An adaptive videos enrichment system based on decision trees for people with sensory disabilities

José Francisco Saray Villamizar
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
jsaray@gmail.com

Benoît Encelle
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
bencelle@liris.cnrs.fr

Yannick Prié
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
yprie@liris.cnrs.fr

Pierre-Antoine Champin
Université de Lyon,
CNRS Université Lyon 1, LIRIS,
UMR5205, F-69622, France
pchampin@liris.cnrs.fr

ABSTRACT

The ACAV project aims to improve videos accessibility on the Web for people with sensory disabilities. For this purpose, additional descriptions of key visual/audio information of the video, that cannot be perceived, are presented using accessible output modalities. However, personalization mechanisms are necessary to adapt these descriptions and their presentations according to user interests and cognitive/physical capabilities. In this article, we introduce the concepts needed for personalization and an adaptive personalization mechanism of descriptions and associated presentations is proposed and evaluated.

Keywords

adaptive videos enrichment system, videos accessibility, videos annotation, decision trees, people with sensory disabilities.

1. INTRODUCTION

The amount of Web videos is continually growing up [17] and, as a result, challenges a lot of accessibility problems for people with disabilities [3][17]. The ACAV project [15] aims to explore how accessibility of Web videos can be improved and tackles two research questions [15] i) *what is required to make a video accessible on the Web and how can it be achieved*?, and ii) *how to increase the number of accessible videos on the Web*?

The ACAV approach consists in providing a tool for describing key visual/audio elements of a given video and another tool for presenting these descriptions during the playing of the video in an accessible way (according to the user disabilities) [15]. The work outlined in this article contributes to the development of the second tool: the question we address is how to provide relevant descriptions and relevant descriptions presentations to a given user during the visualization of a video?

Section 2 formally describes the research problem we tackle. Section 3 introduces the adaptive personalization mechanism we use and Section 4 evaluates this mechanism. Section 5 deals with related works. Finally, we discuss our work and highlight some future work in Section 6.

2. PROBLEM DESCRIPTION

Improving videos accessibility for people with visual and hearing impairments requires video *annotations* – i.e. additional descriptions (electronic texts) about key visual or audio elements attached to temporal intervals of the video [13]. For instance, according to [9], several *types* of visual elements have to be described: information about settings, actions, etc. –with as many complexities as possible [6] (hereafter called *Levels of Detail (LoD)* - e.g. for a setting annotation, “house” will have a LoD of 1, “house with two rooms” will have a LoD of 2, etc.).

As a result, an annotation contains a *typed* textual description of a given *LoD* and can be presented using a given *output modality* [15]: e.g. using a female synthetic voice (Text-To-Speech), or using a refreshable Braille display (with regular or contracted Braille). A video with the presentations of its associated annotations is called an *enriched video*.

However, predefined presentations of annotations for a given enriched video may not fully satisfy the needs of each kind of users: e.g. some visual impaired may want deeper details about characters and settings while others want to get details about actions. Some read Braille and others not. An adaptation mechanism, capable of transforming the presentations of annotations during video visualization is thus needed.

Generally speaking, two distinguishable approaches for performing adaptation exist: adaptive and adaptable ones [4]. “Pure” adaptable approach is not really a good option for us because the end-user has to assume the adaption mechanism by setting up explicitly her preferences before playing the video. On the other hand, pure adaptive approach offers automated adaptation based on the analysis of the user behavior and user-system interaction traces, what better suits our needs.

Our approach is close to an adaptive mechanism: the user interacts with the system through common actions in a video player (e.g. the PLAY/PAUSE events) and through new actions. As an example, a “FEEDBACK” event - fired by a spacebar key press, is used to indicate to the system when the user dislikes the presentation(s) of annotation(s) (i.e. *rejected* if event is fired otherwise *accepted* by the user).

2.1 Formal description: definitions

Presentation attributes: Presentation attributes PA=$\{A_1,\ldots, A_n\}$ discussed in [19] are attributes concerning an annotation presentation. For our case, considered presentation attributes are annotation *type* (e.g settings, actions), *output modality* (e.g. Braille or Text To Speech (TTS)) and its *LoD*.

Domain of a presentation attribute: for a presentation attribute A_i we define the attribute domain Dom(A_i) as the set of discrete values $\{a_1, a_2, ..., a_n\}$ we can choose Ai from. For example if A_i = "Modality ", a possible Dom(A_i) will be {'Braille', 'TTS', etc.}.

