VOX POPULI: Automatic Generation of Biased Video Sequences

Stefano Bocconi
CWI
P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Firstname.Lastname@cwi.nl

Frank Nack
CWI
P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Firstname.Lastname@cwi.nl

ABSTRACT

We describe our experimental rhetoric engine *Vox Populi* that generates biased video-sequences from a repository of video interviews and other related audio-visual web sources. Users are thus able to explore their own opinions on controversial topics covered by the repository. The repository contains interviews with United States residents stating their opinion on the events occurring after the terrorist attack on the United States on the 11th of September 2001. We present a model for biased documentary statements, such as interviews, and explain in detail how this model facilitates the automatic generation of rhetorical arguments on a micro-level. We outline the required representations of relevant rhetorical structures and the way they can be processed. The processes are described via examples generated by our experimental engine. The first example shows how to logically counter an opinion using semantics contained in the audio tracks from the database, while the second example describes the generation of an emotional counterargument using visual material.

Categories and Subject Descriptors

H.5.4 [**Hypertext,Hypermedia**]: Architectures, Navigation, User issues; I.7.2 [**Document Preparation**]: Hypertext/hypermedia, Multi/mixed media

General Terms

Design, Experimentation, Human Factors

Keywords

Media semantics, media rhetorics, automated video editing, multimedia presentation generation, video documentaries

1. INTRODUCTION

With production equipment becoming smaller, better and cheaper, video presentations, such as documentaries, can now be produced by semi- and non-professionals. Additionally, new ways of distribution, such as the Internet, increase the trend towards audio-visual information supply.

An example of the above trend is the work by IWA, a group of independent and non-professional filmmakers. They present their work to a wider audience, for example in form of documentaries, such as *Interview with America*, that presents interviews with United States residents on the events happening after the terrorist attack on the 11th of September 2001 (www.interviewwithamerica.com). The documentary is not a simple presentation of the interviews themselves but rather an artful and subjective construct by the filmmakers that has its own inherent logic, dynamics and enunciation and thus not necessarily be a representative sample of the recorded, and thus potentially available, opinions.

The filmmakers now wish to make the complete data available in combination with other web material related to the events and aftermath of 9/11. The aim is not only to inform the user by providing an objective presentation of information (e.g. on how many people support the decision to attack Afghanistan and why), but also to challenge the visitor of the site, who most likely has an own opinion about this controversial historical event. The challenge is to generate rhetoric statements that provide the requested information but in a form that facilitates users to explore their own opinions.

In this paper we illustrate how our experimental rhetoric engine *Vox Populi* utilizes the IWA repository of video interviews and other related audio-visual web sources to generate biased video sequences.

We base our approach on existing research that adopts the documentary form to automatically present media content relevant to the information needs of the user (see [8, 11, 18]). Here the aim is twofold: for the user, to allow an automatically-guided visual navigation of the content; for the author, to provide a framework for gathering content and making it available without having to specify explicitly how, and in what order, the user should view the material. The research focus of these projects was, therefore, to provide utilization of the syntactic aspects of the material for its automated presentation in the documentary form.

While structure is also important for our approach, we lay more emphasis on the semantics of the material. As the aim is to present biased arguments, the system has to be in the position to understand different views and, if required, to strengthen or weaken a point of view depending on the user's requests. The aim of our work is to model this manipulation mechanism and to define a minimal set of semantic annotations for the audio-visual material. We provide insights in how to achieve contextualized integration of video material into a presentation. We are inspired by work on automatic generation of audio-visual material in general, as provided by [16, 7, 9, 14].

We first present our model for biased documentary. To generate biased documentary sequences automatically, we have to construct a rhetorical argument, that works both on a micro-level (the argument is conclusive in itself) as well as on a macro-level (the argument is conclusive in a chain of arguments over time). In this paper we focus on the generation of media-based arguments on a micro-level to explain the required representations of relevant rhetorical structures and the way they can be processed. The processes are described via examples generated by our experimental system. The first example shows how to logically counter an opinion using semantics contained in the audio tracks from the database, while the second example describes the generation of an emotional counterargument using visual material. We evaluate our approach and conclude with a section of future work.

2. DOCUMENTARY STRUCTURES AND ELEMENTS

A documentary is the creative treatment of actuality [19]. This means that the documentary form is a subjective construct of mixed historical or contemporary footage with interviews, about actual people or events.

It is impossible to come up with a formula that fits all documentaries (for a detailed descriptions of styles and forms see [19],pp. 3 - 36, and pp. 315 -366). There are, however, a number of concepts that are central to this presentation style. We briefly describe these and determine which are of interest for our work on the automated generation of biased video sequences.

2.1 Documentary and organizational structure

The basis of a documentary is, similar to its fictional equivalents, a story. The story imposes an order, which demonstrates a cause and effect relation between events structured around an underlying point of view.

