Published at the Conference of Associazione Italiana Scienze della Voce (AISV) 2006.

*NEEDLE: Next gEneration sEearch engine for Digital LibrariEs

Giuseppe Riccardi and Marco Ronchetti
Department of Information and Communication Technologies, Università di Trento
Via Sommarive 14, 38050 Povo Italy
{giuseppe.ricccardi, marco.ronchetti}@unitn.it

1. ABSTRACT

Next-generation information access will take advantage of (un)structured multimedia databases. We want to address the problem of searching, presenting and evaluating multimedia content fruition in the context of e-learning domain application. In particular the use of navigable streamed digital video as a teaching and learning resource became in recent years an attractive option for many educators as an innovation which expands the range of learning resources available to students by moving away from static text-and-graphic resources towards a video-rich learning environment. Streamed video allows remote access to lectures and, when integrated into a multimedia package, creates a rich, accessible, interactive and controllable teaching resource.

Our research project is based on indexing streaming audio related with other kinds of resources (video, desktop activities recording, PowerPoint presentations, interactive whiteboard tracks etc.). The difficulty is to extract information from the multimedia and allow searching among them. The first problem is to index speech, video and metadata streams and exploiting the channel correlations.

The second part of the problem is the information extraction; the third part is the information organization. We will face the first problem exploiting the video audio through voice recognition techniques for allowing index and search inside documents with an audio description based on a generated transcript. The idea is to train a speaker-independent speech recognition tool and to synchronize the transcript files with the video and the related learning materials.

The third part of the problem is about data organization based on metadata conveniently organized in ontological classifications for enriching the searching capabilities. This activity can be relevant in the digital libraries context and as an advanced knowledge management tool.

2. INTRODUCTION

Since several years traditional, frontal lectures are streamed over the Internet. For example, at the University of Western Australia (Fardon et al. 2005) video streaming has given over 6000 students access to recordings of over 1800 lectures. The University of Sydney (Wozniak et al., 2005) is also using streamed video to deliver lectures both synchronously and asynchronously. Several similar cases are available: however, in these examples streamed video is mostly being used for transmitting unenhanced recordings of live lectures. At the University of Trento we started recording lectures a few year ago (Ronchetti 2003). We operated on a smaller scale, but used a richer format. We developed a software called LODE that is composed by an acquisition part that allows to inexpensively record learning events (e.g. lectures), and a fruition part that allows students to navigate the video while putting their cognitive focus in the “right” place: the video

itself, the slide that was projected in the classroom, additional learning resources. The system has been described elsewhere (Dolzani & Ronchetti 2005 a &b) but we shall mention here its general features.

