Computing and Informatics, Vol. 38, 2019, 1375–1402, doi: 10.31577/cai_2019_6_1375

FRAMEWORK FOR KNOWLEDGE DISCOVERY IN EDUCATIONAL VIDEO REPOSITORIES

Jorão Gomes Jr., Laura Lima Dias, Eduardo Rocha Soares

Postgraduate Program in Computer Science
Federal University of Juiz de Fora
36036-900, Minas Gerais, Brazil
e-mail: {joraojunior, laura.lima, eduardosoares}@ice.ufjf.br

Eduardo Barrere, Jairo Francisco de Souza

LApIC Research Group, Department of Computer Science
Institute of Exact Sciences, Federal University of Juiz de Fora
36036-900, Minas Gerais, Brazil
e-mail: {eduardo.barrere, jairo.souza}@ice.ufjf.br

Abstract. The ease of creating digital content coupled with technological advancements allows institutions and organizations to further embrace distance learning. Teaching materials also receive attention, because it is difficult for the student to obtain adequate didactic material, being necessary a high effort and knowledge about the material and the repository. This work presents a framework that enables the automatic metadata generation for materials available in educational video repositories. Each module of the framework works autonomously and can be used in isolation, complemented by another technique or replaced by a more appropriate approach to the field of use, such as repositories with other types of media or other content.

Keywords: Semantic annotation, knowledge discovery, video repositories

Mathematics Subject Classification 2010: 68-P20

1 INTRODUCTION

The ease of creating digital content coupled with technological advances enables institutions and organizations to increasingly adopt distance learning [16, 41]. Current efforts at distance education are geared towards a more individualized and personalized education. Researchers are interested in observing and modeling the profile of students, making it possible to adapt the learning according to the personality and the needs of students [33]. This factor allows the effective use of distance education systems and the permanence of students in the offered courses [6].

Another factor that also contributes to this effective use of e-learning systems is the correct administration of didactic materials, in order to make them available according to the learning needs of each student [42]. Teaching materials have also received attention from the researchers, since it is difficult for a student to obtain an adequate learning material by himself. A high effort and previous knowledge about the material and the repository are necessary to be succeeded in the search task. This difficulty is further compounded by the growth in the number of learning materials in the repositories [26], which can cause many irrelevant materials to be returned in the search. The fact that, even with advances in technology, students still cannot obtain what they want by doing searches in web repositories indicates the relevance of the studies focused on the understanding of these repositories in order to improve the search and the administration of the didactic materials, especially when we talk about videos the natural language of which is often vague and uncertain [15].

In addition, it must be kept in mind that most of the time spent on learning online courses is dominated by student interaction, unlike in-person courses where the instructor dominates most of the time. Another point to note is that the classroom is a network in which students from different geographic locations interact socially, sharing information and resources, and even performing joint projects [20]. Therefore, for a correct administration of learning materials, it is also necessary to consider an adequate creation of these shared information spaces. The work of [34], for example, presents a good alternative for the creation of globally shared information spaces, named the Linked Data initiative. This initiative is an interesting option for the discovery and understanding of Open Educational Resources (OER) data. Linked Data consists of a set of practices for publishing, sharing and interconnecting data in Resource Description Framework (RDF) format. Educational repository administrators are realizing the potential of using Linked Data for describing, discovering, linking and publishing educational data on the Web [34]. The Linked Data is based on the use of Uniform Resource Identifier (URI) references to identify digital documents, as well as real content and even abstract concepts [22]. Thereby, obtaining materials that use this format allows a greater flexibility and connection within the repositories of learning objects, as is the case of the works [29, 9].

This work proposes the creation of a framework that enables the generation of metadata for the materials available in educational repositories of videos and texts, in order to facilitate the creation of these shared information spaces. In addition, this research intends to present a way for administrators of repositories and course creators to better know their repository as a whole, through a panoramic view of the areas of concentration. In this work, we consider scenarios for the use of this framework on repositories with materials that have little or no previously associated metadata. A case study of the use of the framework on a real repository of educational videos will also be presented. Besides, the present work shows how didactic materials are related within the repository at the end of the process. These relationships give a big picture of how the areas of knowledge present in video lectures are related.

2 RELATED WORK

Many works in the literature relate to ways to better explore and gather information about the media contained in repositories. Metadata are important for integrating learning objects from different repositories, and recent integration strategies are heavily dependent on metadata previously associated with learning objects. For example, in [7, 40], the authors present a recommendation system for the Moodle platform that indexes learning objects from different repositories and searches in their stored metadata. A re-design of the Moodledata module functionalities is presented in [10]. The authors aim to share learning objects between e-learning content platforms, e.g., Moodle and G-Lorep, in a linkable object format. Their proposal allows a semantic description of the learning objects. However, we argue in this paper that metadata of learning objects in open repositories, especially video lectures, are usually poorly descriptive. In Section 3.1.1 we present an analysis of metadata quality in an open Brazilian video lecture repository. Automatic metadata extraction techniques are important to enrich learning object information. In [38], natural language processing techniques are used to improve browsing and searching within the BBC's radio program repositories. Some papers in the literature also rely on technologies such as Automatic Speech Recognition (ASR), that is the process that allows computers to receive a speech signal as input and convert it into natural language text. This technique is used to automatically extract information from the audio track of multimedia content on the web. The extracted information, generally, has two main uses: in the composition of multimedia applications (e.g., closed caption, personal assistants, other ways of accessibility) or in understanding multimedia content in order to improve the search for it on the web [44, 3]. In the second case, ASR has great importance on extracting content from an audio signal that can be useful for media representation. Researches such as that of [44], for example, make use of ASR to extract the spoken content of videos in order to extract keywords through semantic annotation. Then, the extracted keywords are used to recommend those videos through similarity calculations. The present work, however, uses the result of ASR as input to a series of natural language processing techniques to associate semantic resources with educational videos in order to help to identify videos with similar content and calculate similarity to discover and understand the repository.