Imposing the order can be achieved in various forms. We are interested in the form that persuades the audience to adopt an opinion. This is termed as rhetorical documentary [2] and has the following goals:

1. it addresses the viewer openly, trying to move her or him to a new intellectual conviction, to a new emotional attitude, or to an action;
2. the subject is not a well established truth but a matter of opinion; the documentary is controversial;
3. the filmmaker often appeals to our emotions, rather than presenting only factual information;
4. the film will often attempt to persuade the viewer to make a choice that will have an effect on his or her everyday life (see [2], p. 122).

Of the five canons of rhetorics we focus on the first, namely *invention*(according to [6][1]). The invention defines the five demands of a rhetorical situation, namely:

1. The audience and their needs/desires/thoughts regarding the situation;
2. Types of evidence (facts, testimony, statistics, laws, maxims, examples, authority) to employ with the particular audience;
3. Appeal to the audience (logic, emotions, character);
4. Topics to examine the situation and generate ideas;
5. Timing and proportion for communication;

As our approach addresses a rhetoric situation that persuades, the relevant rhetoric forms are:

1. **Logos** appeals to logic or reason. There are a number of rhetorical figures that support subject-centered arguments (such as enthymeme or syllogism, see [4]).
2. **Ethos** appeals to the reputation of the author or character.
3. **Pathos** appeals to the emotions of the audience, including: fear, sadness, contentment, joy, pride. Pathos does not concern the truthfulness of the argument, only its appeal.

Though *Vox Populi* covers all five demands, we concentrate on exploring logos and pathos as the basis for the generation of rhetoric arguments. Logos is of importance as it provides a means of establishing a structured way of argumentation. Pathos, on the other hand, facilitates the creation of emphasis within an argument. During our discussion we also indicate where ethos can play a role.

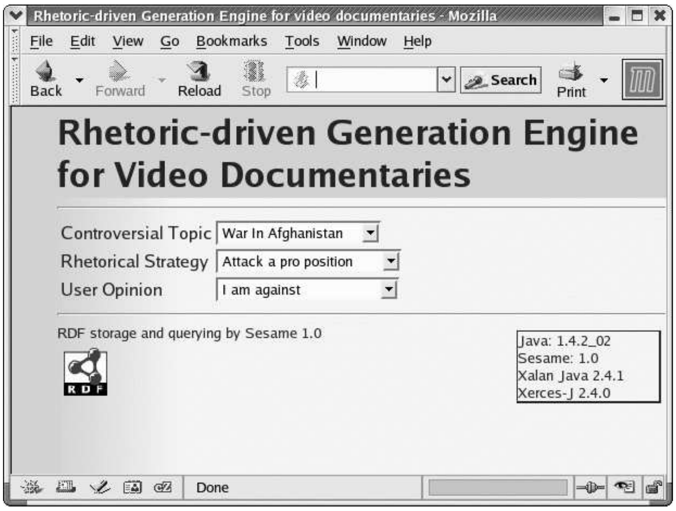
2.2 Documentary and point of view

As stated earlier, a documentary is a subjective construct that represents a personal interpretation of an event. This interpretation is the point of view. There are distinct categories that can be applied, which are not pure but incorporate each other (see [19], p.323).

1. **Single point of view** the argument or chain of arguments is channeled through by one character.
2. **Multiple points of view** represents various viewpoints, of which none predominates. This is well suited to expose cause and effect in an interdependent group, such as a class of society.
3. **Omniscient point of view** represents a collective rather than personal vision.
4. **Reflexive point of view** tries to present the material as a coherent whole, where the viewer is made aware of relationships.

In an automated system the user's request triggers the point of view of the presentation. For the ongoing discussion in this paper we are mainly interested in the *single* and *multiple points of view* as they cover the essential elements to establish the choice of a rhetorical strategy.

[1] see `http://www.rhetorica.net/textbook/invention.htm`



Rhetoric-driven Generation Engine for video documentaries - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop Search Print
Done

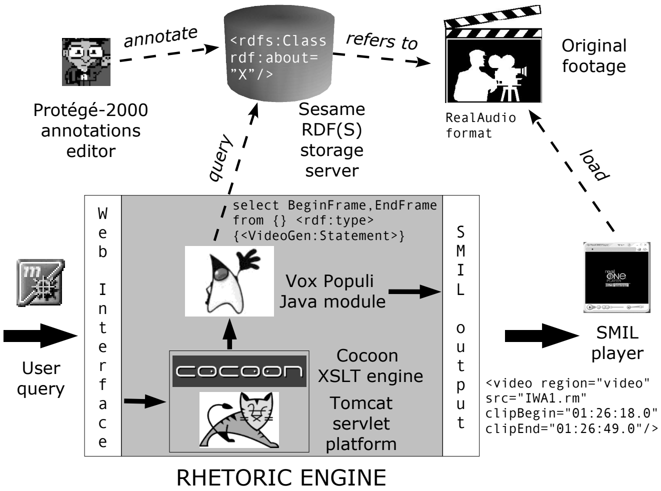
Rhetoric-driven Generation Engine for Video Documentaries