Annotation presentation: an annotation presentation P is a set of values $\{V_1,V_2,\ldots,V_n\}$ for presentation attributes $\{A_1,A_2,\ldots,A_n\}$ such that $V_i \in$ Dom(A_i) .

W4A2011 – Communications, March 28-29, 2011, Hyderabad, India.
Co-Located with the 20th International World Wide Web Conference.
Copyright 2011 ACM 978-1-4503-0476-4...$5.00.

Rendering of annotation(s) presentation(s) at time t: a rendering R at instant t is the set of annotation presentations that are currently rendered at time t. $R(t) = <P_1,P_2,\ldots,P_m>$

Rendering Section at time t:

A rendering section is defined using a time interval $t_{\_begin}$, $t_{\_end}$ ($t_{\_begin} \leq t \leq t_{\_end}$) and a rendering of annotation(s) presentation(s) at time $t_{\_begin}$.

$S(t) = <t_{\_begin}, t_{\_end}, R(t_{\_begin})>$ where $t_{\_begin} \leq t \leq t_{\_end}$

$t_{\_begin}$ and $t_{\_end}$ indicate two consecutive changes concerning the presentations of annotations during the playing of an enriched video. As a result, a new section is created each time something changes on presentations of annotations (e.g. a new annotation presentation appears, an annotation presentation ends).

As a consequence, an enriched video contains *o* rendering sections ($o \geq 0$). For instance, Figure 1 presents the cutting of an enriched video in terms of rendering sections that are numbered (i.e. 5 rendering sections, numbered from 1 to 5).

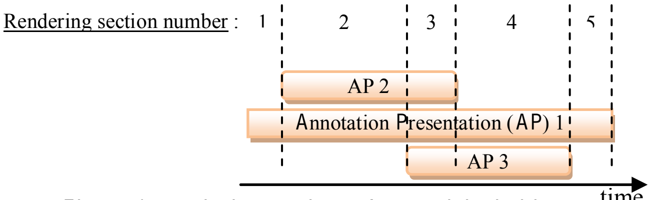

Figure 1: rendering sections of an enriched video

SCard: for a given numbered rendering section S_i, $SCard_i$ is the number of concurrent presentations it contains (i.e. $Card(R_i)$).

Feedback tuple F: A feedback tuple F is associated to each rendering section. $F_i = <S_i, f_i>$ where i is a section number, f_i the associated feedback: $Dom(f_i)$ = {"accepted", "rejected"}.

For a given rendering section S_i, if explicit negative feedback is provided by the user during S_i, it means that she dislikes what is being presented: $F_i = <S_i$, "rejected">, otherwise $F_i = <S_i$, "accepted">.

3. PERSONALIZATION PROCESS

3.1 Preliminary Discussion

Knowledge represented using predefined set of rules can be employed in personalization engines [20], moreover data-mining and artificial intelligence techniques (e.g. decision trees, Bayesian networks, etc.) are used to analyze patterns of user-system interactions and, hence to perform adaptivity.

As a consequence, we propose an approach where the system learns in a first stage (hereafter called the "*learning phase*") from given user feedbacks. After this stage, an "*adaptation stage*" is performed: the system uses what it learned about the user to predict her feedback (i.e. *accepted* or *rejected*) for any incoming presentation(s) of annotation(s). If the prediction is *accepted*, then the presentation(s) of annotation(s) is shown as it is, if the prediction is *rejected*, then some transformations has to be made before rendering the annotation(s).

The problem, specified in Section 2, has three specific features. Firstly, we hardly know almost anything about the user (e.g. maybe her disabilities and native language). Secondly, the user is the expert: she is the only one to know about her preferences, interests and we do not want her to introduce explicitly these information before video visualization (i.e. adaptable approach) in order to avoid this time-consuming initialization step. Third, adaptation has to be performed in real-time.

As a result, adaptation based on predefined rules [5] is discarded as we cannot easily model adaptation to all users with a fixed rule set. Thus Bayesian networks [12], decision trees (DT) [7][10][14], or any other supervised learning technique [20] seem to be suitable approaches. In this paper, we have chosen to use decision trees [7][10] as they have some interesting advantages in terms of learning and prediction speed. They are best suited to problems where instances are represented by attribute-value pairs and target function has discrete output values [14] (our case). In user modeling, decision trees can be used to classify users or documents in order to employ this information for personalization purposes [10]. Any decision tree induces a rule set on the problem domain that can be transformed in real time by non-expensive tree re-learning at any moment: this is compatible with current ACAV implementation and is also a reason for this choice.

