Content Augmentation Aspects of Personalized Entertainment Experience

Nevenka Dimitrova[1], John Zimmerman[2], Angel Janevski[1], Lalitha Agnihotri[1], Norman Haas[3], and Ruud Bolle[3]

[1] Philips Research, 345 Scarborough Rd., Briarcliff Manor, NY 10510, USA
{nevenka.dimitrova,angel.janevski,lalitha.agnihotri}@philips.com

[2] Human-Computer Interaction Institute, Carnegie Mellon, Pittsburgh, PA, USA
johnz@cs.cmu.edu

[3] IBM T.J. Watson, 30 Saw Mill River Road, Hawthorne, NY 10532, USA
{nhaas,bolle}@us.ibm.com

Abstract. We present personalization aspects in information and video content retrieval and augmentation applications. Our goal is to explore the value of linking related content among different media such as TV and Web. Metadata extraction can make a great difference for personalized user experience. Annotations that provide title, genre, description, and cast can be greatly enriched with detailed information at subprogram level, i.e. at the scene and story level. This enhanced metadata can be matched against the user profile for prioritizing information and entertainment.

1 Introduction

For many years, television has provided rich content, delivering news, information and entertainment. More recently, the web has developed as a rich content source, more overwhelming to navigate and master. Linking the related information between these two media, it seemed to us, could create an enhanced, more personalized, TV viewing experience. We set out to explore the possibilities, in this project.

At a high-level, we wanted to explore how information linking and personalization can bring value to users. We call this research direction *Content Augmentation*. As an example, imagine a viewer is watching a TV news story about Brazil. Because a TV program is expensive to produce, the content creators can only afford to play a short story, of around two minutes. After watching the story, a viewer may want to know more about Brazil. A content augmentation application could understand that the news story is about Brazil, and provide the user with appropriate, summarized, and targeted information, as well as references (e.g. web links) for further exploration. In addition, this application could employ a user profile, personalizing the linked content by prioritizing the types of links a user is likely to want. To test this model, we developed a pilot system. We began by focus group-testing several concepts, and, based on the

group's reaction, designed and implemented a personal news application (MyInfo) and a movie information retrieval application (InfoSip).

While the majority of TV personalization research conducted today focuses on examining TV shows at a program level, our system utilizes video content analysis to process TV programs at a sub-show level. It first processes the video in order to create content-based metadata at a subprogram level and stores this metadata with the TV program. At the same time, the system extracts relevant web content based on a personal profile. MyInfo extracts specific web content listed in the profile and can display a personalized TV news program based on the web content and prioritized, individual TV news stories. This approach is a step further than existing news retrieval systems in the literature that are focused on broadcast news [Merlino et al. 1997]. InfoSip identifies actors in individual scenes. It then extracts web data, including the latest filmographies and biographies. While watching the movie, users can press a button and see a list of all actors in the current scene. In addition, they can view filmographies based on their TV viewing history. These content augmentation applications offer a new direction for personalization research, where the source of the content is less important than the actual delivered information to the user.

2 Augmented User Experience

Existing TV user experience is based on a 50-year-old tradition. Broadcasters tailor TV programming to capture mass audiences. They used both demographic data on viewers and input from advertisers to determine which programs to play at the various times of day. More recently, with the emergence of niche-based TV channels such as CNN (news), MTV (music), and ESPN (sports), viewers have had a little more control over when they view the content they desire. Until recently, viewers had only printed guides to help them plan their viewing and select TV shows. More recently, electronic program guides allow viewers to browse the program offerings by genres, date, time, channel, title and, in some cases, search using keywords.

Next, we describe the existing TV user experience in a broadcast setting and the potential of connected media delivery modes for enhanced user experience.

2.1 The TV Experience of Current EPG, PVRs, and Recommenders

Current electronic program guides found in products such as DirecTV and EchoStar's digital satellite settop boxes, cable settop boxes from Time-Warner Cable and CableVision, and personal video recorders by TiVo and ReplayTV, offer users several methods for searching and browsing TV listings. All these systems hold one to two weeks worth of TV data, which users can view by time and channel. Higher end systems allow users to browse by show title and keywords. Finally, the TiVo system even offers a recommender to help users find something to watch.