Concepts addressed in the materials of a repository can be discovered and retrieved by several processes. Other works have also explored the task of automatic indexing or automatic annotation, as in [21, 44, 45]. In [21], the authors demonstrate the effect of extracting and combining visual and audio information into the search process using part of the TREC 2001[1] Video Recovery Trail for evaluation. Among the analyzed information are: speech recognition, face detection, text extraction via OCR and the use of image similarity matching. OCR techniques are also used in [25] to summarize fixed-camera video lectures by detecting handwritten content. In [45] the authors present a visual navigation system for exploring bio-medical OER videos, while the present work explores linked data for the discovery of the main topics and relations among videos. Word Embedding models were used in [11] to detect segment boundaries in video lectures. The authors argue that classical scene detection algorithms are useless for segmenting video lectures because this kind of video is usually recorded in one shot.

The calculation of similarity between media is used by some systems to support the content recommendation process. Iris AI[2], for example, makes the recommendation of scientific articles from an initial user's indication. Like this work, Iris AI calculates the similarity between documents through the relations in knowledge bases. However, this is done with the focus on the recommendation of the articles, while our work is also concerned with the existing knowledge in the repositories where the video lectures are stored. In [30], the authors closely approximate our research by employing ontologies and automatic metadata annotation, retrieving information to use recommendations and also by aggregating related content. However, the results are still experimental and limited to the relations defined in the SCORM standard.

The work of [12] reports that one of the main challenges in the implementation of technological services in repositories is the visualization of information and that current research in this area is directed towards the improvements in the retrieval of scientific and academic information. Finally, the works [13, 35] report that new orientation strategies for service innovation and the functionality of technological means in institutional repositories are needed, as well as the effort to solve problems such as obsolescence and guaranteeing the satisfaction of academic communities based on the usefulness of the repositories and the experience and usability of students and users. So [13] builds a systematic review about the user experience in institutional repositories and discusses possible solutions to collaborate with those needs.

Considering the recent and important concerns with the use and knowledge of educational repositories, our work aims to make it possible to learn about educational video repositories, providing automatic metadata to support their usability. We present a framework used to extracted metadata from video lectures based on speech information. We compare different techniques used for automatic semantic

[1] https://trec.nist.gov/

[2] http://iris.ai

annotation. Unlike other works, we present how the extracted semantic metadata can be used in two tasks: similarity calculation and clustering. In similarity calculation, we propose an algorithm to extract and compare extrinsic metadata information provided by the DBpedia ontology. In the clustering task, we show how the label propagation algorithm can be used on the knowledge graph to identify items in similar domains. We make all data available for further research.

3 CONCEPTS AND TECHNIQUES

Currently, several techniques can be used for video indexing, such as color histogram sorting, shape recognition, action recognition, face recognition, text extraction through Optical Character Recognition (OCR), among others. Regardless of how data is collected from videos, these videos can be associated with pre-existing concepts from a knowledge base. These concepts can be seen as the main topics. This process is called semantic annotation. The semantically annotated videos are related to entities, which, in turn, are part of a network of relationships with meanings, such as an ontology or a thesaurus. Thus, media annotation facilitates the search and recommendation processes in repositories and many researchers have been working on improvements in it for various media [5, 36].

The following sections aim to clarify the concepts and techniques used in the proposed framework. They are organized to present the process of indexing and retrieving information, starting from video files until the relationship of the videos with close contents.

3.1 Process Architecture

It is a difficult task for students to find satisfactory didactic materials. This activity demands effort and knowledge about the material and the repository. The difficulty is worsened by the growth in the number of learning materials [26]. Therefore, educational video repositories must deal with these difficulties, especially when they are constantly fed with new materials.

Since the proposal of this work is to facilitate the understanding of these materials and repositories, we use as the setting for our experiments a real repository of educational videos produced by Brazilian universities. The application scenario is contextualized within the Video Advanced Search Group (GT-BAVi) of the Brazilian National Research and Educational Network[3] (RNP). The objective of the GT-BAVi is to develop a prototype to facilitate the semantic enrichment of the video repository and to facilitate the search. For this, a framework was implemented in order to accomplish this task automatically.

The developed solution can be divided into three main steps: Content Processing, Context Association, and Knowledge Graph. For the Content Processing step, an ASR system was used, since the main focus of the framework is on processing

[3] `https://www.rnp.br`

videos. We can, then, extract the video lecture content in a text format through its audio track. For the Context Association step, we used an Automatic Semantic Annotation method to attach concepts from a knowledge base to the videos. Finally, in the Knowledge Graph step, we obtain the relationships between the videos according to their similarity related to the extracted concepts.

Figure 1 represents the solution steps. Each step is implemented by a corresponding process.