3.2 A Decision-Tree Adaptation Approach

3.2.1 Learning stage

During a learning time *LT*, the system presents rendering sections S_i to the user and stores feedbacks until enough amount of feedback tuples F_i have been collected [14]. Next, rendering sections tables are built. Each table contains rendering sections that have the same SCard and each table line contains information used for the rendering of a given number of (overlapped) presentation(s) of annotation(s) (cf. Tables 1 & 2).

Table 2 shows how the renderings sections with SCard=2 are classified ($R_1(t_1)$ = <{Place,Braille,1},{Cloth,TTS,1}>, $R_2(t)$= <{ Character , TTS,1},{History,Braille,2}>).

Table 1. Rendering sections table for SCard = 1

Type	Modality	LoD	Feedback
Place	Braille	1	Accepted
Place	Braille	1	Rejected
Characters	Text to Speech	2	Rejected
Clothes	Text to Speech	2	Accepted

Table 2. Rendering sections table for SCard = 2

Type	Modality	LoD	Feedback
Place, Cloth	Braille, TTS	1,1	Accepted
Character, History	TTS, Braille	1,2	Rejected

After storing an important quantity of tuples (c.f. [14]) on these tables, a decision tree for each table is induced using the J48 algorithm. A decision tree as the one presented in Figure 2 is built (SCard = 1). In this tree, renderings predicted as "accepted" are represented as 0-tagged leaves, renderings predicted as "rejected" are represented as 1-tagged leaves. For a given leaf, the number between parentheses is the frequency in the learning sample of the path between root and that leaf. If there are two numbers in parentheses, leftmost number is the frequency and rightmost number is the number of opposite feedback(s) found in the learning sample for that path.

For example, the second leaf from left to right (prediction "rejected") means that the rendering section with R=<speech, L1, character> appears four times in the learning sample and

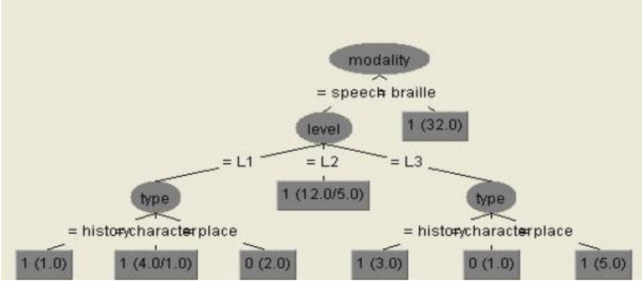

Figure 2. Example of an induced decision tree

was accepted just once by the user (and thus was rejected three times).

3.2.2 Adaptation

Once the prediction trees are built, the adaptation step begins: the system adapts any incoming rendering sections of SCard = n, predicted as “rejected”, using the decision tree from table of SCard = n, because it represents the user knowledge about section renderings acceptation. The tree on figure 2 classifies any possible rendering section of SCard = 1 (cf. Table 3).

Table 3. Group and associated section rendering templates.

Accepted	Rejected
<Place,Speech,1>	<*, Braille, *>
<Character, Speech, 3>	<History, *, *>
	<*, *, 2>
	<Character,Speech,1 or 2>
	<Place,Speech,2 or 3>

An intuitive adaptation solution consists in transforming any section rendering predicted as “rejected” using the most similar rendering template (taking SCard and rendering attributes values), located in the “accepted” column. However, a restriction has to be taken into account: the type of an annotation cannot be changed and its LoD can be changed only if different(s) LoD for this annotation exist(s).

As a consequence, if a section rendering predicted as “rejected” (SCard=1) matches a rejected section rendering template <X, *, *> (e.g. <History, *, *> in the Table 3), that means the user is not accepting any annotation of this type and the annotation is not presented. Concerning others cases, the following distance/similarity measure, based on [14] is suggested.

