An open source platform for indexing and retrieval of multimedia information from a digital library of graduate thesis

Ricardo Acosta-Díaz,
Haydeé Melendez Guillén
Miguel Ángel GarcíaRuiz
Armando Román Gallardo
Jorge Rafael Gutiérrez Pulido
Pedro Damián Reyes
Universidad de Colima
Universidad Autónoma de Baja California
México
Contacto: acosta@ucol.mx

Abstract: In the MIND (MIxed-media Networked Digital Library) project, we have developed an environment to explore efficient and effective mechanisms for indexing and retrieving Mixed-media information from a digital library of thesis. We have stored and indexed not only the text from the dissertations, but the images contained in the text, the slides used for the thesis defense, and the audio and video of the presentation of the thesis itself. Full-text indexing of the whole document and the text in the slides, and event-based indexing of the presentation and the video will be used to create a rich Mixed-media library with support for queries that incorporate different media and navigation between these media. For instance, looking at the thesis defense and asking for more details and being redirected to the appropriate place in the document or video. In this paper we focus on strategies for identifying the segments within thesis document that correspond to each slide in the presentation.

Introduction

This paper describes the work on the MIND (MIxed-media Networked Digital Library) project. The objective of this work is to provide an environment for the automatic capture, indexing and retrieving of mixed-media information from a digital library of graduate thesis.

We have defined the requirements and created a prototype for the automatic capture of the media, and to test preliminary algorithms for mixed-media retrieval using media alignment, events and keywords.
The organization of the paper is as follows. First, we introduce the term "digital library" and its advantages, second we present the background to the field of mixed-media information retrieval, third we present the description of the architecture of our first prototype. Finally, we mention on-going and future work.

What is a digital library?

The term digital library means different things to different people. For some, it simply suggest the computerization of the traditional libraries. For others, who have studied the science of the libraries, it is the execution of the functions of the traditional library in a new form, holding back new types of information new forms of getting data and new preservation and storage methods, more confidence in electronic systems and networks, and a dramatic change in the economical, organizational and intellectual practices.
For many professionals in the computer sciences, a digital library is simply a text-based distributed information system, a collection of distributed information services, a distributed information interlinked space, or a

multimedia information system in the Internet. For the users of the Web, a digital library suggests more improvement in the performance, organization, and usability.
A digital library is composed of three classes of elements: data, metadata and process [15]. Data is the contest of the library (Hypernovel, scientific visualization, and computers programs). Metadata is information about the library and its data (dynamic index, personnel structures, and annotations). And the processes are active unctions that execute on the elements of the library (Search in the text, search in images and videos, retrieval).

Advantages of the digital library

The digital libraries offer the following advantages:

1. They contain information in different media (traditional books are not able to include animations, audio and video).
2. Search engines facilitate information retrieval.
3. The information stored in the digital library can be accessed at the same time by several users (unlike traditional books where we need a copy for each of them).
4. The material can be adapted to each user (letters color, font size).
5. They offer information that can be used without the restrictions of intellectual property (in some cases).

Digital library of thesis

Libraries play a central role in education and learning, the same is the case for the digital libraries. Between the principal roles of the library in education are [8]:

1. To share resources.
2. To preserve and to organize artifacts and ideas.
3. They have a social and intellectual role for bringing together ideas and people.

One of the main information resources of universities are the thesis produced there, principally postgraduate thesis. Those documents report good part of the research work that is produced in the institution. Besides, they support courses and are often good introductions to current topics, because generally they present a description of the state of the art in a particular field.

With financing from the US Department Education's Fund the NDLTD (The project Networked digital Library of Thesis and Dissertations) was created in United States. The project had the purpose of promoting the electronics publication of thesis between postgraduate students and publicizing the work that it made in the universities.
In contrast with the MIND project the NDLTD only stores the thesis documents an supports only text retrieval.

Creating a postgraduates digital library of thesis has several advantages:

1. Higher circulation of the material providing access by Internet.
2. Useful and current information .The postgraduates thesis generally include a complete revision of the state of the art in the topic study that can be of great value for those that are initiating in this area.
3. Most of the theses, nowadays, are written on a computer, so it reduces the cost of capturing them.
4. The information in the document is property of the educational institution. The educational institutions are interested in promoting its products.

Mixed-media Information Retrieval

The possibility of including media other than text in digital libraries has originated research in the last years relative to the indexing and content-based retrieval of images, audio and video.

For example, Content-Based Image Retrieval [11] studies mechanisms for retrieving digital images giving as attributes parameter of the image (colors, forms, texture, etc.), or other images similar to those interest. For this algorithms of image processing, comparison of histograms [4] and the wavelet transformed [7] have been utilized, between others.