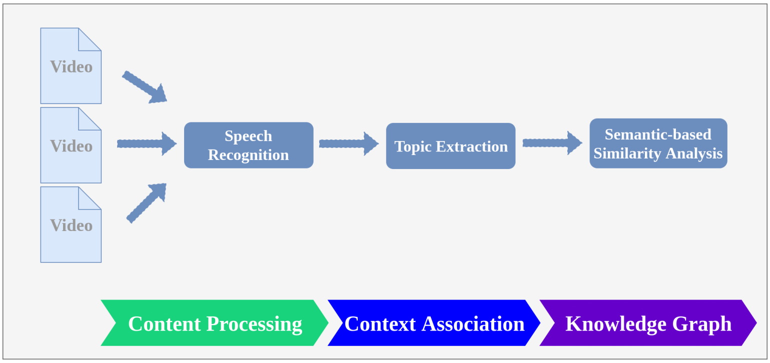


Figure 1. Processing Framework

3.1.1 Analysis of the RNP Repository

Among all the RNP repositories, the video repository VideoAula@RNP[4] is the focus of our research, since it is currently used by Brazilian academic institutions to store video lectures. An analysis of the pre-existing video metadata in the RNP repository was performed. This analysis is relevant to an initial validation of the prototype. These metadata are tags manually created by video editors to ensure that the video is found through keyword searches. The information collected for this analysis was the number of videos within the repository, the total amount of metadata used in the repository, how many of them were different, how many tags on average each video received, the number of videos with useless tags, as tags out of context or very generic. Examples of useless tags are “video”, “video lectures”, “tag”, “test”, “teacher description”. The information collected is in Table 1.

The RNP repository had 858 video lectures. Each video had an average of 2 to 3 metadata, which totaled 2 225 metadata in the repository. However, only about a third of them were unique, indicating that many of the metadata were repeated. Thus, many videos would be returned after searching for some keywords. Furthermore, 604 videos had useless metadata, i.e., they did not add a specific identification to the videos to which they were attached.

[4] `http://www.videoaula.rnp.br`

Information	**Collected Values**
Number of videos	858
Total of tags	2 225
Total of distinct tags	849
Average tags per video	2.59 ± 1.34
Number of videos with useless tags	604
Number of videos that did not have tags	2

Table 1. RNP scenario using only manually associated metadata

After collecting more specific information about the metadata, the following data were also found: 540 videos had only 2 metadata, being "video" and "video lectures", making it impossible to identify any of these videos by their content. One of the videos had only the "test" metadata. Still, 16 videos had only the "Teacher description" metadata. These data show that it is impossible to find a specific subject in any of these videos just by searching for terms. The data collected also indicates that the metadata attached to the video often limit the potential of a search in the repository. This situation occurs mainly due to informalism and little dedication during the metadata creation stage when uploading videos, generating inappropriate metadata for a future video search.

Since many videos in the repository have too few metadata, and their titles are generally vague (e.g. "Exercise_5"), this repository is a good choice to show the potential of our proposal based on the ASR and semantic annotation. We want to show that our framework is able to automatically extract meaningful information that can be used to improve the search and recommendation of these videos from which we previously had no information.

3.2 Description of the Framework

We have defined the framework as a process composed of three steps that work in isolation and these steps will be validated individually. Since the audio is the main source of information in video lectures, the Subsection 3.2.1 presents an ASR system trained for this work. The Subsection 3.2.2 presents the Context Association supported by Natural Language Processing techniques that allow automatic semantic annotation. In this section we will present two options for this task and discuss the results for both. Finally, the Subsection 3.2.3 presents the Knowledge Graph supported by the similarity calculation between videos by walking in the category graph of a knowledge base. In addition, the parameters used in this walk are discussed in order to demonstrate the best options considering the accuracy and computational cost.

The differential point of the process presented in this work is the possibility to develop the process steps as different services that can be instantiated as needed, such as the ones regarding the need of specific processing steps and the type of material utilized.

3.2.1 Content Processing

The automatic semantic annotation process discussed in this paper focused on the video lectures scenario. In this context, even more than in others, most of the information is present in the teacher's speech. For this reason, the automatic semantic annotation process depends primarily on ASR. According to [14], ASR systems generally are built on three main models: **acoustic**, **lexical** and **language model**.

The acoustic model is responsible to allow the ASR system to determine which sequences of speech unit (generally, phonemes) have more similarity with the vectors of acoustic characteristics that were extracted from the audio signal. Thus, the acoustic modeling is done by training an algorithm to predict the probabilities of phonemes to be related to an audio segment. Some of the most popular algorithms to do acoustic modeling are the Hidden Markov Models (HMMs) [18] and Deep Neural Networks (DNNs) [24]. For training an acoustic model, it is necessary to provide a corpus of speech containing well segmented audio files with speeches from the specific target for which the ASR is being designed. Furthermore, the training also requires their respective ground-truth transcriptions.

A lexical model is basically a dictionary that maps words of a vocabulary to a sequence of phonemes. This dictionary is used by the ASR system so that it can convert the sequences of phonemes recognized through the acoustic model into words. To create this model, there are phonetic converters that take a sequence of characters and return a sequence of correspondent phonemes. For example, in [39], the authors have used Long Short-Term Memory (LSTM) recurrent neural networks to do the automatic grapheme-to-phoneme conversion.

The language model is not essential for the operation of an ASR system. However, its use significantly improves the accuracy of those systems because the acoustic model is not enough to obtain a satisfactory transcription. The acoustic model only infers sequences of phonemes that are then converted into a sequence of words, through the lexical model, without any grammatical restrictions. The language model acts exactly on this issue by calculating conditional probabilities of words from the vocabulary to be recognized after others. With this, it is possible to restrict the possibilities of recognized sequences of words. Thus, ungrammatical sentences have low probability to be formed, and that reduces the search space, decreases the time for recognition, and improves acoustic ambiguities resolution [43]. Therefore, it is a consensus that in systems which deal with wide vocabularies, like the ASR system for continuous speech that is used in this work, the language model is extremely important. The training of the language model is done through a text corpus where the word frequencies and conditional probabilities of the word sequences are extracted.