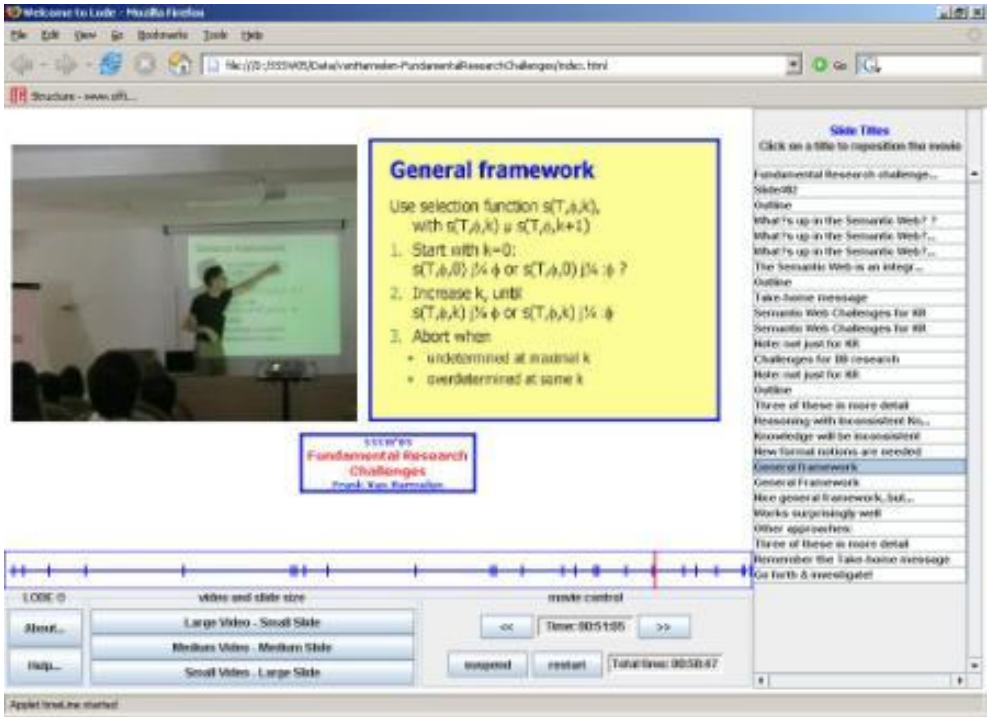

Figure 1: The Lode interface.

As shown in Figure 1, the final user sees at the same time a video of the educational event and a reproduction of the slide (if any) that was been projected in the classroom. On the right hand side, a list of slide titles (or talk sections) allows to quickly reposition the video (and to correspondingly synchronize the slide). The time bar (an horizontal blue line below video and slide) allows navigating the lecture as desired. Vertical bars on the time bar indicate events (such as slide changes).

The dimensions of the video and of the slide areas can change on the flight on user's demand. The user can choose among different video resolutions: small (144x180 pixels) medium (288x384 pixels) or large (520x562 pixels). To make the enlargement of the video possible, the document area is correspondingly reduced; in any case by double clicking on the document area the user can have an enlarged version presented in a pop-up window. The user can at any time switch to a different configuration with just one click on a suitable button. The blackboard becomes easily readable when the largest video area is chosen, so that fully traditional lectures can be supported.

Over the years, we gathered more than 500 hours of videos collected from university lectures, seminars and summer schools. The appendix describes the structure of our database and gives some statistical information about it.

The videos can be also played on the go with an ipod-like mp4 player, but their best use is on a computer where they can be navigated by slide title or through a time bar. We found that students find such system very valuable and useful: not only they can recover lost lectures, but in most cases they use the system to find and review critical information, e.g. when discussing with their peers while preparing an exam or for checking their notes. The starting point of the present work was the desire to ease the finding of such critical information. We were looking for an agile way to search for content in the whole body of lectures. We aimed at a first approximation that should be a simple and intuitive, Google-like tool, while future implementation will probably include clustering, ontology indexing etc. We therefore built the Needle prototype, where the acronym means "Next gEneration sEearch engine for Digital LibrariEs" in recognition of the fact that the issue is quite general and goes far beyond the e-learning case that we were starting from.

3. INFORMATION EXTRACTION AND ORGANIZAZION

Needle is based on the idea of exploiting all the information that is contained in each recording to allow searching. Each recording is composed by the video stream, the audio stream and a series of documents, each having a time stamp that refers to a given time in the recorded event. At present the most relevant document is a Power Point presentation, with each single slide temporally annotated.