Controversial Topic: War In Afghanistan
Rhetorical Strategy: Attack a pro position
User Opinion: I am against

RDF storage and querying by Sesame 1.0

Java: 1.4.2_02
Sesame: 1.0
Xalan Java 2.4.1
Xerces-J 2.4.0

Figure 1: The interface to access the *Vox Populi* engine



```
<rdfs:Class
rdf:about=
"X"/>
```

```
select BeginFrame,EndFrame
from {} <rdf:type>
{<VideoGen:Statement>}
```

```
<video region="video"
src="IWA1.rm"
clipBegin="01:26:18.0"
clipEnd="01:26:49.0"/>
```

Figure 2: The *Vox Populi* engine and its supporting framework

2.3 Documentary and range of form

The rhetorics used for the generation of an argument are applied to existing material, that essentially can be separated into two forms. One sort of content contains in itself an argument, such as sequences where people talk to each other; or interviews where people answer formal, structured questions. The other type of content is action footage, still photos, graphics, sound effects or music that only in combination with other items of this type establish an argument.

The material we are interested in is video and single images. It needs to be made clear, though, that we understand a video as the combination of two separate streams, namely the visual and audio stream. This distinction is important as the different rhetoric forms focus on different media. For example, the rhetorical form of *logos* mainly supports the continuity of an argument and thus addresses the main media continuity supporter, in the case of interviews the audio track.

Having introduced the key constituents for our approach toward the automatic generation of biased video statements we are now in the position to describe the underlying processes and structures in more detail.

3. SCENARIO AND ARCHITECTURE

We base our discussion on opinion generation using material from the IWA database, which contains 8 hours of video footage, mostly interviews with people of different socioeconomic groups and some location material.

3.1 Scenario

For the sake of clarity we explain the opinion generation process for a user who wishes to see an argument about the war in Afghanistan (see the topic in Figure 1). Note, the current interface was built to allow us to experiment with the underlying rhetoric engine. Issues of interface design have not yet been considered.

Additionally, the user can state the rhetoric strategy to be used. The range of possible requests goes from simple, such as "Show all opposing statements", where a line of opposing statements is created, to complex requests, such as "Attack a pro position"(see the rhetoric strategy in Figure 1). In our scenario we describe the complex requests because its rhetorical mechanisms include those used for easier cases.

Finally, the user can provide an own opinion, such as "I am against", "I am in favour", or "I am undecided" (see the user opinion in Figure 1). The user opinion is important as the aim of the system is not only to provide information that is correct but to challenge the viewer's point of view. As the system cannot yet rely on a detailed user model, this is the way to retrieve basic but essential information from the user.

Our scenario query is, thus, that the system should provide a video statement about the war in Afghanistan, clearly providing an opposing opinion, which is shared by the user, as portrayed in Figure 1.

The outcome of the query is a generated video sequence, which takes either a multiple or single point of view. The multiple point of view is chosen if the engine can create a chain of arguments based on available interviews that attack the favour statement. The single point of view is chosen if the system cannot find appropriate material and thus has to generate the argument itself. Both examples will be explained in the discussion of section 4.

3.2 Implementation

The architecture in which the *Vox Populi* engine works is described in Figure 2.

The user is presented with a Web interface that is generated by Cocoon[2], a web development framework that allows XML transformation via an XSLT [5] engine and HTML serialization. Cocoon runs on top of Tomcat[3], a servlet container, and calls *Vox Populi* functionality to handle the user request. *Vox Populi* is implemented in Java and queries (via HTTP) Sesame [3], an RDFS [22] storage server, containing all the RDFS-encoded metadata about the video material. We use Protege-2000 [10] to create annotations and the query language SeRQL [1] to query the repository. *Vox Populi* output is encoded in SMIL-2 [21] and is presented