To create a robust ASR system it is important that the system is trained with a large data volume that covers the main characteristics and variations (e.g. noise in the audio, accent, intonation, gender, age) present on the speeches of the system's target public. The biggest challenge in designing ASR systems is to obtain a speech and text corpora that are proper for training. The creation of these training bases is a costly process, as there is not enough free and open properly catalogued audio samples available, and the process of creating such samples is expensive in terms of time, space and money. When we talk specifically about training ASR systems for Brazilian Portuguese, that difficulty is even bigger, which makes those systems perform poorly when dealing with different accents and with different types of noise and distortion in the speech signal [32]. An alternative to obtaining a better accuracy in ASR with few data is to train it to be a specialized system by using databases that contemplate only the target scenario of the final application.

That is why in this work we trained our own specialized ASR system for Brazilian Portuguese video lectures. To train the acoustic model, we use a speech corpus extracted from subtitled video lectures in Brazilian Portuguese that are made available for free by Coursera[5], with a total of 55 hours of audio. We also added to our dataset a total of more than 2 hours of audio from corpora made available for free by VoxForge project[6] and by the Signal Processing Laboratory of UFPA (LAPS-UFBA) from Brazil[7]. The acoustic model is based on Deep Neural Network (DNN), which has 440 neurons in the input layer and 6 hidden layers of 2048 neurons each. The output layer has around 4000 neurons. For language model training, we use a union of text corpora from multiple sources such as subtitles from Coursera video lectures, open and free text corpora like CETEN, OGI and LapsFolha that are also made available by the LAPS plus Wikipedia articles. In total, the text corpus used to train our language model has about 13 million sentences.

To obtain the final ASR model that we use in the framework proposed in this work, we performed experiments aiming to explore different training configurations and pre-processing steps, separately or combining them, in order to verify which of them are responsible for obtaining a complete ASR model that has better recognition accuracy in our test data. For the acoustic model, we evaluate the impact of changing the training algorithm and audio sample rate. We have also evaluated the impact of segmenting the audio into smaller chunks and aligning them with their respective transcriptions. For the language model, we evaluated the effects of the variation of the model order[8], text pre-processing and probability smoothing methods.

To evaluate the trained models, we build manually an evaluation dataset composed of 2 hours of audio extracted from different parts of video lectures that are not in our training data. To build this evaluation dataset, we transcribed manually

[5] `https://pt.coursera.org/`

[6] http://www.voxforge.org/home

[7] `https://laps.ufpa.br`

[8] The order is related to the dependence of each word given the $n-1$ words which precede it. This means that for a model of order 3 (trigram), the probability of occurrence of a word is related to the two that precede it. For the order 4 (4-gram) model, the probability of occurrence of a word is related to the three that precede it, and so forth.

each audio in it, which resulted in about 581 spoken sentences with a ground-truth transcription. The accuracy metric used to evaluate the models was the Word Error Rate (WER). This metric represents the number of modifications (insertion, replacement or removal of words) that are necessary for the recognized sentences to transform them into the correct ones. That is, the lower its value, the better the accuracy of the recognizer [32].

After the experimentation, we get the best WER of 45.5 % with the following training settings:

- The training of the **best acoustic model** was done using the WAV codec, audio sample rate of 8000 Hz, MONO channel, and DNNs as the training algorithm. In this model, we have also performed the segmentation and alignment of audios with their respective transcriptions.
- The best language model was obtained with the interpolation of two other models, of 4-gram and 3-gram, with a normalized training corpus. The following tasks were applied in the normalization step: setting all words to lowercase; transforming all dates, times, percentages, Roman numerals, cardinal and ordinal numbers, acronyms, abbreviations and monetary values to their full forms (e.g., if it was in English, "5°" and "100 %" would become "fifth" and "one hundred percent", respectively). Furthermore, the applied probability smoothing was the Kneser-Ney method [31].

The WER of 45.5 % in speech recognition that we obtained in this work are close to those obtained in commercial systems such as Google[9], Microsoft[10] and IBM[11]. For our video lectures evaluation dataset, Google obtained 35.9 %, IBM 73.7 % and Microsoft's model achieved a result of 44.7 %.

Since ASR is used as the basis for the following processes of our framework, it is important that the recognition error rate be the smallest possible. The presented results in this subsection showed that our specialized ASR model is capable of obtaining a good performance in the context of this work. However, it is still not an error-free process. Therefore, in the following subsections we analyze and discuss ways of performing the following processes on the noisy video lecture transcriptions from ASR.

3.2.2 Context Association

There are several approaches aiming at video search improvement through semantic annotation. The approaches to assign these annotations can be divided between those that make use of external data related to the video and those that use only the information contained in the media. In the first group, for example, the use of

[9] `https://cloud.google.com/speech/`

[10] `https://docs.microsoft.com/pt-br/azure/cognitive-services/Speech/API-Reference-REST/BingVoiceRecognition`

[11] `https://www.ibm.com/watson/developercloud/speech-to-text.html`

texts around images is used by [17] to verify correspondences between images and texts and thus to find the interrelated sets of terms and topics, instead of simply annotating texts. For works that only use information contained in the media, we can cite the work of [1], which presents some approaches with event-based content (using the visual content of the videos). Among the approaches presented, there are mechanisms such as limit detection of takes, keyframes extraction for representation of important parts of the video, structural analysis and scene segmentation combined with OCR techniques for extraction of textual resources and creation of tags. Although these approaches have good classification results, they are limited to specific types of videos, usually those with a well-defined content and temporal structure.

