on [3] except they used an additional penalty term if the segment lengths are too long or too short. The attributes "too long" or "too short" are based on experimental analysis of the average length and standard deviations of segment lengths. However, we did not find that using this additional penalty term helped. In addition, they used all words, while we used nouns.

The algorithm works better if the chunk of text is as short as possible. Hence, it is desirable that the chunks be the sentences between commercial boundaries. While it is true that sometimes coherent segments of text cross commercial boundaries, the before-commercial text is typically a transitional phrase such as "Coming up next…" and this can be readily detected and removed from the optimization using context-free grammars or other natural language processing techniques.

Clip Relevance Ranking

A second phase of content preparation occurs after all ingested content has been segmented and stored in a multimedia database. The database maintains the full textual contents of the clips, along with associated metadata such as air date, program title, genre, etc. Thumbnail images extracted from the video and program icons are also available for user interface rendering.

The user interests are expressed as queries which may be as simple as a small set of search terms (words and phrases) or may involve more detailed information derived from other content which the user has identified as being relevant. The query is executed against the database using standard information retrieval (IR) techniques including stemming and basic logical operations and a set of candidate content clips is returned. In addition to full text search, temporal and genre metadata restrictions are applied to increase the relevance of the resulting content subset. We have used a window corresponding to the most recent 30 days of acquired content.

While this content set represents all content in the database that may be of interest to the user, most applications require further filtering to generate suitable content sets. We use a relevance ranking based on term-frequency. In contrast to traditional IR, we can use a true frequency (occurrence of query terms per second) since the content duration is available a priori. Only clips with rank greater than a minimum are included in the final content presentation for the user. This hit rate parameter can be tuned for different applications. For example, a service for delivering news clips to wireless users may be configured so that only highly relevant clips are sent so as to minimize data service costs. For our application, we have used a value of 0.002 hits per second which ensures that a user's query term will appear in the video at least once every 8 minutes on average.

We have found that a further rules-based filtering stage results in a more desirable user experience. Specifically, we impose the constraint that at least one of the user's query terms occur in the video presentation within a predetermined maximum time limit. A value of 30 seconds is suitable for most users and content types that we have encountered.

A final level of content set filtering is applied to remove duplicate clips. This situation arises when sets of user queries are processed as described above and there is an intersection in the resulting clip sets. For example, a user may be interested in video about President Bush and video about China. If the President makes state visit to China, the same video clip may be relevant to both topics. For some applications it may be desirable to display the clip twice to the user, but in our case, where we view the content presentation in aggregate across all topics that the user is interested in, we choose to suppress the repeated content.

Prototype System Design

To determine the feasibility of the proposed method for automated content personalization, we have designed and implemented a fully automated end-to-end system that indexes, transcodes and personalizes content from a number of content sources on a daily basis. We also describe a set-top box client for lean-back consumption.

Acquisition and Processing

Consumer DVRs provide robust, cost effective MPEG-2 encoding and storage as well as EPG metadata capture facilities. The encoding quality is not professional grade but some DVRs also support direct capture of digital TV. The video is transcoded using Direct Show filters from 6Mb/s MPEG-2 to 2Mb/s MS WM9 while preserving the spatial resolution of 720x480 to enable efficient streaming from a MS Windows Media server for retrieval. The closed caption (CC) is extracted, aligned with the audio and case-restored to improve readability. Pixel domain video processing identifies shot boundaries and representative frames and 160x120 resolution JPEG thumbnails are generated to provide a visually browsable summary. The content segments are identified as described above. For each video program, the processing results are stored in an XML representation. Collections of video programs are maintained on a media server. More information on the acquisition and processing system is available in reference [7].

User Preference Interface

Users can self-register with the system using a web application designed for desktop PC or laptops. This application also supports specifying the interest profile by allowing users to set up topics and query terms for each topic. Other user preferences such as genre or selecting content providers for each topic are supported. Users typically create 5 or fewer topics and the system is configured to generate and maintain 10 clips per topic, but the user can configure this parameter. While it is possible to use a set-top application for this purpose, it is too cumbersome for the user to enter text using a simple remote control. This hybrid system approach, combining both the desktop and the set-top is a significant aspect of the system design that users find to be highly desirable.