Although TiVo is currently the only commercial product with a recommender, much personalization research has been done in this area. Das and Horst developed

the TV Advisor, where users enter their explicit preferences in order to produce a list of recommendations [Das et al. 1998]. Cotter and Smyth's PTV uses a mixture of case-based reasoning and collaborative filtering to learn users' preferences in order to generate recommendations [Cotter et al. 2000]. Ardissono et al. created the Personalized EPG that employs an agent-based system designed for set top box operation [Ardissono et al. 2001]. Three user modeling modules collaborate in preparing the final recommendations: Explicit Preferences Expert, Stereotypical Expert, and Dynamic Expert. And Buczak et al. developed a based recommender that uses a neural network to combine results from both an explicit and an implicit recommender [Buczak et al. 2002]. What all these recommenders have in common is that they only examine program-level metadata. They do not have any detailed understanding of the program, and cannot help users find interesting segments within a TV program.

The Video Scout project we previously developed offers an early view of personalization at a subprogram level [Jasinschi et al. 2001, Zimmerman et al., 2001]. Video Scout employed a GUI construct called “TV magnets” (Figure 1). Users can specify financial news topics and celebrity names. The system then watches TV and extracts segments, searching the contents of talk shows for matching celebrity clips and searching the contents of financial news programs for matching financial news stories.

Figure 1. Financial news magnet screen with four stored clips from two TV shows.

2.2 Content Extraction: Beyond High-level Metadata

The technology described in the previous section provides mostly coarse tools for content personalization and navigation. However, with the advancement of video content analysis, indexing, and retrieval, we can offer the consumer a more powerful set of tools for an abundant set of media sources including TV, Web, and radio.

The system diagram in Figure 2 shows the high-level chain of content processing and augmentation. Unannotated or partially annotated content is delivered to the service provider (e.g. content provider, broadcaster) where generic analysis and augmentation is performed. The first step extracts features and summarizes the content and content segments, generating descriptive metadata. A more detailed description of this step is given in Section 3. The generated metadata in conjunction with any pre-

existing metadata, is then used to augment the content with additional information from web sources. This augmentation is general in that it is not based on any personal profile. Following broadcaster augmentation, the content with the complete metadata is formatted and delivered to the consumer device.

The consumer device ("client" in Figure 2.) has the capability of storing content and metadata, and also contains the user profile in conjunction with the prioritization module. This is used to perform a secondary augmentation with web information, but this time based solely on user preferences. The information obtained is stored together with the content and is presented to users as if it were a part of the original program.

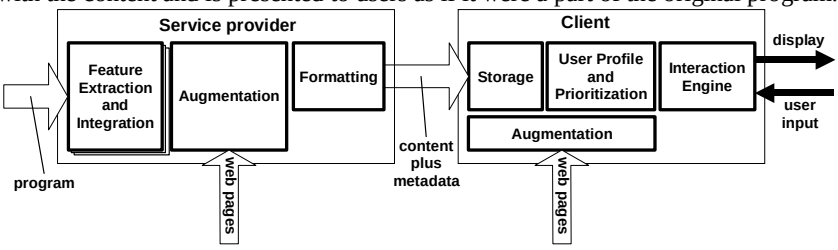

Figure 2. High-level system diagram

2.3 Applications

Detailed metadata extracted locally or delivered from a server or a peer device offer a host of possibilities for personalized applications. We divide these applications into retrieval and augmentation.

There are two distinct modes for a retrieval application: (i) the user proactively searches through the metadata in order to find segments of interest, such as querying a digital library, (ii) a user profile acts as a query against continuously updated metadata of the latest stored news, narrative or other content, such as Video Scout – a content-based PVR [Jasinschi et al. 2001, Zimmerman et al. 2001].

Content augmentation enhances the user experience by automatically utilizing the user profile. We have researched and prototyped two applications: MyInfo and InfoSip. MyInfo delivers news from TV and web channels according to a user profile. To access the news, users select one of the six "content zones": weather, traffic, sports, financial news, headlines, and local events (Figure 3). InfoSip answers most frequently asked questions, such as who, when, where, and what. Using a remote control and the current play context, users indicate their query, such as "Who is that actor?", "What is this song?" and the system overlays the answer onscreen, allowing users to sip relevant information while watching movies (Figure 4).

3 Metadata Extraction for Personalized User Experience

Methods for automatic metadata extraction can be divided into coarse- and fine-grain segmentation and abstraction. In this section, we briefly introduce the methods used.