Although there are several approaches for video annotation, there are not many studies that analyze the quality of information used to semantically annotate educational videos. Textual quality information is commonly present in video lectures repositories. For example, almost all Videoaula@RNP videos were recorded as an expository lesson, containing slide projection throughout the video. However, most of the information content of the video is in the teacher's speech. Therefore, search engines that only use video metadata (title or abstract) cannot help the user when he/she wants to find a video using terms that appeared during an exercise or an example that the teacher approached. In this view, we analyzed the impact of the semantic annotation process on several data sources extracted from the Videoaula@RNP and how the quality of the new tags created can influence the knowledge about the video lectures; for use of the search engines, for example, and general knowledge of the repository.

The automatic semantic annotation experiment was performed using two approaches: an Entity Linking Approach and a Topic Extraction Approach. The Entity Recognition Approach has a natural language text as input and produces a set of $(term, entity)$ pairs that represent the concept (entity) associated with the term present in the text. For this process, natural language processing techniques are typically used to tokenize the text and identify the correct terms. We used DBpedia Spotlight [28], which makes use of DBpedia to create a map of candidate entities for each term found and to disambiguate the term. In turn, the Topic Extraction Approach produces a set of entities that represent the main subjects of the text. For this task, the approach proposed in [38] was adapted, which makes use of the DBpedia category graph to identify the entities with the greatest relation to the text.

The process of automatic semantic annotation was performed with each data source combination with both annotation approaches. The results were measured using the recall and TopN measures. The TopN is measured as follows: considering a document with a total of N_r manual annotations not ranked for a given video and that were associated with N_k correct annotations by the algorithm, and let $rank_i$ be the position of the i^{th} correct annotation of the response set, thus the TopN is defined as the Equation (1). A constant $\alpha = 0.8$ was adopted to adjust the penalty associated with the correct annotation position in the response. The TopN is used to verify not only whether the algorithm returned correct results, but also how close these results are to the top positions.

$$TopN = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\alpha^{rank_i}}{\sum_{j=1}^{rank_i} \alpha^j}. \tag{1}$$

A dataset with manually annotated videos from Videoaula@RNP was created. The test dataset has 39 videos in Portuguese about areas such as computer science, statistics, chemistry and physics, with a total duration of approximately 6 hours. These videos were watched by experts invited to accomplish the manual annotation process. Each specialist assigned a DBpedia feature for each subject explicitly spoken during the video, without repetition[12]. There was no restriction on the number of resources for each video that the specialist could assign. During the process of creating the dataset, it was verified how expensive the manual annotation process is. For every 1 hour of video the experts took an average of 4 hours of manual labor, totaling approximately 24 hours to write down the entire base. We have evaluated the following data sources to choose which are the best data sources for semantic annotation:

Metadata: The texts included by the user, which includes the title, abstract and the keywords extracted from Videoaula@RNP.

Summary: Each video lecture has a summary that describes the topics that will be addressed throughout the lesson.

Speech Recognition: The audio was extracted and the automatic audio transcription was generated using the techniques discussed in Section 3.2.1.

Subtitle: If available, subtitles can be used instead of the automatic transcription. However, subtitles are a manual transcription adapted to better suit the reading. Subtitles are not always present in video repositories due to the high cost of production. This data source has been inserted into the experiments to simulate a speech recognition process with optimal word error rate.

Text Recognition: Text is often present in video lectures, recorded during the slide show, inserted in post-production or in video-related PDF files. OCR algorithms can be used to extract text from video frames.

Table 2 shows the results. Subtitles and speech recognition results demonstrate that the subtitle generates a low TopN, especially in the Entity Linking approach, because it generates a very large set of text and many words that were annotated were not among the entities of the video. In this case, the use of the summary or metadata is more appropriate. In the Topic Extraction approach, the combination of subtitles and transcription generates a high TopN because the Topic Extraction approach is more influenced by the frequency of words in the text. In the Entity Linking approach, the lack of the subtitles can be suppressed by using

[12] `https://github.com/ufjf-dcc/LAPIC1-benchmark`

another data source for higher recall. Like the Topic Extraction approach, there is a satisfactory recall when combining OCR, subtitles and automatic transcriptions.

Source	*Entity Linking*		*Topic Extraction*	
	Recall	TopN	Recall	TopN
M	0.214	0.102	0.071	0.076
M + O	0.611	0.041	0.286	0.132
M + Sb	0.838	0.019	0.410	0.171
M + Sm	0.304	**0.102**	0.132	0.118
M + T	0.614	0.019	0.287	0.160
M + O + Sb	0.838	0.013	0.432	0.334
M + O + Sm	0.630	0.041	0.291	0.129
M + O + T	0.713	0.017	0.351	0.165
M + Sb + Sm	0.838	0.019	0.410	0.185
M + Sb + T	0.838	0.017	0.387	0.162
M + Sm + T	0.656	0.020	0.305	0.161
M + O + Sb + Sm	0.838	0.013	0.432	**0.372**
M + O + Sb + T	0.838	0.012	0.454	0.337
M + O + Sm + T	**0.720**	0.017	**0.356**	**0.162**
M + Sb + Sm + T	0.838	0.016	0.387	0.172
M + O + Sb + Sm + T	0.838	0.012	**0.454**	**0.355**
O	0.568	0.042	0.281	0.109
O + Sb	0.838	0.013	0.432	0.303
O + Sm	0.587	0.042	0.285	0.120
O + T	0.688	0.017	0.347	0.145
O + Sb + Sm	0.838	0.013	0.432	0.362
O + Sb + T	0.838	0.012	0.454	0.327
O + Sm + T	**0.694**	**0.017**	**0.356**	**0.145**
O + Sb + Sm + T	0.838	0.012	0.454	0.344
Sb	**0.838**	**0.020**	0.387	0.156
Sb + Sm	0.838	0.019	0.387	0.194
Sb + T	0.838	0.017	0.387	0.151
Sb + Sm + T	0.838	0.017	0.387	0.172
Sm	0.166	**0.175**	0.098	0.098
Sm + T	**0.590**	0.019	**0.285**	**0.145**
T	**0.531**	**0.017**	**0.264**	**0.145**