We can use information retrieval techniques to determine user preferences even when there is no information in a user profile. By tracking the segments that are viewed, we can use the existing metadata in those segments to enhance the user preference profile. We can combine the list of segments viewed fully, partially and segments that were skipped, into a set of user profile metadata that can be used to generate better personalized video segments. This technique is an informal use of relevance feedback: rather than have the user make explicit choices, we use the implicit choices as our feedback data.

For each registered user of the system, a user interest profile is stored in XML format to encode their interests as well as other information such as contact information. Whenever the profile is updated by the user, or when new content is added to the media server, the following operations are performed:

1) the user profile data is read and queries are formed,
2) the queries are executed against the multimedia database and the results are filtered as described above,
3) the resulting personalized content representations are maintained as a persistent data structure.

A key feature of the service is to automatically aggregate the clips of diverse sources, providing a multimedia experience that revolves around the user's provided profile. This means that users have direct access to their specific items of interest. The content representations for each user are compact since they consist primarily of clip metadata and pointers to media assets which are effectively shared across many users. A hierarchal structure is maintained in XML format as follows:

```
<user>
     <topic>
          <clip>
            [metadata,
             media pointers]
          </clip>
     </topic>
</user>
```

This representation is then used as a basis for multiple renderings for various consumption scenarios through the application of different XSL templates. The user may view the content immediately on the desktop web application, but the preferred consumption model involves a set-top environment. The profiles make searching personalized content easier as the user does not need to retype search strings every time the service is used. Figure 1 depicts a personalized content summary screen designed for display on a standard definition television monitor.

Figure 1: Results screen with excerpts from several video programs related to a topic of interest

Set-top Retrieval System Architecture

To support rapid prototyping, we have chosen Windows Media Center Edition (MCE) [5] as a client development environment. This allows the use of the XBOX 360 gaming platform running as a Media Center Extender to serve as the set-top box. While the capabilities of the XBOX exceed that of typical IPTV set-tops deployed today, we have designed the client application as a lightweight hosted-HTML application which is suitable for less capable platforms as well. The hardware components for the retrieval system are shown in Figure 2. The XBOX 360 with a universal media remote control drives an SD or HDTV monitor. This is connected to a PC or laptop running MCE which can also run the demonstration using an LCD monitor at a resolution of 1024x768 and using a USB IR receiver. The user interface is generated by an HTTP application server, and a media server delivers the streaming video to an embedded video player on the MCE. This is the same player that is used for playing back live or recorded television, so consistency of the user experience is maintained. The software architecture for the prototype is based on available consumer electronics equipment and leverages standards-compliant components such as XSL, XML, HTTP, etc. where possible.

The set-top software architecture is shown at a high level in the left side of Figure 3. Client side scripting with dynamic HTML behaviors is used to perform navigation amount the UI elements using the remote control for input. Key presses are mapped to keyboard-codes (this design allows the use of keyboard arrow keys as an alternative input modality if desired) and elements are highlighted to indicate the current focus. The client scripting also responds to the remote "select" or "ok" key (or keyboard "enter" key) and launches URLs to navigate the user interface or initiate video playback using a shared viewport which is similar in function to an embedded Windows Media player. The right side of Figure 3 depicts the server software architecture. The primary operation of the server web application is to apply an XSL transformation to the XML content descriptions that have been generated previously for the user. This transformation could also be implemented to take place on the client to further offload the server if required.

The graphical display is greatly simplified as compared with typical desktop applications due to the limited spatial resolution of television displays. Interlacing and TV overscan are additional issues that had to be taken into consideration in the UI design. A fully featured system would include accommodations for 16:9 aspect ratios and customized versions of 720p and 1080i or 1080p displays in addition to standard definition displays.

