Generating Media Stories – Play it again, Sam

Frank Nack and Stefano Bocconi
CWI, Amsterdam
Kruislaan 413, P.O. Box 94079
1090 GB Amsterdam, The Netherlands
E-mail: {firstname.lastname}@cwi.nl

ABSTRACT

New types of knowledge spaces, such as the web, allow for yet unexplored forms of knowledge exploration and social relationships. Such an interactive, open and multimodal system sustains the activation of articulation expressions that form the basis of adaptive discourses. In this paper we look at the problem of how stories in such environments can be established. We are in particular interested what story generation requires, in such a context, from authors and how that reflects on the accessibility of the information to users. The examples are taken form the domain of documentary making.

1. Introduction

A great deal of research has been directed towards the development of engines that seek to interpret, manipulate or generate audio-visual stories either in a semi-automatic, or automatic way [].

All these approaches focussed either on particular intrinsic aspects of a media that the authors wished to represent or the works concentrated on a particular process that can be performed on or with the investigated media. The problem with those approaches is that they are, as many AI applications from the nineties, knowledge intensive and closed systems.

More recently, however, it became apparent that a more holistic view is required as we are heading towards a cyberspace as described by William Gibson in his novel *Neuromancer* [11] and envisioned by McLuhan in his work *Understanding Media* [15].

Here the aim is to provide open knowledge environments that facilitates new forms of knowledge exploration and social relationships, mediated through communication networks. Such an interactive, open and multimodal system sustains the activation of articulation powers, which in general represent parts of a semiotic continuum, where verbal, gestical, musical, iconic, graphic, or sculptural expressions form the basis of adaptive discourses.

This type of environment requires from an authoring point of view that content can be gathered and be made available without having to specify the order explicitly. This means a shift in the author paradigm from the provider of linear material to a provider of explorable content. This requires, though, that the author can anticipate various ways of discourses each featuring various sorts of rhetorics. For the user of such an environment it is important that she can be automatically-guided through the content The challenge is to generate media statements that provide the requested information but in a form that facilitates users to explore their own opinions. These two aspects are what we address in our research.

2. Look back without anger

Our current work on Vox Populi [] is based 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] but mainly tries to overcome problems we face in older work.

The AUTEUR system [] is a planner based approach to the application of video semantics and theme representation to the automated editing of visual stories at the level of events. AUTEUR could fully automatically generates humorous, non-verbal video sequences drawing on film theory (in order to define methods to perform automated editing, to model the fundamental units of the image and the conceptual relationships between image, shot and sequence); on narrative and humour theory (to attempt to automatically generate emotion provoking and credible film narrative); and on Artificial Intelligence (planning, story generation, and knowledge representation). Though AUTEUR could handle arbitrary material it was too knowledge intensive and mainly performed only in a controlled working environment.

DISC [5], on the other hand, could make use of ontology-controlled multimedia repositories as well as domain ontologies to create to create multimedia presentations on request. The aim of the approach is to build a multimedia presentation about a certain topic by traversing a semantic graph. The semantic graph consists of domain ontology of classes, instances and relations between them together with the media material related to those instances. To create a discourse structure the system contains a set of rules defining

- what kind of genre can be applied to a certain main character
- what types of narrative units are relevant for a certain genre
- what types of characters can appear in what narrative unit
- what types of domain relationships are relevant for those characters.

The DISC architecture allows the development of a story inside a narrative unit since a number of related characters it describes depends on the information found in the semantic graph. The problem with the approach is that rules about what defines a genre, such as they exist in AUTEUR, do not exist explicitly. Moreover, no order is defined for organizing narrative units into a discourse structure and for defining appearance of related characters inside a narrative unit. Finally, DISC cannot handle situations where multiple media items can be annotated with the same concept.

Our current work tries to integrate both approaches, namely the genarting advances from AUTEUR with the ontology and semantic web based flexibility of DISC. The focus is to establish means thyat allow the reuse of media material. Reuse here means recreating a context in the presentation for different information items that were created for a different purpose, so that the presentation shows a coherent structure as human-authored ones do.