Image processing techniques also have been used for video indexing [2, 12]. For example, algorithms of analysis of movement and morphologic methods were used to index scenes from a digitalized soccer game [14]. Additionally, voice recognition algorithms and natural language processing have been used for indexing audio [2, 5].

More recently research have began to explore synergistic effects of integrating techniques from the retrieval of different media in mixed media databases [1, 3]. This work is oriented to information databases that include different media, like news repositories, digital libraries in the WWW, records of medical patients, etc., they are indexed in various media and make use of the complementary and redundant information to facilitate the retrieval process. For example, identification of an radio anchorman using voice-processing algorithms can help the segmentation of a news video.

In a recent work, techniques for content based image retrieval were combined with text retrieval techniques of the text associated to the images in a digital library of animals and plant species from Baja California in the QBICAT system [Figure 1]. Queries to the system can be made using keywords and/or images drawn [9].

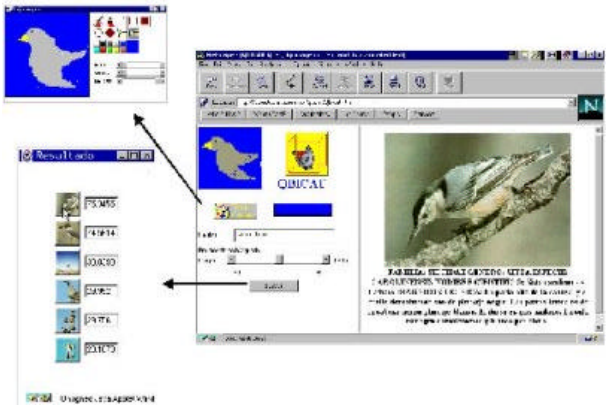

Figure 1: Image retrieval using image-content and associated text with QBICAT.

MIND system description

The complete thesis documents were digitalized and stored, including text and images, as well as the thesis defenses (audio, video and slides). In this way it is possible to access the thesis documents from the presentation. That is, the user can provide a list of keywords to find a relevant thesis and retrieve the part of the document or the slides where these words appear. If he decides to see the digital video, he will use media aligning to change from one media to another. (Figure 2).

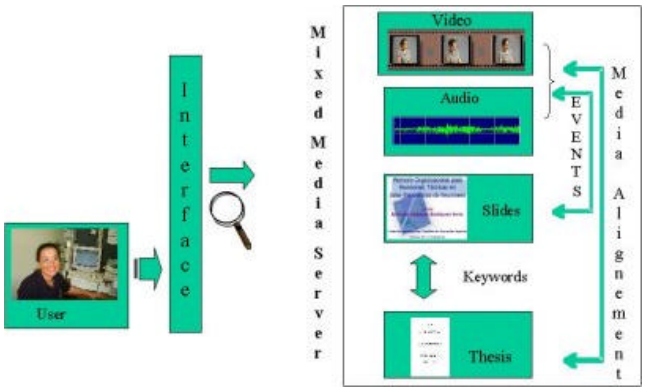

Figure 2: MIND system architecture

1). Capture

Audio and video are captured automatically by the system when the student makes the presentation using WP-SICREP[16], but we must provide the following information about the presentation:

1. General information, name of the presentation, author, etc.(Figure 3)

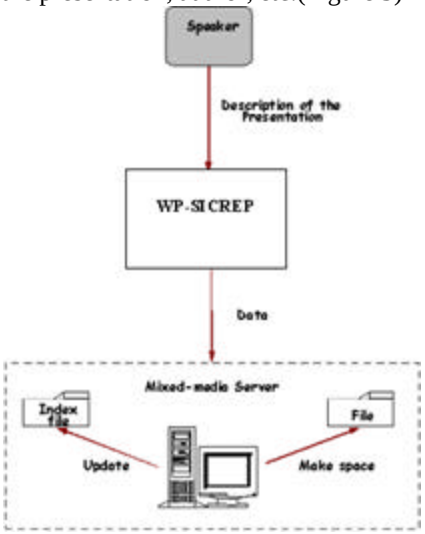

Figure 3: Procedure for the capture of information about the presentation

The system creates a file in the mixed-media server, that will store the information of the presentation, after this the system will be ready to receive the events (change of slides, in this case) and they will be added to the information file.