3.1 Segmentation

We have developed a high-level segmentation method, in which we first find the commercial breaks in a particular news program and then we perform story segmentation. For stories, we use the story break markup (“>>>”) in the closed captioning. This method has been used in the literature before [Maybury 2000]. For commercials, we use a transcript-based commercial detector. In part, this relies on the absence of closed captioning for 30 seconds or more, and in part, it relies on the news anchors using cue phrases to segue to/from the commercials. We look for onset cues such as “right back”, “up next” and “when we return”, in conjunction with offset cues, such as “welcome back” and the “new speaker” markup (“>>”). We tested commercial detection on four financial news and four talk show programs, totaling 360 minutes, with 33 commercials totaling 102 minutes. Our algorithm detected 32 commercials totaling 104 minutes. Of these, 25 were exactly right. Only one commercial was completely missed. We detected 4 extra minutes spread out over seven commercials. The resulting recall and precision are 98% and 96% respectively.

In addition we have developed a single descent method for story segmentation that relies on multiple cues: audio, visual, and transcript. The single descent method relies on unimodal and multimodal segment detection. Unimodal – within the same modality – segment detection means that a video segment exhibits “same” characteristic over a period of time using a single type of modality such as camera motion, presence of certain objects such as videotext or faces or color palette in the visual domain. Multimodal segment detection means that a video segment exhibits a certain characteristic taking into account attributes from different modalities. Next, we use multimodal pooling as a method to determine boundaries based on applying votes from multiple modalities. In the single descent method, the idea is to start at the top of the stacked uniform segments and perform a single pass through the layers of unimodal segments that in the end yields the merged segments. In this process of descending down through the different attribute lines, we perform simple set operations such as union and intersection. For example, one approach is to start with the textual segments as the dominant attributes and then combine them with the audio and visual segments.

3.2 Summarization

Each broadcast news story must be summarized in order to use (i) the abstracted data for matching against the personal profile and (ii) the summary for presentation browsing. Although there are different forms of summaries, in our case a summary consists of a sentence of text and a representative image (keyframe). Another implicit form of summarization is categorization (assignment to one of six “content zones”).

The closed captioning text sent with each frame of the story is collected. Since it is usually mono-case, it is recapitalized. (We used our own algorithm for this, but more thorough algorithms have been developed [Brown et al, 2002]). Then IBM TextMiner document summarizer is applied, to select the single best sentence in the “document” [TextMiner]. “Best” in this context is a weighted metric involving the “salience” (position) of the sentence, its length, and other factors. Usually the first sentence of a

news story is selected; this sentence is normally both a comprehensive summary and a good introduction. However, if a non-useful sentence occurs first ("Hello, I'm Dan Rather."), TextMiner often catches these and makes a better selection.

We also use a TextMiner engine for document (story) classification. It works on the basis of frequency of occurrence of words in the story. The classifier was trained off-line with a corpus of labeled exemplar stories.

Finally, we try to find a representative keyframe for each news story. Each story is composed of an anchor shot followed by *reportage* shot(s). The anchor shot is the same for all stories, so it does not provide any value. In order to select an image from only the *reportage*, we developed an anchorperson detector, based on the Kolmogorov-Smirnoff test. Basically, it finds the face, which is the largest face in view in the majority of the frames of the entire news program. Further, we have empirically determined that outdoor shots are more important for news stories than indoor shots. We use the indoor/outdoor detector developed by Naphade [Naphade et al. 2002]. Subject to these constraints, we select a frame that is deemed to be the most "interesting" by the algorithm that considers other attributes such as color, etc.

3.3 Person Identification

A rich "frequently asked question"-answering application relies on manual annotation or automatic detectors. For example, to answer the "who is this person" question in a movie, documentary, or home video, we need to know the people that are present in the each of the scenes. The challenge is to robustly identify persons from different views, distance, lighting conditions, in various background noise conditions. We have used automatic face and voice identification methods for this task [Li et al. 2001].

A person identification approach is constructed based on the joint use of visual and audio information. First, in the *analysis* phase we perform visual analysis for detection, tracking, and recognition of faces in video. Face trajectories are first extracted and the Eigenface method is used to label each face trajectory as one of the known persons in database. Due to the limitation of existing face recognition techniques and the complex environmental factors in our experimental data, the recognition accuracy is not high. Next we employ audio analysis, which operates by speaker identification. Both audio and visual analysis have their advantages under different circumstances, and we studied how to exploit the interaction between them for improved performance. In the fusion phase, two strategies have been employed. In the first strategy, the *audio-verify-visual fusion* strategy, speaker identification is used to verify the face recognition result. The second strategy, the *visual-aid-audio fusion* strategy, consists of using face recognition and tracking to supplement speaker identification results.

3.4 Generic Detectors

We have developed a number of other detectors that in the future could be integrated into the system. For example, a natural scene vs. graphics detector can be used to differentiate information-rich natural image from a relatively information-poorer syn-

thesized graphic [Naphade et al. 2002]. Other event detectors can be of great value, for example, highlights in sports, explosions, and music segment detectors.