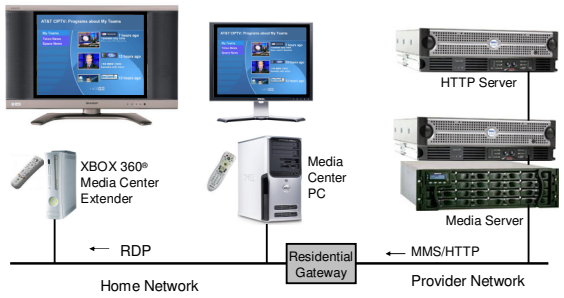

Figure 2: Prototype system components (retrieval)

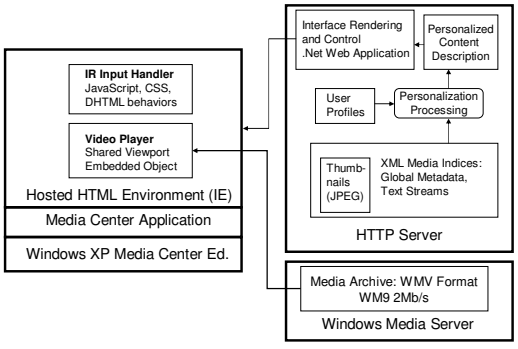

Figure 3: Prototype system software architecture

Conclusions

As media consumption models evolve to take advantage of technology advances such as time shifting and RSS content syndication, on-demand will supplant live broadcasting as the dominant method of viewing.

As this transition occurs, content personalization will move from an optional feature to a necessary feature and finally to an expected feature. The need for personalization increases as the amount of available on-demand content increases.

Our proposed system and media service concept leverages available metadata, augments it with automatically extracted data to provide a rich base of information for content personalization. The generated content presentations have been well received by users in our initial small-scale trials. The system is stable and has been running with minimal operational support for several months in a lab environment. We have found that highly structured content such as nightly news programming is most amenable to automated methods, but other content types can be successfully segmented as well using our method. The user interface design is critical to providing an overall acceptable user experience.

We anticipate that as IPTV and production workflow automation asset management systems become more commonplace, more detailed manually authored metadata will be available with the media. Also, an increasing amount of short-form content is becoming available on the internet. These trends suggest that personalization research should be directed more towards automated content organization and structuring and less on segmentation in the future. However, content segmentation will continue to be an important system capability, particularly for second tier content such as training material and enterprise content.

The methods presented for automated content organization based on user interest profiles provide a good first step toward alleviating the problem of content overload that media consumers face today. The implemented system provides a solid testbed for prototyping content personalization service concepts.

References

[1] N. Dimitrova, R. S. Jasinschi, L. Agnihotri, J. Zimmerman, T. McGee, D. Li, The Video Scout System: Content-based analysis and retrieval for Personal Video Recorders, in Handbook of Video Databases, CRC Press, June 2003, ISBN 0-8493-7006-X.

[2] D. Gibbon, L. Begeja, Z. Liu, B. Renger, and B. Shahraray, "Creating Personalized Video Presentations using Multimodal Processing," Handbook of Multimedia Databases, Edited by Borko Furht, CRC Press, pp. 1107-1131, June 2003, ISBN 0-8493-7006-X.

[3] P. Fragkou, V. Petridis, and A. Kehagias, "A Dynamic Programming Algorithm for Linear Text Segmentation", *Journal of Intelligent Information Systems*, vol 23, #2, pp. 179-197, September 2004.

[4] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, B. Shahraray, *Automated Generation of News Content Hierarchy By Integrating Audio, Video, and Text Information,* 1999 IEEE International Conference On Acoustics, Speech, and Signal Processing Phoenix, Arizona Volume: 6, March 15-19, 1999, Pages: 3025-3028.

[5] Microsoft, "Microsoft Windows XP Media Center Edition 2005 Reviewers Guide," 2004.

[6] A. Hauptmann and M. Witbrock, Story Segmentation and Detection of Commercials in Broadcast News Video, *ADL-98 Advances in Digital Libraries*, Santa Barbara, CA, April 22-24, 1998.

[7] Z. Liu, D. Gibbon, B. Shahraray, "Multimedia Content Acquisition and Processing in the MIRACLE System," in *IEEE CCNC,* January 7, 2006.