[2] http://cocoon.apache.org/

[3] http://jakarta.apache.org/tomcat/

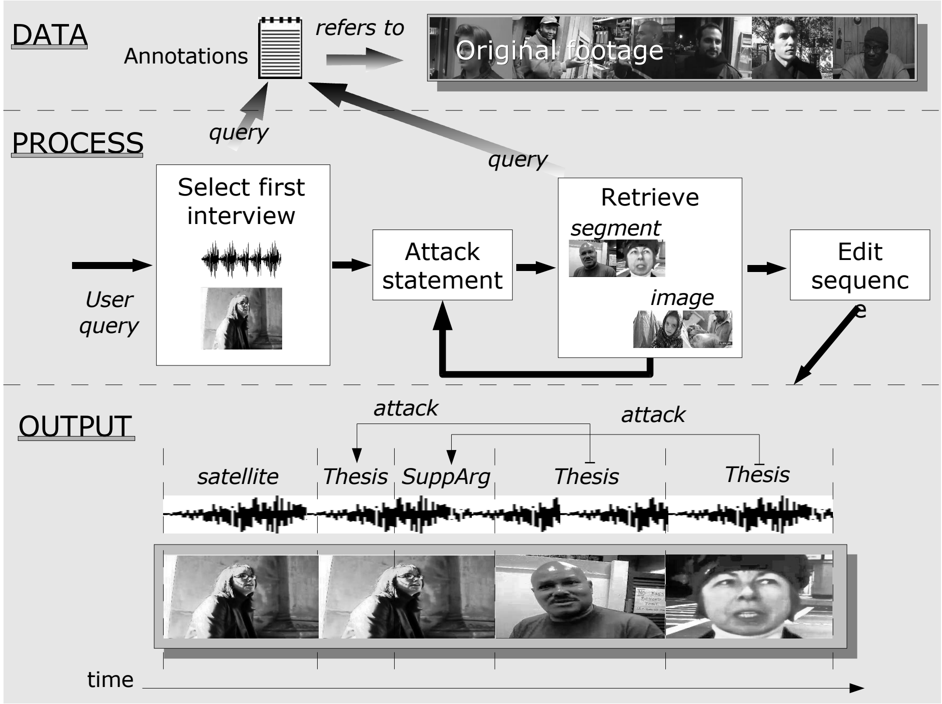


Figure 3: Multiple point of view based opinion generation process

in a SMIL-2 player, which is in our environment RealONE from RealNetworks.

4. BIASED SEQUENCE GENERATION

As *Vox Populi* works with interviews, the engine first addresses the audio material available in the IWA database. It picks, therefore, from the list of ordered *points of view* the highest, and thus most complex, one associated with the required rhetoric goal, namely *attack-pro*. The system tries the complex structure first because the finally generated statement should be a challenge for the viewer. Note, this approach from more complex to less complex strategies is always applied by the engine.

In cases where it cannot generate the required argument the engine swaps to the next less complex strategy until it reaches the point where it cannot find any suitable solution with the material available in the repository. It then tries to generate a biased statement by itself. Note, on its way to the final argument the engine might shift between different rhetoric forms and thus also shift focus on the media representations it addresses.

We now describe the generation of a multiple point of view based counter argument and outline in section 4.2 how the system generates a single point of view opinion.

4.1 Multiple point of view based generation

The task of the engine at this stage is to generate a particular biased video sequence on a micro-discourse level. This means that the engine first has to find one interview that it can attack according to the user request. This process is subdivided into 4 stages, namely:

1. deciding about an interview suitable to be attacked,
2. establishing the rhetoric elements to attack,
3. retrieving the required material,
4. editing the final video sequence.

The relation between the 4 process stages and their integration in the general framework, as described in section 3, is portrayed in Figure 3. Each stage is described in more detail in the following sections. Our discussion is based on the settings in Figure 3, namely *Topic* (War in Afghanistan), *Rhetoric strategy* (Attack a pro position), and *User opinion* (Against).

4.1.1 Interview Selection

As stated previously, the first action for the engine is to find a suitable target interview sequence. To achieve this *Vox Populi* draws on all three request parameters.

It uses the topic to search for appropriate material on the semantic level. This means that all interviews annotated with the concept *War Afghanistan* become relevant. Additionally, it applies the *Rhetoric Strategy* to retrieve an interview that is annotated as *pro* topic. Finally, the engine utilizes the user's opinion to establish the start interview.

Since the goal of the engine is to challenge the viewer's opinion, it tries to identify an interview that not only provides arguments for the war but also anti-war statements. For this, the engine uses constructs from Rhetorical Structure Theory (RST [13]). RST facilitates the analysis of texts spans based on the function they perform, such as antithesis, evidence, condition, or concession. For example, in a *concession* a speaker provides a proposition about her own beliefs and additionally a set of propositions that contradict these beliefs. The aim is to acknowledge the incompatibility between the propositions presented, but make the listener recognize the propositions as compatible and to respect these contradicting propositions.

Since the engine is looking for a combination of pro and con statements, an interview statement annotated as *concession* is a suitable choice.

In our scenario the engine selects the following interview with a young well-educated woman in the garden of Harvard, Cambridge, because it introduces a concession that provides the highest number of propositions (3), among those interviews that are pro war (see Figure 3). Note, each statement in the interview is annotated to indicate its RST role.