2. Information about its content (audio,video and slides), (Figure 4)

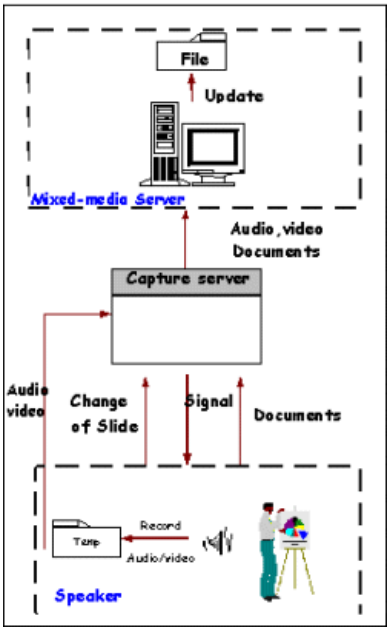

Figure 4: Scheme for audio, video and text capture

For each slide there is a corresponding speaker's audio and video file which indicates to the server when the speaker changes the slide. The server notify the change to the client and stores the audio and video. At the end of the presentation the audio, video, URL of the slides uses and the general information about the presentation are sent to the mixed media server.

2). Indexing

Audio and video are associates to their slide using events. During capture, the system stores the slides (HTML pages) in the mixed-media server, later it processes each slide to obtain the relevant words that identify each one. The thesis documents also are converted to HTML format and then they are indexed. The procedure for accomplishing the indexing is based on the TFIDF algorithms (Term Frequency Inverse Document Frequency) [13].

3). Information retrieval

We analyzed the process required to access the digital library by interviewing potential users and librarians and we developed an interface for information retrieval. (Figure 5).

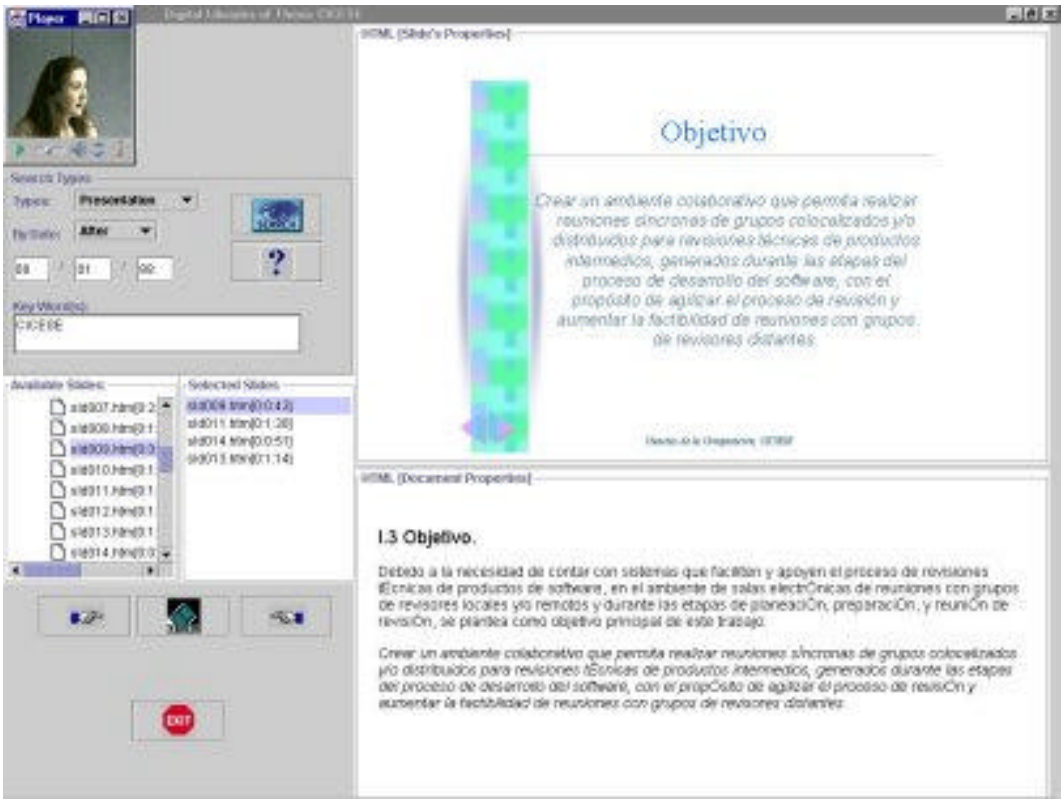

Figure 4: The MIND system information retrieval interface

So far we have implemented a retrieval methods:

1. Using keywords, we access the thesis documents and defense presentation files.
2. By date, which allows the retrieval of stored documents between a range of dates.
3. From the presentation slide of the defense the user can access detailed information in the thesis document.
4. From the thesis document the user can access the part of the defense that refers to that section.