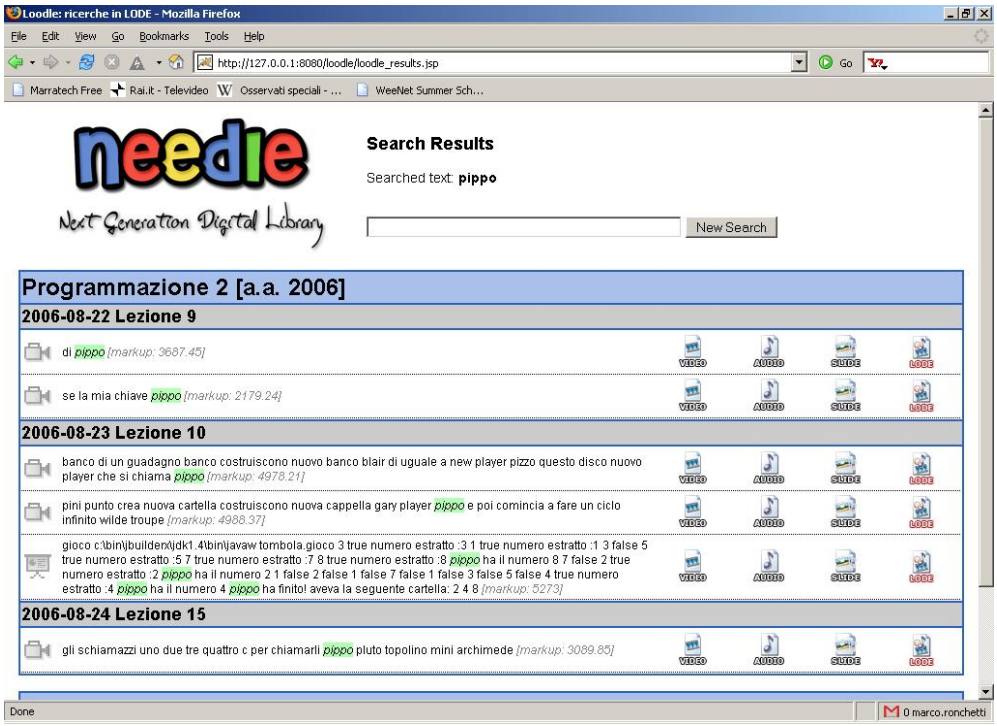

Figure 2: The Needle interface.

We extract information from the slides and from the audio stream. The first is quite obvious. For the second we experimented with various automatic speaker-independent speech recognition systems. We obtain a temporally marked transcript. A search engine produces indexes, that allows performing quick multimodal search.

The user interface of the resulting system is a web page with an input field for entering the search string. The string can be a single word or multiple words. Logical operators and wildcards are allowed. The result page displays the results arranged by course and lecture. For each lecture, results are shown inside a table that allows the user to easily understand which results belong to each lecture. On the left most part of the table an icon shows in which media the result has been found. In the present version of the prototype the search result can stem from the slides or from the audio track of the video. Right after the icon there is a text that contains the portion of text representing the hit. The searched words are highlighted inside the textual description with different caption colours. For video hits, the text is a part of the transcript file, for slides it would be part of the text present in the slide, or of its comment when present.

After the textual description four modes for accessing the multimedia results are displayed. The first link displays the video segment of the lecture in MPEG 4 format. To limit bandwidth usage we show a segment of video that lasts only a few minutes, but the user has anyway the option of viewing the segment in the context of the entire video. The second link points just to the audio file of the lecture in MPEG 3 format. Both audio and video files are displayed with the navigation bar. The third link is for displaying the slide that is temporally correlated with the video and it may contain the search hit. Also the slides are displayed with a navigation bar that allows going back to the previous or forward to the next slide. The forth media mode is a link to the LODE user interface that was described in the previous section.

We describe elsewhere (Fogarolli, Riccardi & Ronchetti 2007) the architecture of the Needle system, as well as the ASR experiments that we run. Here we just mention that the although the ASR performance were not fully satisfactory, the final result is quite good: even in presence of transcription errors, it is possible to perform quite efficient searches. In fact, it is not necessary to have perfectly transcribed phrases: what matters is just the transcription of the single words being searched: in unlucky cases the search will be incomplete, but the result is in any case adequate. We just mention that we did not tune the ASR we used at all: by simply tuning the vocabulary the result could be easily enhanced.

The system was actually tested with a set of 60 students to verify the soundness of our approach and the usability of the system. The encouraging results have been published elsewhere (Fogarolli & Ronchetti 2007).

4. CONCLUSIONS

We used speaker-independent ASR to generate time-marked transcriptions of the speech associated with video lectures. These transcription were used, together with other sources like the original power point slides used during a lecture, to perform searches in a body of recorded educational events (lectures, seminars, summer schools). We implemented a running prototype.

We consider this work as a first step in a very promising direction. After tuning the ASR we envision using the system for indexing the whole body of material that we collected. More information might be extracted by an analysis of the video sequences. The

addition of ontologies will further empower the user. Moreover, this approach can be applied beyond the original e-learning domain and extend to the large body of knowledge that will compose future digital libraries.

5. APPENDIX: THE STRUCTURE OF THE DATABASE

Our database contains five main entities: Actor, Event, Series, View and Document. The main relations among these entities are the following:
- An Event (e.g. a lecture) is linked to one or more Actors (e.g. the teacher) and belongs to a Series (e.g. a course).
- Views are the images that are shown at fixed times during the reproduction of an Event (e.g. the single slides shown during a lecture).
- Views are typically extracted from a Document (e.g. a MS PowerPoint presentation).
- Documents can contain also additional information connected to an Event or to a Series (e.g. suggested readings) and are referenced in the database since they can be used to give more context information (e.g. for refining a language model).