INTERVIEW 1. *" I am never a fan of military action [Satellite-Thesis]. In the big picture I don't think it is ever a good thing [Satellite-Supporting Argument1], but I think there are circumstances in which I certainly can't think of a more effective way to counter this sort of things [Nucleus-thesis]. I suppose there is a point in which certain people play by certain rules and you have to go to their level [Nucleus-Supporting-Argument1]. I do not think there is any way to resolve this conflict diplomatically [Nucleus-Supporting-Argument2]."*

4.1.2 Rhetoric Analysis

Once the target interview has been identified, the engine has to establish the rhetoric elements that can be attacked. The engine uses annotations that describe the rhetoric aspects of the audio track of the target interview. An annotation is composed out of a claim the interviewee is making (the thesis) and the arguments the interviewee uses to support the thesis (the supporting arguments). Both *Nucleus* and *Satellite* can have a thesis and supporting arguments.

The building block of a thesis or of a supporting argument is the *Statement*, which is encoded within *Vox Populi* as an RDFS class with the following properties:

- *statement identifier*
- *statement role*
- *subject* (e.g. war)
- *modifier* (e.g. best)
- *predicate* (e.g. solution)
- *object* (e.g. terrorism)
- *fileLocator* containing the physical location of the video footage with the statement
- *beginFrame* time stamp within the filename of the start of the statement
- *endFrame* time stamp within the filename of the end of the statement
- *mediaProperties* describing features of the media, such as framing or colour for visuals, or pitch for audio.

The first six properties encode the rhetoric semantics of the statement, while the last four properties are used to physically locate the footage and to facilitate the editing stage of the opinion generation process (see section 4.1.3).

In order to attack the interview, the engine first identifies the relevant statements within the target interview, namely those of the *Nucleus*. The *Satellite* is not touched as it provides the same opinion as the one of the viewer, namely being against war.

Once the targets, thesis and supporting arguments, are identified, the engine tries to collect for each target statement another interview containing a statement in disagreement. Disagreement is understood here in terms of difference between either the subject, the modifier, the predicate or the object of a statement.

The instantiation of a statement is evaluated by instances of a special *Concept* class that has the following properties:

- *logicalSimilar* (e.g. war has *logicalSimilar* military actions)
- *logicalOpposite* (e.g. war has *logicalOpposite* diplomacy)
- *negativeAspect* (e.g. war has *negativeAspect* victims)
- *positiveAspect* (e.g. war has *positiveAspect* liberation)

In our scenario the engine makes use of the *logicalOpposite* property to establish the concept to be retrieved. The basis for this process is rule that instantiates the triple [targetConcept, logicalOpposite, differenceConcept]. Here the targetConcept represents the value *war* of the currently addressed property of the target statement, such as *Subject*. *LogicalOpposite* is the parameter that determines which concept to look for. *DifferenceConcept*, finally, is instantiated with the matching concept, such as *diplomacy*. The differenceConcepts for subject, predicate and object form the set to search for the appropriate material in the annotation repository.

The engine must also be able to differentiate between concepts. Thus we introduced an additional class that facilitates such a differentiation based on the modifiers, such as *good* and its opposite *not_good*, only and its opposite *not_only*, or *same* and its opposite *not_same*.

Thus, modifiers not only allow the establishment of opposition (as good and not_good), but also facilitate the gradual differentiation between disagreements, as in "war best solution", "war not_only solution", "war bad solution" and "war worst solution". This allows building up an argument gradually from supporting a thesis to contradicting it. Thus, the engine can now generate a hierarchical set of search statements that all attack the thesis of the target statement [war best solution]. The search set might take the form: [diplomacy best solution, war bad solution, war worst solution, war not solution], where the order determines the importance (the first element of the list is the most important).

4.1.3 Retrieval and Presentation

The retrieval of relevant interviews is now a matter of matches between the various options in the search set and the available rhetoric-based annotations of the audio tracks. Assume that the retrieval process found two suitable interviews.

Interview 2 features an African-American man, who owns a shop in Stanford. This interview was retrieved due to the annotation of its thesis as *subject:* war; *modifier:* not; *predicate:* solution.

INTERVIEW 2. *"War has never solved anything. What is it going to solve? What will it achieve? I cannot achieve anything. You have to sit down and decide what we are going to do about this, who can I talk to, who will listen."*

Interview 3, portraying a white woman on a street of Boston, was retrieved because its thesis was annotated as *subject:* violence; *modifier:* not; *predicate:* solution.