Vox Populi [] is a rhetoric-driven presentation engine that utilizes an audio-visual repository to automatically generate short video sequences that make a point and show argumentation progression. The repository used is provided by ‘Interview with America’, an initiative of a group of independent, non-professional filmmakers to present American people’s opinion on the events happening after the terrorist attack on the United States on the 11th of September 2001. The gathered material contains 8 hours of video footage, mostly interviews of people of different socioeconomic groups and some location material that highlights the interviewee’s identity.

Vox Populi utilizes two types of annotations: descriptive and rhetorical. The descriptive annotations cover the who, where, when and what in the video. The rhetoric annotations are based on the verbal information contained in the audio channel, identifying the claims the interviewees make and the argumentation structures they use to make those claims. To encode the argumentation structures we use the Toulmin Model [19].

Vox Populi generates meaningful video sequences by selecting and ordering video segments using rhetoric-based strategies, such as opposition and similarity. Those strategies traverse the graph of typed relations between video segments (the Semantic Graph) deriving video sequences from this structure. In our case the Semantic Graph is the product of the automatic link generation process using the annotation schema.

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.

3. Where from here

Further work is required to establish a wider range of rhetoric forms for the micro-level and macro-level of the presentation. Compared to an engine such as Terminal Time [], our engine provides the advantage of explicit rhetoric rules for generating the argument . Moreover, they are not embedded in the material organisation. Second, Terminal Time is 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.

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 AUTEUR [16],. 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.

4. ACKNOWLEDGMENTS

5. REFERENCES

[1] ACM Communications (2003). A game experience in every application. Communications of the ACM, July 2003, Volume 46, Number .

[2] Aguierre Smith, T. G., & Davenport, G. (1992). The Stratification System. A Design Environment for Random Access Video. In *ACM workshop on Networking and Operating System Support for Digital Audio and Video*. San Diego, California

[3] Brooks, K.M. (1999). "Metalinear Cinematic Narrative: Theory, Process, and Tool. " MIT Ph.D. Thesis.

[4] Chatman, S. (1978). *Story and Discourse: Narrative Structure in Fiction and Film*. New York: Ithaca.

[5] DAVENPORT, G., and MURTAUGH, M. ConText: Towards the Evolving Documentary. In ACM

[6] Davis, M. (1995) Media Streams: Representing Video for Retrieval and Repurposing. Ph.D., MIT.

[7] Geurts, J., Bocconi, S., Ossenbruggen, J.v. and Hardman, L. (2003). Towards Ontology-driven Discourse: From Semantic Graphs to Multimedia Presentations. Proceedings of the 2nd International Semantic Web Conference (ISWC2003). Sundial Resort, Sanibel Island, Florida, USA, 20-23 October 2003. http://iswc2003.semanticweb.org/

[8] Gibson, W. (1986). *Neuromancer*. Phantasia Press, 1st Phantasia Press ed. West Bloomfield

[9] McLuhan, M. (2001). Understanding Media. Routledge Classics, London and New York.

[10] Nack, F. (1996) "AUTEUR: The Application of Video Semantics and Theme Representation in Automated Video Editing," Ph.D., Lancaster University, 1996.

[11] Nack, F. & Hardman, L. (2001) Denotative and Connotative Semantics in Hypermedia: Proposal for a Semiotic-Aware Architecture. *The New Review of Hypermedia and Multimedia 2001, Vol. 7, pp. 39 - 65.*

[12] Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. New York: Cambridge University Press.

[13] Barry, B. (2003) The Mindful Camera: Common Sense for Documentary Videography. In Proceedings of ACM Multimedia Conference. Berkeley, California, USA

[14] Michael Mateas, "Generation of Ideologically-Biased Historical Documentaries," in *Proceedings of AAAI 2000*, July 2000, pp. 36–42.