Table 2. Recall and TopN using Entity Linking and Topic Extraction approaches. Read M as Metadata, O as OCR, Sb as Subtitle, Sm as Summary, and T as Transcription.

It is worth mentioning that any evaluation study is subject to the quality of the test data. Although the dataset used has been established by experts, it is possible that terms that have been annotated correctly by both approaches were not found on the dataset. The test dataset was created to evaluate how similar the result of the semantic annotation approaches was from manual annotations. Although other analyses can be performed to verify the accuracy of the experiments, the test dataset is adequate for the evaluation that was proposed.

It was possible to see how distinct sources of information can improve the semantic annotation process of educational videos, associating new information that was not previously present in these repositories. The new information represents the video content and would help to understand how the repository is semantically structured. Considering that manually created subtitles are not widely present in videos, automatic transcriptions are used as the main input for our automatic semantic annotation process.

3.2.3 Knowledge Graph

The issues in searching for a specific content as well as the lack of knowledge of the administrators about contents within the repository appear due to the increase of the videos in the repository and a low quality of the tags filled by the users who are disseminating the video. In this type of scenario, even if someone succeed in an initial search for some type of content, it is very difficult to find related content without having to perform a new search. On the other hand, several methods can be used to allow the user to browse the contents of the repository after the initial search.

Some authors propose the use of knowledge bases to help identify similarity between video lectures and other types of videos. In this case, it is common to use text associated with the video, such as titles, abstracts and other metadata, as well as captions or tags filled by users. For example, in [4] the authors manually explore the Wikipedia categories to find the best categories according to the user-created tags and video titles. Next, the Wikipedia categories are used to improve the video categorization.

After the Context Association task, each video is associated with a set of DBpedia resources, i.e., URIs that identify instances in the DBpedia ontology. We use the DBpedia graph of resources and categories in order to identify related videos through a similarity function. The encyclopedic content of DBpedia is suitable for calculating similarity in educational content repositories by having good coverage in the main teaching topics. The Wikipedia corpus covers different fields of knowledge and the organization of its category graph enables the linking of concepts belonging to different domains [19].

The Algorithm 1 is used to calculate the similarity of two videos and to create a relationship among the most similar videos. Let the video v_i be defined as an n-tuple $v_i = \langle r_1, r_2, \ldots, r_n \rangle$ where r_j are DBpedia resources, $j \in [1..n]$. The list C of categories of v_i is the union of the direct categories of each resource $r \in v_i$, the broader (α) and the more specific (β) categories of each direct category of r. We say that the videos v_i and v_j will be related if $sim\left(C_{v_i}, C_{v_j}\right) > \omega$, where $\omega \in [0, 1]$ is a predefined constant.

The similarity between two videos can be calculated as a generic function sim. The Sorensen-Dice coefficient calculation was used as similarity function. This

Algorithm 1: Algorithm for relation prediction

```
Input  : $\omega$, $\alpha$, $\beta$
Output: Set $R$ of related videos
begin
    $R \leftarrow \emptyset$
    for each video $v_i \in$ repository do
        $C_{v_i} \leftarrow \emptyset$
        for each $r \in v_i$ do
            $c \leftarrow$ getDBpediaCategories($r$)
            $C_{v_i} \leftarrow C_{v_i} \bigcup^{\alpha,\beta} c$
        end
    end
    for each videos $v_i, v_j \in$ repository do
        $List_{v_i} = \emptyset$
        if $sim\left(C_{v_i}, C_{v_j}\right) > \omega$ then
            append($List_{v_i}, v_j$)
        end
        $R \leftarrow R \bigcup List_{v_i}$
    end
    return $R$
end
```

method can be seen in the formula below where Ω represents the percentage of related categories between two videos v_i and v_j.

$$\Omega = \frac{2(|C_{v_i}| \bigcap |C_{v_j}|)}{|C_{v_i}| + |C_{v_j}|}.$$

The relationships between videos, resources and categories are represented in a simplified way by the flowchart in Figure 2. Videos are linked to DBpedia resources by the Contextual Association step, and these resources are associated with categories in the DBpedia graph. Therefore, the similarity is calculated not only by the number of resources that the videos share, but by the number of categories associated with those resources that are gathered after a walk in the graph.