INTERVIEW 3. *"Every American and every other person has to remember that if we participate in their hate, that we are doing the same thing they are, we cannot allow our souls to be taken over by the kind of hate that is taking over the souls of the people who have been doing these actions and if we can somehow stop the hate, that's what we have to do, and I do not know how."*

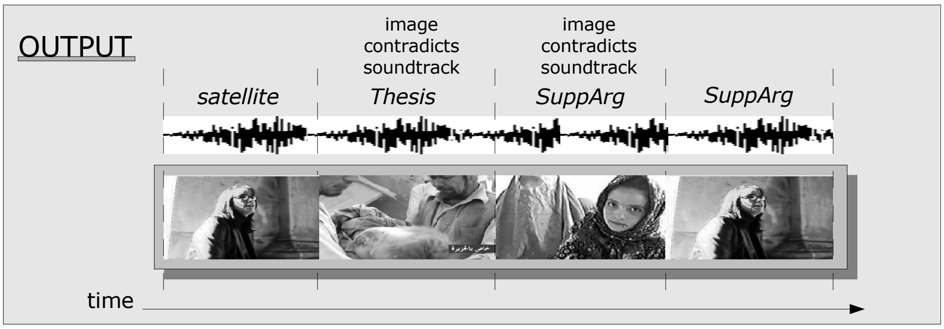


Figure 4: Output generated by a single point of view process

At this stage the material for the final presentation is collected. The editing unit of the *Vox Populi* engine has to determine which statements are finally used and in which order they are presented.

In section 4.1.2 we identified the relevant thesis and supporting arguments that need to be attacked. The retrieved results are grouped based on their relation to the associated argument from the target interview they were retrieved for. In case there is more than one statement per target, the system uses the *mediaProperties* of the statement class to decide which media item to make use of. In our scenario, for example, the editing unit compares the framing of the target video with that of the attacking statement. The framing of a video provides the viewer with an idea of importance of the image. A wider shot gives an overview whereas a close-up provides emphasis. Our engine uses the hierarchical model of shot ratios, developed in [16], to allow the engine to compare shot ratios and select the one for the attack statement. For example, the framing of the target video is described as *medium*. As the aim is to differentiate the attack, the engine prefers those statements that have a framing of type *close-up*. In our scenario this rule is not really applicable as the engine only retrieved one attack sequence for each target statement.

It still remains the problem of ordering the selected attack statements. At the moment the engine goes for clarity first. This means the engine presents the segments attacking the thesis statements and then those that attack the supporting arguments (see output part in Figure 3). The reason is that the most direct attack should be displayed first to have a bigger impact. Another option is to have a slower build-up of the argumentation, first attacking the supporting arguments and then giving the *coup de grace* to the thesis. Both options can be generated by our system, but as yet we have no mechanism to choose between the two.

As mentioned in section 3, the system's output format is SMIL. We chose this language because it support the needs of automatic editing as described in this paper: a SMIL player can play specific video segments within a file specifying with clipBegin and clipEnd the segment's position. Another useful feature is the possibility to insert cross transitions between different video segments, making the editing process easier and smoother. Here is an example showing the two above-mentioned features:

```
<video region="video" src="IWA1.rm"
 clipBegin="01:26:18.0" clipEnd="01:26:49.0"
 transIn="trans" transOut="trans" fill="transition"/>
```

The output of the engine is portrayed in the lower part of Figure 3.

4.2 Single point of view based generation

The engine cannot always retrieve the required material from the interview database. In these cases it has to come up with a solution where it attacks or supports a selected statement itself.

Assume the engine is not be able to retrieve suitable interviews, after it has performed the steps *Interview selection* (see section 4.1.1) and *Rhetoric analysis* (see section 4.1.2). The engine marks *audio* as not valid for manipulation and then backtracks within the ordered list of *points of view* until it finds one that allows for a presentation using a different media, such as the *single point of view* strategy set, that supports the manipulation on an audio as well as visual level.

The engine still applies the rhetoric goal "attack a pro statement", and makes use of the rhetoric analysis of the target interview. The aim is now to establish the material to be retrieved. As the engine cannot use audio material, it sets the target media to visual, which includes image and video.

In section 4.1.2 we introduced the *Concept class*, which allows the system to identify disagreement or agreement. The relevant properties the engine utilizes now are *negativeAspect* and *positiveAspect*.

These properties provide relations that facilitate the engine to build the search query for relevant material. In every statement to be attacked in the target interview the engine isolates the *subject* attribute to identify the concept to attack. In our scenario these are *war*, *diplomacy* and *violence* (see analysis of Interview 1).

It then uses the *modifier* to establish which of the two properties in the *Concept class* to address. Take the thesis of the Nucleus of Interview 1 as an example. Here the *best* modifier indicates that the statement is positive with respect to the subject. As the engine has to attack the statement, it goes for the *negativeAspect* property, resulting in the selection of *victims* and *destruction* as the concepts for which the engine tries to find visual material, either in its own database or on other repositories, such as the web.

The results, such as the photos of a dead Arab baby and of a suffering Arab girl portrayed in Figure 4 need now to be included in the target video clip.

The engine first establishes that the main character of the interview sequence, i.e. the interviewee, can be seen long enough so that the viewer knows who is talking. For that the engine uses rough estimates based on the framing of a shot. In our scenario example, the shot framing is described in the *Statement class* instantiation as (*mediaProperties:framing* = medium) for which a viewer usually needs around 3 seconds to perceive the main details. In our scenario sequence, the first four seconds are covered by the *Satellite* part of the concession. The satellite is, however, not to be touched at all, and thus the engine can process with the next step to replace existing video material with the newly collected images.