4 Personalized News

Personalization provides one of the greatest benefits of the MyInfo application. However, it is also one of the greatest risks. From our focus group sessions we learned that users must feel that they are in control of the system, or they will quickly abandon the application. MyInfo personalizes the segmented and summarized news (see section 3.2) in two ways. For the Internet data, the system parses and extracts information from web sites as specified in the user profile. For the TV news stories, the application prioritizes individual stories based on both topics of interest listed in the user profile and on cues broadcasters use to indicate the importance of a story.

4.1 Personal Profile

In designing the personalization of the Internet content, we focused on providing users with maximum information using minimal input. The traffic section of the profile contains the user's home address. In addition, it contains a set of destinations and "hot spots". Destinations include towns or prominent structures such as malls, stadiums, airports, etc. Hot spots include points of constriction like bridges and tunnels, which notoriously have traffic delays. Once selected, the system extracts web traffic information on the specific hot spots and on the major roads between the users home and the selected destinations. For sports, the profile contains team names. As a default, once a zip code has been entered, MyInfo begins to track the teams in the local area. When users select the sports zone, they see the latest scores and upcoming events for the monitored teams. For financial news, the profile contains the names of indexes, stocks, and mutual funds. For local events, it contains a set of keywords describing events users like most. The system prioritizes a listing of local events based on their distance from the user's home and the match to the keywords.

Personalized web information improves the TV news experience in two ways. First, it makes it faster for users to get information. For example, if users just want to know the current temperature or a stock price, the information is a single button push away. They don't even have to wait for the news anchor to tell them. Second, the Internet-extracted data is of more benefit to users than traditional TV content because it is much more relevant to their information needs. For example, the local TV news can only afford to devote so many minutes of broadcast time each day for traffic information. This prevents them from relaying information on all routes during each clip. Often, they skip the specific routes that are important to an individual user. The personalized web data takes care of this issue, while still allowing users to see the traditional TV traffic news to get an overview of the worst spots in their area.

MyInfo collects personalized web information through Information Extraction (IE) [Janevski et al. 2002]. IE is a process of extracting structured data from a relatively

unstructured input (e.g. English free-form texts or a web page). An IE system contains one or more *rules* that describe generic ways of processing natural language texts and extracting the correct data. For a specific input document and output data, the rules are instantiated as *tasks*. In general, rules have a set of parameters and specific parameter values that define a task.

Our system obtains personalized web data using specific web IE tasks that depend on the user profile. For example, the current weather conditions and a weather forecast can be obtained from the weather.com site with a URL that contains a particular zip code. Here, the rule specifies how to extract the weather information, but the zip code defines a particular task with a customized page location. Similarly, the traffic information can be found for a geographic area. Once an area has been selected, information for the user's hot spots is extracted by instantiating an IE task with a list of hot spots of interest and for the particular time of the day/week. The system can extract information on different routes for the morning/afternoon and weekend commutes.

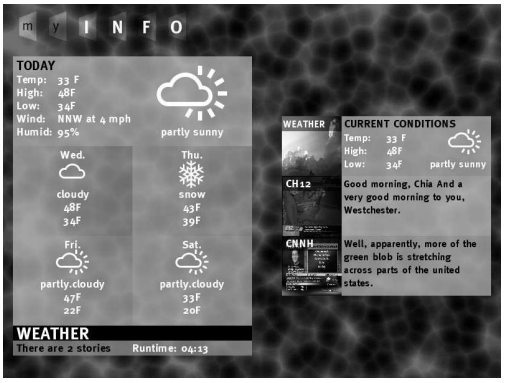

Figure 3. The Weather content zone screen.

4.2 Prioritization

MyInfo further personalizes TV news stories by prioritizing them. The system tries to balance topics users have specified in the profile with cues the broadcaster uses to indicate a story's importance. Use of the broadcaster information is very important. Users have no way of predicting every kind of news story that might be important to them. They may know they are interested in China, and therefore add this topic to their profile. However, it is hard for them to predict major events that affect many people, such as earthquakes, elections, etc. By allowing the broadcasters' editorial content decisions to play a role, users get a much better mix of information.

MyInfo determines broadcaster importance of a story from three different characteristics: (i) duration, (ii) location in the newscast, and (iii) teaser announcing a story will play later in the broadcast. Since broadcast time is limited, a longer story will be more important. Location in the broadcast and use of a teaser are subtler. The most important TV news stories generally appear at the beginning; however, broadcasters

place other stories they think many viewers want to see at the end. Then they use teasers to keep the viewers from switching channels. At this time, our measurement of broadcaster importance is far from perfect; however, using this value in conjunction with stories that match the user profile creates a richer viewing experience.