A test dataset was created with manual relationships defined by experts in the area of Exact Sciences in order to evaluate the approach. This version of the dataset is available to other researchers[13] and contains the same set of videos used in the previous section. The experts were free to define relationships between videos. An expert might consider that a video is related by addressing exactly the same topic of a video, but another expert might relate videos considering that they contain complementary information. Altogether, 211 relationships were defined manually,

[13] `https://github.com/ufjf-dcc/LAPIC1-benchmark`

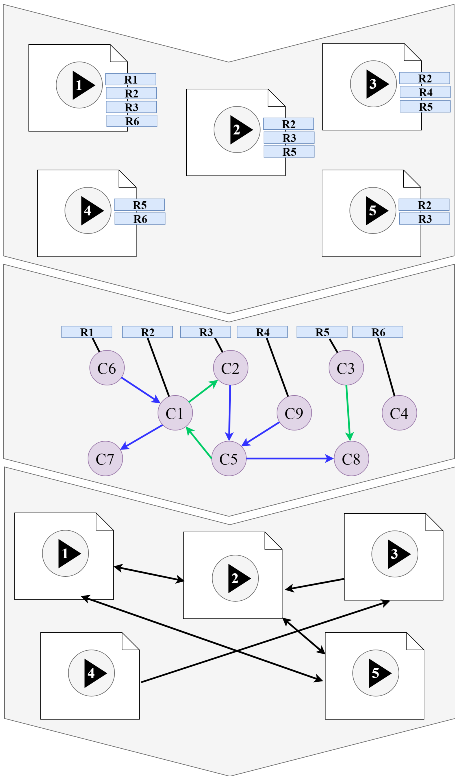
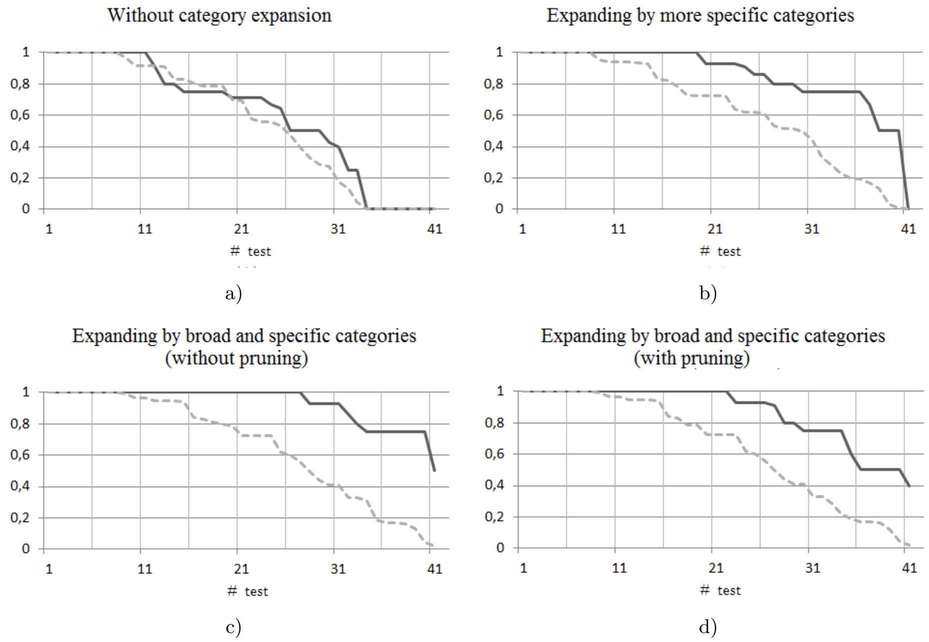


Figure 2. Similarity flowchart

with a mean of 5 relationships per video. There is no video without relations. There are 10 videos with only 1 relationship, and 17 videos with at least 5 relationships.

The approach is expected to find new relationships beyond the existing ones. In the experiments, some combinations of α and β parameters of the algorithm were tested in order to retrieve different levels of DBpedia category information in each of the experiments and to analyze how the amount of information influences the evaluation metrics.

Figure 3 presents the recall and TopN (Y-axis) for each video (or test) in dataset (X-axis). The solid line represents recall values and the dashed line represents TopN values for each experiment. To identify the proportion of videos with good and bad results in each experiment, the videos were sorted by TopN in descending order. A desirable but intangible algorithm that always finds the correct videos would have its constant curves at 1. Figures 3 a), 3 b) and 3 c) show that the recall increases as more information are processed by the algorithm. For instance, Figure 3 a) shows that 9 videos reached a maximum recall (videos numbered from 1 to 9), while Figure 3 b) shows that the maximum recall was reached in 20 videos. It is verified that walking in broader (α) and more specific (β) categories produces high recall. The recall does not change with greater depths according to the obtained results. Although the use of the categories can help in a high TopN, increasing the depth does not imply in higher TopN. By increasing the depth, more common categories of videos will be used, and it will be more difficult to rank the result correctly through the number of categories in common.



a) b) c) d)

Figure 3. Recall (solid line) and TopN (dashed line) for each experiment

The best configuration found for the parameters was $\alpha = 1$ and $\beta = 1$ (Figure 3 c)), resulting in 32 videos with a recall of approximately 0.9. According to the TopN, the algorithm was able to return the correct videos in the first results in more than half of the dataset. It is also possible to verify that the TopN follows the trend of generating better results as more information is processed by the algorithm. The use of specific categories (Figure 3 b)) presents a better result in relation to the non-expansion approach (Figure 3 a)). Since the list of related videos is ranked for each video, a threshold can be used to limit the set of videos in the result set. Figure 3 d) shows the results of the third experiment retrieving only videos that have at least 10 categories in common. This pruning method did not influence considerably the result of the algorithm. The mean TopN increased from 0.67732 to 0.688215 and the mean recall reduced from 0.93120 to 0.87715.