As there is only one image for the thesis and one supporting argument, the positioning is already decided upon. The engine does not change the order of arguments as it keeps the audio track intact. Yet, the duration of the images needs to be adjusted. In our scenario the engine uses the properties *beginFrame* and *endFrame* of the *Statement class* to calculate the presentation time for each image. Having established that, the engine generates a SMIL file, in which transitions guarantee the right swap of visual material, applying the superimposing feature of SMIL, as described below:

```
<img region="video" src="girl.png" begin="17s"
 dur="13s" transIn="trans" transOut="trans"
 "fill="hold" />
```

The output of the engine is portrayed in Figure 4.

5. DISCUSSION AND CONCLUSION

The described rhetoric processes, which aim to use minimal semantic annotations, demonstrate the feasibility of our approach. The prototype engine is in the position to generate, depending on the user request, acceptable biased statements in various rhetoric forms (see also our test page at `http://homepages.cwi.nl/~media/demo/IWA`). The current engine, however, needs further fine-tuning.

Further work is required to establish a wider range of rhetoric forms for the micro-level of the presentation. The current engine mainly addresses structures of one rhetoric canon, namely invention. Here it covers, however, all five demands of a rhetorical situation (see section 2). In this respect our work is similar to Terminal Time [14], which applies an equivalent canon to generate cinematic experiences for mass audiences. The major differences between the two approaches are two-fold. First, in our engine the rhetoric rules for generating the argument are made explicit and are not embedded in the material organisation. Second, Terminal Time is, as other knowledge intensive but closed AI applications from the nineties, content driven, where our approach is structure oriented. The aim of our engine is to apply access to material based on the connections between ideas, where the connections are grounded in a discourse/argumentation ontology - a strategy also used in hypermedia discourse modeling [20]. We go a step further, though, as we do not apply this technique to present an existing, although complex, discourse but to generate a biased argument on-the-fly.

The price we have to pay for the flexibility we gain in the generation process is a loss in reliability of material use. In Terminal Time the material is especially created and thus complete control about the content and its combinations can be provided, especially because Terminal Time also applies a restricted set of questions the audience can answer collectively. As our engine at the moment mainly generates biased statements on the basis of rhetoric structures it can happen that, depending on the viewer, the content (both on an audio and visual level) of the generated statement can be either unqualified or offensive, which clearly damages the statement if it was intended intended to support the target statement. To avoid such mis-generations it is necessary to introduce some sort of high-level reliability measures that facilitate the use of the material in various contexts. One option to establish a reliability measure for the material is to provide a model that determines the social status of the speaker (e.g. education, age, gender, race) and the correctness of statements within the ranges of a particular culture and can set this in relation to the corresponding views of the user. Exploring these description and processable complexities is part of our ongoing research.

As our engine applies its rhetoric rules on various media we also have to improve the descriptions of various media with respect to their use in a rhetoric context. At the moment we relate structure, thus logos, with the medium that provides the continuity, which is in the case of interviews the sound. Moreover, we associate emotions and images (pathos). Yet, the interplay between rhetoric forms and related media require a more subtle model, in particular if we look at the generation of macro-structures. Our engine can only perform basic linear sequencing techniques. Far more interesting is to provide means that facilitate the intercutting of arguments. For example, the sequence of our first example described in section 4.1 could also be generated so that the attacking statements follow the target arguments straight away.

As the final goal for our engine is to generate an evolving discourse, such as a discussion with a user over a controversial topic in form of a Socratic tutor, we will follow the approach of progression of detail that facilitates navigation based on a given weighted set of descriptors representing a story context on a micro-level (next step in content exploration) as well as on a macro-level (larger contextual units clustering content), as described in [8, 11]. Further research is needed to determine the flexibility of the generic microstructures generated by our engine to facilitate macro structures.

One strength of our engine is that it can manipulate audiovisual material on a physical level. Here it works similarly to the system described in [16], which assembles random video segments and edits them fully automatically to generate slapstick content. Our engine uses a number of the rules from that work. We improved the work, however, as we introduced mechanisms that also address audio within the editing process.

The ability of physical material manipulation is not unproblematic. At the moment the engine performs these tasks but the viewer cannot see that the material is manipulated. Here, we have to investigate visualisation mechanisms that achieve this task without destroying the feel of the documentary. In future work we intend to use Berthold Brecht's defamiliarization effect. This mechanism used in the epic theater, establishes a distance between the viewer and the presented material and thus facilitates the viewer to reflect about the intended meaning to be communicated.