5 InfoSip: Personalized/Augmented Narrative

InfoSip is an example of a "frequently asked questions" answering application. It unobtrusively serves actor information related to the scene. During focus group testing, participants indicated that they wanted supplementary information for movies and TV shows, but they did not want it to interrupt viewing.

Figure 4. Screenshot of the InfoSip showing actor information.

With InfoSip, users interact by selecting a specific query. InfoSip uses predefined categories of questions/buttons such as "who", "where", "what", "when", "why", and "how much". For example, users press the "who" button to ask "who's that actor?". This displays a list of all of the actors in the current scene using annotated data from person identification (see section 3.3) and supplemental data about each one obtained through web IE (Figure 4). Web IE works better than metadata loaded on products such as DVDs, because it always contains the latest information. Filmography information is personalized, based on the user's viewing history. Highlighting the movies in which users have seen this actor increases the chances that they will remember why this person looks familiar.

6 Conclusions

In this paper we presented personalization aspects for content augmentation applications that combine content from multiple media sources. Our pilot applications MyInfo and InfoSip show promise that the technology has come of age. Web Information

Extraction and the segmentation, indexing, and retrieval of video at a subprogram level offer new tools for TV personalization developers. These technologies can improve the viewing experience by better understanding the TV content and by retrieving related material that is more focused at individual users. In the future we plan to evaluate our pilot applications with real users, continue developing video and web retrieval and extraction algorithms and generate more content augmentation concepts.

References

Ardissono, L., Portis, F., and Torasso, P.: Architecture of a System for the generation of personalized Electronic Program Guides. Proceedings of UM '01: Workshop on Personalization in Future TV, Sonthofen, Germany, (2001)

Brown, E.W., and A. R. Coden, A.R.: Capitalization Recovery for Text. A. R. Coden, E.W. Brown, and S. Srinivasan (eds.), Information Retrieval Techniques for Speech Applications, Springer, New York, (2002) 11-22

Buczak, A., Zimmerman, J., Kurapati, K.: Personalization: Improving Ease of Use, Trust, and Accuracy of a TV Show Recommender. Proceedings of AH '02 Workshop on Personalization in Future TV, Malaga, Spain, May (2002) 1-10

Cotter P. and Smyth, B.: PTV: Intelligent Personalized TV Guides. Proceedings of AAAI '00, Austin, TX, USA, (2000) 957-964

Das D. and ter Horst, H.: Recommender Systems for TV. Recommender System, Papers from the 1998 Workshop, Madison, WI. Menlo Park, CA: AAAI Press, (1998) 35-36

Dimitrova, N. Agnihotri L., and Jasinschi R., Temporal Video Boundaries, in Video Mining, A. Rosenfeld, D. Doermann, D. Dementhon eds., Kluwer Academic Publishers, 2003, 63-92.

IBM Intelligent Miner for Text™

Janevski A. and Dimitrova, N.: Web Information Extraction for Content Augmentation. Proc. IEEE Int. Conference on Multimedia and Expo, Switzerland, Aug., (2002), pp 389-392

Jasinschi, R., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J. Video Scouting: An Architecture and System for the Integration of Multimedia Information in Personal TV Applications, ICASSP, Salt Lake City, UT, USA, May 7-11 (2001) 1405-1408

Li, D., Wei, G., Sethi, I.K., and Dimitrova, N.: Person Identification in TV Programs, Journal on Electronic Imaging, special issue on Storage, Processing and Retrieval of Digital Media, October (2001), pp 930-938

Merlino, A., Morey D., Maybury, M.: Broadcast navigation using story segmentation. Proceedings of ACM MM '97, ACM Press, November (1997) 381-388.

Maybury, M. (ed.) February 2000. News On Demand. CACM. Volume 43(2), pp 32-34.

Naphade, M. R., Kozintsev I., and Huang, T.S.: Factor Graph Framework for Semantic Video Indexing. IEEE Trans. on Circuits & Systems for Video Technology, 12(1) (2002) 40-52

Zimmerman, J., Marmaropoulos, G., and van Heerden, C. Interface Design of Video Scout: A Selection, Recording, and Segmentation System for TVs. Proceedings of Human Computer Interaction Intl (HCII) New Orleans, LA, USA, August 5-10, (2001) 277-281