3.2.3 Similarity measure

We want to maximize the user acceptation probability of rendering sections. We use the fact that any decision tree learning algorithm chooses the attribute to split at any level, according to the information gain it can obtain from this attribute (IG(T|A$_i$))[14]. According to [18], information gain of a given attribute can be applied to weight a distance measure. In our case, that would mean that if a rendering is marked as “rejected”, the attribute at the top level of the decision tree is statistically speaking more important in the rejected decision than others. For instance, according to figure 2, for a <braille, L1> rendering predicted as “rejected”, we have better chance (in probability terms) to have an accepted rendering if we change the modality and keep the level of detail rather than if we change the level of detail and keep the modality. As a result, we propose a weighted sum of nominal attribute distances as explained in [14] between the predicted rejected rendering section (S$_1$) and each candidate rendering section predicted as accepted (S$_2$).

$$\mathrm{Dis}(S_1, S_2) = \sum^{n}_{(i=1))} IG(T|A_i)*\mathrm{Diff}(X_i,Y_i)$$

$$\mathrm{Diff}(X_i,Y_i) = \begin{cases} 0 \text{ if } X_i = Y_i \text{ or } X_i \text{ has the same words and in the same order than } Y_i \\ 1 \text{ if } X_i \neq Y_i \text{ or } X_i,Y_i \text{ don't have the same words, or have the same words but in different order and } X_i,Y_i \text{ are nominal types} \\ | X_i-Y_i | \text{ if } X_i \text{ and } Y_i \text{ are numbers} \end{cases}$$

IG(T|A$_i$) is the information gain of A$_i$. X$_i$(S$_1$), Y$_i$(S$_2$) $\in$ Dom(A$_i$)

Information gain [14] is defined as: IG(T|A$_i$) = H(T) – H(T|A$_i$), being H(T) the information entropy of T (how predictable are the different values T can take) and H(T|A$_i$) the conditional entropy of T given A$_i$ (how predictable the different values of T are, given that we know the value of A$_i$). In decision trees, information gain gives an idea of how important is an attribute to predict the value of the target attribute.

In the case that two or more candidate renderings have the same distance with the rejected rendering section, the one with the greatest acceptation probability is chosen. If both have the same probability, the rendering who has the highest frequency value is chosen (i.e. it appeas more times). Finaly, if both have the same frequency value, one is chosen randomly.

In the example used through this section, the similarity measure gives the results shown in Table 4.

Table 4. Similarity measure results

Default section rendering	Assigned section rendering
Any history annotation presentation	None
Character,Speech,1	Character,Speech,3
Character,Speech,2	Character,Speech,3
Place,Speech,2	Place,Speech,1
Place,Speech,3	Place,Speech,1

4. EVALUATION

The previous mechanism has been evaluated through a simulation by facing blind user behaviors with enriched videos. Two scenarios, respectively with SCard=1 and SCard=2, were tested. Annotation presentations and associated parameters were randomly generated and different user models (UM-X) were prepared and tested. We learn the tree with 30 annotations ([0-30]), and we define the *acceptation rate* each 10 annotation presentations as the ratio between the number of accepted annotation presentations and 10. We perform tree relearning each 10 presentations and we noticed that acceptation rate was improving, quickly converging to 94% in average (cf. Table 5).

Table 5. Acceptation rates results

	[0-30]	[30,40]	[40,50]	[50-60]	[60-70]	AVG
UM 1	100%	100%	100%	100%	100%	100%
UM 2	60%	90%	90%	90%	90%	85%
UM 3	60%	80%	100%	100%	100%	90%
UM 4	100%	100%	100%	100%	100%	100%

5. RELATED WORK

HYPERVIDEO AND ACCESSIBILITY- For audiovisual documents, some approaches to make hypervideo accessible were already implemented: some projects in [2][8] use the adaptable approach, letting the user manipulate what he is going to see before the audiovisual document rendering starts. Hypervideo has also been used for deaf people to convey Sign Language on the Web [21]. In [16] an adaptive approach was implemented for video annotations in hypervideos using P2P technologies and interests groups. Adaptation was in this project performed in terms of group of interests, neglecting people with disabilities and personal preferences.

ADAPTIVE HYPERMEDIA SYSTEMS (AHS)- AHS study how an hypertext/hypermedia system can be adapted to a user through her interactions with the system, by modeling and inferring user characteristics like goals, preferences and knowledge [11][1]. As we attempt to develop an adaptive videos enrichment system based on user interactions and because it is a special case of a hypermedia system, AHS theory seems to be very interesting.

6. CONCLUSION / FUTURE WORK

Enriched videos, i.e. videos enriched with multimodal presentations of additional descriptions (i.e. annotations), can improve videos accessibility for disabled users, all the more if these descriptions and their presentations are user-relevant.

We suggest a formal description of the problem we tackle (i.e. the adaptation of presentations of annotations) and an adaptive mechanism based on decision trees for the adaptation of presentations of annotations during videos visualization.