All solutions to the described problems, however, require that the engine has access to high quality, though not necessarily excessive, annotations of the media units. Most annotations described in this paper can be provided during the production of the material (see [17, 12, 9]). Yet, a substantial part of the annotations need to be provided after the material is established, such as the rhetorical annotations. At the moment the engine uses a basic rhetoric ontology that adopts elements from RST [13]. Moreover, the engine uses a link to Wordnet [15] to find synonyms and antonyms, to extend the search space.

The tools used in our prototype to provide the rhetoric annotations were used by the IWA group to annotate their material. At the beginning the group would need 10 times the time of the statement to annotate. Once they got used to the annotation environment they reduced the annotation time to five times the duration of the statement. For the statement of Interview 1 this means that it takes roughly 5 minutes to annotate it. Nevertheless, we still have to investigate better ways to make the material available.

6. ACKNOWLEDGMENTS

This research was funded by the Dutch national ToKeN2000 I2RP and CHIME projects. The authors are grateful to the IWA team for their permission to use the material from *Interview with America* for this paper. The authors also wish to thank in particular their colleagues Lynda Hardman, Jacco van Ossenbruggen, Lloyd Rutledge, Yulia Bachvarova and Joost Geurts for useful discussions during the development of this work.

7. REFERENCES

[1] Aidministrator Nederland B.V. *SeRQL user manual*, April 4, 2003.

[2] D. Bordwell and K. Thompson. *Film Art: An Introduction*. McGraw-Hill, 7 edition, 2003.

[3] J. Broekstra, A. Kampman, and F. van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In I. Horrocks and J. Hendler, editors, *The Semantic Web - ISWC 2002*, number 2342 in Lecture Notes in Computer Science, pages 54–68, Berlin Heidelberg, 2002. Springer.

[4] G. Burton. Silva Rhetoricae. http://humanities.byu.edu/rhetoric/.

[5] J. Clark. XSL Transformations (XSLT) Version 1.0. W3C Recommendation, 16 November 1999.

[6] A. R. Cline. The Rhetorica Network. http://www.rhetorica.net/.

[7] M. Crampes, J. P. Veuillez, and S. Ranwez. Adaptive narrative abstraction. In *The Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, pages 97–105, Pittsburgh, PA, June 20-24, 1998. ACM, ACM Press. Edited by Kaj Grønbæck, Elli Mylonas and Frank M. Shipman III.

[8] G. Davenport and M. Murtaugh. ConText: Towards the Evolving Documentary. In *ACM Multimedia '95, Proceedings*, pages 377–378, November 1995.

[9] M. Davis. *Readings in Human-Computer Interaction: Toward the Year 2000*, chapter Media Streams: An Iconic Visual Language for Video Representation., pages 854–866. Morgan Kaufmann Publishers, Inc., 1995.

[10] W. Grosso, H. Eriksson, R. Fergerson, J. Gennari, S. Tu, and M. Musen. Knowledge Modeling at the Millennium (The Design and Evolution of Protege-2000). Technical Report SMI Report Number: SMI-1999-0801, Stanford Medical Informatics (SMI), 1999.

[11] G. Houbart. *Viewpoints on Demand: Tailoring the Presentation of Opinions in Video*. PhD thesis, Massachusetts Institute of Technology, 1994.

[12] IBM research. VideoAnnEx - IBM MPEG-7 Annotation Tool, 2002.

[13] W. C. Mann, C. M. I. M. Matthiesen, and S. A. Thompson. Rhetorical Structure Theory and Text Analysis. Technical Report ISI/RR-89-242, Information Sciences Institute, University of Southern California, November 1989.

[14] M. Mateas. Generation of Ideologically-Biased Historical Documentaries. In *Proceedings of AAAI 2000*, pages 36–42, July 2000.

[15] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An On-line Lexical Database. *Journal of Lexicography*, 3(4):234–244, 1990.

[16] F. Nack. *AUTEUR: The Application of Video Semantics and Theme Representation in Automated Video Editing*. PhD thesis, Lancaster University, 1996.

[17] F. Nack and W. Putz. Designing Annotation Before It's Needed. In *Proceedings of the 9th ACM International Conference on Multimedia*, pages 251–260, Ottawa, Ontario, Canada, September 30 - October 5, 2001.

[18] D. B. Nitin Sawhney and I. Smith. HyperCafe: Narrative and Aesthetic Properties of Hypervideo. In *Proc. of the Seventh ACM Conference on Hypertext*, pages 1–10, 1996.

[19] M. Rabiger. *Directing the Documentary*. Focal Press, 1998.

[20] S. B. Shum, E. Motta, and J. Domingue. ScholOnto: an Ontology-Based Digital Library Server for Research Documents and Discourse. *International Journal on Digital Libraries*, 3(3), August/September 2000.

[21] W3C. Synchronized Multimedia Integration Language (SMIL 2.0) Specification. W3C Recommendation, August 7, 2001. Edited by Aaron Cohen.

[22] W3C. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 10 February 2004. Edited by Dan Brickley and R.V. Guha.