An Event is a recording unit. It has a video stream, an audio stream, a file containing all what is needed to reproduce the event in a browser (as defined by the Lode Architecture). It is characterized by a title, a type (lecture, seminar etc.). It belongs to a Series and it is linked to one or more Actors, and to one or more Documents. Its attributes are the spoken language, the date of the event, the number of Actors, the number of attached Views, the total time length.

An Actor is the representation of a person participating to an Event. Its data include (besides obvious fields like name, surname and title) also his/her native language.

A Series is a collection of Events related among themselves. It is characterized by a type (e.g. university course, seminar series, summer school, stand-alone event etc.). It has a name, an edition (since there might be multiple instances). It may have related Documents.

A Document can be any file. Typically it will be the source of a collection of Views (like in the case of a PowerPoint presentation). In other cases it will be documents related to an Event or a Series, that user can employ as additional resource. In the context of ASR they can be used for improving the knowledge of the context. Document attributes are its title, the format, the language.

A View corresponds to an image (typically a jpeg file) that has a temporal association with an Event. Its attributes are the image itself, the language, the title, a text and notes (some of these fields may be empty).

A table in the database keeps track of the associations between an Event and a View. It records the time at which a View has to be shown during an Event (a View can be shown multiple times during an Event). If the View originates from a PowerPoint file (in which case its origin is a PowerPoint slide) title, text and notes are automatically extracted. In future we might implement also other view extractors that capture data from other sources (like pdf or html files etc.).

At present our multimedia database comprises 17 series for a total of approx. 260 events corresponding to a total of approximately 500 hours of recordings and 50 speakers (30 in English and 20 in Italian). Most of the topics (88% of the total time) are in Computer Science, but there are some on different areas (Meteorology, Sociology, Engineering). 64%

of the speakers and 38% of the hours refer to events in English, 36% of the speakers and 62% of the hours to Italian.

Details are the following:

- 3 International summer schools on Computer Science (Semantic web, Web Engineering) for a total of 80 hours and 30 speakers. In English;
- 2 Workshops ("Human Language Technology and "Noise on the workplace", 15 hours, the first in English the second in Italian;
- 2 Schools on Meteorology, 60 hours, 15 speakers, partly in Italian and partly in English;
- 8 Bachelor courses in Computer Science, 250 hours, 7 speakers, in Italian
- 2 Master courses in Computer Science (Web architectures and Machine Learning), 80 hours, 2 speakers, in English
- 1 Master course on Sociology of Tourism, 20 hors, 1 speaker, in Italian
- 6 Seminars on scientific topics (Physics, Mathematics, Computer science), 6 speakers, 6 hours, 3 in Italian and 3 in English.

6. REFERENCES

Dolzani M., Ronchetti M., (2005a) "Video Streaming over the Internet to support learning: the LODE system". WIT Transactions on Informatics and Communication Technologies, 34, p. 61-65.

Dolzani, M., & Ronchetti, M. (2005b). Lectures On DEmand: the architecture of a system for delivering traditional lectures over the Web. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2005*, 1702-1709

Fardon, M.F. and Williams, J. (2005), "On-Demand Internet-transmitted lecture recordings: attempting to enhance and support the lecture experience", *Proceedings of the Association for Learning Technology, 12th International Conference*, 2005, Totton, England, Hobbs the Printers, 1, 153 – 161

Fogarolli, A., Riccardi, G. & Ronchetti, M. (2007), Searching information in a collection of video-lectures: an advanced e-learning application. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2007*, Vancouver, Canada, June 22-28.

Fogarolli, A. & Ronchetti, M. (2007), Case study: evaluation of a tool for searching inside a collection of multimodal e-lectures. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2007*, Vancouver, Canada, June 22-28.

Ronchetti M., (2003) Has the time come for using video-based lectures over the Internet? A Test-case report. *Proceedings of CATE - Web Based Education Conference 2003*, Rhodes (Greece), June 30 - July 2, 2003

Wozniak, H., Scott, K.M., & Atkinson, S. (2005), The balancing act: Managing emerging issues of e-learning projects at the University of Sydney. Balance, Fidelity, Mobility: Maintaining the Momentum. *Proceedings of the ASCILITE Conference* 4-7 December 2005.