4). Presentation of the information

The list of slides are presented in form of a "tree", where one is able to navigate on the slides of all the presentations stored. Additionally, when the user selects a slide of any presentation its contents will be displayed in the panel at the upper right side, in this way the user is able to navigate over the contents of all or part of the presentations, in the same way the lower frame displays the part of the thesis document to which the slide is referencing, as well as a list of related documents to that theme. So the user is able to check the contents of any document and decide whether it is of interest before reproducing it. Another advantage of the display of more than one presentation as user ar be able to select slides from the different presentations to which they belong to be reproduced at a later time.

Conclusions and Future Work

We are currently testing and extending the prototype based in the architecture described here. Although we have presented the entire MIND system on this paper, we have concentrated our work in a few directions:

- First, we have developed an environment that allows automatic capture, indexing and retrieval of mixed media using events and keywords.

- Second, this prototype is being used as a laboratory to propose and test indexing and retrieval algorithms. As first instance, we have used keyword to align the slides with the thesis document.
- Third, the system provides users with synchronized information media (slides, text, audio and video) retrieval, allowing them to find more detailed information from one topic in another media.

Mixed-media Information Systems are very useful applications, particularly for Digital Libraries. MIND is a simple architecture developed for carrying out retrieval tasks on a number of media.The work presented in this paper is just the first step towards our larger goals. On the future We will be exploring semantics and how it can contribute to enhance retrieval for MIND. For instance, the so-called Semantic Web (Berners-lee et al., 2001) will be populated with semantic structures allowing smart-serching, agent-based computing, and multi-lingual applications. Semantics therefore can also be helpful to Mixed-media Information Systems, particularly Digital Libraries.

References

[1] Proc. of the “Workshop on Intelligent Integration and Use of Text, Image, Video and Audio Corpora”, AAA-97 Spring Symposium Series, Marzo 1997, Stanford, CA.

[2] Brown, M. Foote, J., Sparck-Jones K. And Young, S., "Automatic Content-Based Retrieval of Broadcast News". Proc ACM Multimedia 95, 35-43, ACM Press. Nov. 1995.

[3] Proc. of the Workshop on Mixed Media Databases. Conf. on Automated Learning and Discovery, Pittsburgh, PA, Junio 1997.

[4] Falatous, C et all., "Efficient and Effective Querying and Image Content". Journal of Intelligent Information Systems, Vol. 1, No. 3, 231-262 1994.

[5] Hauptman, A., Witbtrock, M. Rudnicky, A y Reed, S, "Speech for multimedia information retrieval". Proc. of User Interface Sofware Technology (UIST-95), Pittsburgh, PA, ACM.

[6] Ibarra, M., "Interpretación de la estructura del discurso escrito en lengua española". Tesis de doctorado. CICESE. Diciembre 1999.

[7] Jacobs, C., Finkelstein, D. And Salestin, D., "Fast multiresolution image querying". Proc. ACM SIGGRAPH´95 pp. 277-286, Agosto 1995.

[8] Marchioni, G y Maurer, H. "The roles of Digital Libraries in Teaching and Learning". CACM, Vol. 38, No. 4, pp 67-75, 1995.

[9] Meza, V. Y Favela, J., "Integrating Image Content and its Associated Text in a Web Image Retrieval Agent". AAAI Wkshp. On Intelligent Integration and Use of Text, Image, Video and Audio Corpora, Stanford, CA., pp. 52-56, 1997

[10] Meza, V. y Favela, J. "Information Retrieval by Image Content and its Associated Text in a Digital Library". Proc. of the Conf. on Automated Learning and Discovey, Pittsburgh, PA, Junio 1998.

[11]Niblack, W. Et all., "The QBIC project: Query image by content using color, texture and shape". Storage and Retrieval for Image and Video Databases, pp. 173-187, San Jose 1993. SPIE.

[12] Pentland, A., "Machin Understanding of Human Behavior in Video, In Intelligent Multimedia Information Retrieval" Maybury, M. (ed.), MIT Press, 175, 190, 1997.

[13] Salton, G. "Automatic Text Processing" Addison-Wesley. 1988.

[14] Uriostegui, A., "Indexado de video digital para la recuperación de escenas de interés aplicado a un evento deportivo". Tesis de maestría. CICESE. Septiembre 1996.

[15] Nurnberg, P., "Digital Libraries: Issues and Architectures", Texas A&M University.

[16] Garcilazo, J., "Sistema para la captura y recuperación de cursos electrónicos"., Tesis de maestría. CICESE, Septiembre 1998.

[17]
Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The Semantic Web. Scientific American. May 2001
http://www.ryerson.ca/~dgrimsha/courses/cps720_02/resources/Scientific American The Semantic Web.htm