As a proof of concept, a first evaluation of our proposition tends to confirm that the suggested adaptation mechanism fits our requirements. However, real user studies in real video viewing settings have to be conducted in order to evaluate the suggested adaptive mechanism in a more realistic situation. Moreover, the adaptation accuracy of presentations can still be improved and others machine learning techniques have to be evaluated.

7. ACKNOWLEDGMENTS

This paper was supported by the French Ministry of Industry (*Innovative Web call*) under contract 09.2.93.0966, "Collaborative Annotation for Video Accessibility" (ACAV).

8. REFERENCES

[1] P. Brusilovksy. Methods and Techinques of adaptive hypermedia. *User modeling and User Adapted interaction, 6*(2-3):87-129, 1996.

[2] D.C.A. Bulterman. User centered abstractions for adaptive hypermedia presentations. In *Proceedings of the sixth ACM international conference on multimedia,* 247-256, 1998.

[3] S. Burgstahler. Creating video and multimedia products that are accessible to people with sensory impairments. DO-IT .University of Washington, 2008.

[4] D.C.A. Bulterman. L. Rutledge, L. Hardman, and J. van Ossenbruggen. Supporting adaptive and adaptable hypermedia presentation semantics. In *the Working Conference on database semantics (DS-8)*:5-8, 1999.

[5] T. Tran, P. Cimiano, and A. Ankolekar. A rule based adaptation model for ontology based personalization. *Series studies in computational Intelligence*, Springer, 2007.

[6] F. Tarpin–Bernard, and H. Habieb-Mammar. Modeling elementary cognitive habilities for adaptive hypermedia presentation. *User modeling and user adapted interaction*, 5(15):459-495, 2005.

[7] G. Paliouras, V. Karkaletsis, C. Papatheodorou, and C.D Spyropoulos. Exploiting learning techniques for the acquisition of user stereotypes and communities. In *proceedings of the seventh international conference on User Modeling*, 169-178, 1999.

[8] A non profit organization to make visual media accessible for the blind and partially sighted. http://www.hoerfilm.de/.

[9] J. Turner, and E. Colinet. Using audio description for indexing movie images. *Knowledge organization 31(4)*:222-230, 2007.

[10] M. Stoltze, and M. Ströbel. Utility based decision tree optimization: A framework for adaptive interviewing. In *proceedings of the 8th international conference on User Modeling,*105-116, 2007.

[11] P. Brusilovksy, and E. Millan. User models for adaptive hypermedia and adaptive educational systems. *In The Adaptive Web, LNCS 4321:3-53,* Springer, 2007.

[12] R.E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003.

[13] O. Aubert, and Y. Prié. Advene: an open-source framework for integrating and visualizing audiovisual metadata. *Proceedings of the 15th international conference on multimedia,* 1005-1008, 2007.

[14] P. N. Tan, M. Steinbach, and V. Kumar. Introduction to data mining. Addison Wesley Longman Publishing , Boston , 2005.

[15] P-A. Champin, B. Encelle, N.W.D. Evans, M Ollagnier-Beldame, Y. Prié, and R. Troncy. Towards Collaborative Annotation for Video Accessibility. In *7th International Cross-Disciplinary Conference on Web Accessibility (W4A 2010)*, Raleigh, USA. 2010.

[16] R. Fagá Jr, B. C. Furtado, F. Maximino, R. G Cattelan, and M. de Graca Pimentel. Context information exchange and sharing in a peer-to-peer community: a video annotation scenario. In *proceedings of the 27th ACM international conference on Design of communication* (2009).

[17] W3C, WAI. Providing audio that describes the important video content and describing it as such.

[18] H. Nuñez, M. Sanchez Marré, and U. Cortés. Improving similarity assessment with entropy based local weighting. In *Lecture Notes in Computer Science* (LNCS), Volume 2689/2003, Springer, 2003.

[19] Y Cao, and A. Nijholt. Modality planning for preventing tunnel vision in crisis management. In: *Symposium on Multimodal Output Generation (MOG 2008) at the AISB 2008 Convention "Communication, Interaction and Social Intelligence"*, 6-9, 2008.

[20] E. Frias-Martinez, S. Y. Chen, and X. Liu. Survey of data mining approaches to user modeling for adaptive hypermedia. In *IEEE Transactions on Systems, Man and cybernetics*, 36(6):734-749, 2006.

[21] DI. Fels, J. Richard, J. Hardman, and DG. Lee. Sign language Web pages. *American Annals of the Deaf*, 151(4):423–433, 2006.