Figure 4 presents the dispersion of the recall and TopN for the best experiment ($\alpha = \beta = 1$). Each point represents the recall and TopN values of a specific test.

The number of points is equal to the number of tests in Figure 3 c) (X-axis). It can be observed that in the same recall value, the TopN values can vary by up to 20 percentage points. There is a trend in Figure 4, where the higher the recall, the better the ranking of the videos.

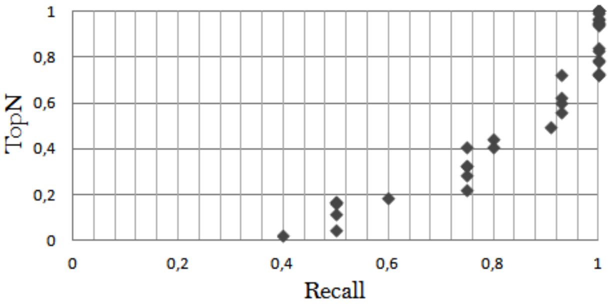


Figure 4. Dispersion of the recall and TopN in the experiment with pruning

Although the results of the experiments were satisfactory, it is important to analyze the false positives. The dataset was created manually by experts following personal criteria to determine which videos should be related. As a result, the test dataset contains videos with few relationships or videos related to others that do not have any resources in common. Thus, the test dataset presents relationships and resources that are not skewed by some set of information or selection methods. Taking into account these particularities, our experiments showed that some false positive relationships that have a large number of categories in common are, in fact, relationships that are absent on the test dataset. That is, the algorithm is able to identify relationships that are not always easily identifiable by a person. Take as an example the video identified as "fis2tempcalor", which addresses concepts of temperature and heat. This video was manually related only to the video identified as "fis2cap18-part2". The algorithm, in turn, related 15 videos with "fis2tempcalor", among them videos about physics and chemistry that address concepts indirectly related to temperature and heat.

The experiments presented in these three sections were carried out with the purpose of demonstrating the feasibility of the proposal. In the following section, we discuss how the approach can be applied to find the main topics in a real video lectures repository.

4 RESULTS AND DISCUSSIONS

In this section, we discuss how the proposed framework can be used in a real video repository. This dataset is composed of 93 randomly selected video lectures from VideoAula@RNP (about 11 % of the repository at the time of the experiment), totaling 3 604 minutes of videos. This repository was created by RNP to make available video lectures produced by associated educational institutions from Brazil.

The videos were transcribed with our ASR system. In the semantic annotation process, each video received up to 5 semantic annotations. In our experiments, the top 5 ranked annotations are enough to generate metadata with high precision. The topic extraction approach was chosen for the Context Association because of the results discussed in Section 3.2.2. Then the knowledge graph was created with $\alpha = 1$ and $\beta = 1$ and no threshold (Section 3.2.3). As a result, each vertex is linked to all others and the edges represent the degree of relationship of the videos (vertices). Figure 5 shows an example of the undirected complete graph. Blue vertices represent a video and green vertices are the categories of the video. The categories related to the videos were extracted from the DBpedia resources automatically annotated in each video and the final graph does not contain the DBpedia resources. The video v3 has a strong relationship with the video v2 because of the number of categories in common. On the other hand, the video v3 shares a few categories with the video v4 and the edge between them has a low value.

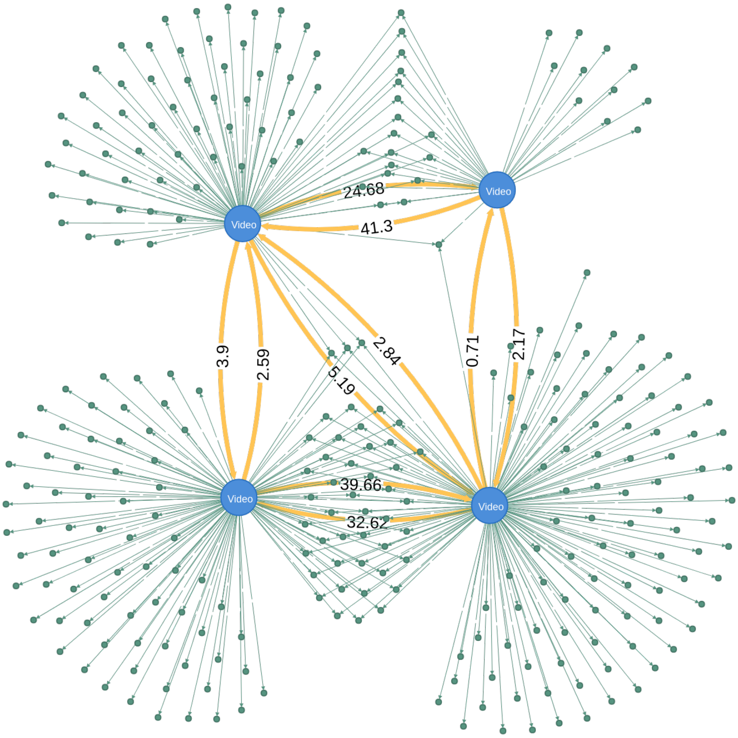


Figure 5. Video relations subgraph. Blue vertices represent videos and green vertices represent categories. Yellow edges link two videos and are weighted. Green edges are unweighted and link one video to one category.

The graph is used to analyze what kinds of content has been produced by users of the repository. The first step consists of identifying communities in that knowledge graph. In this case, communities are clusters of videos that are densely connected internally, that is, groups of videos that share a large number of categories. Therefore, by identifying these communities, we are finding groups of videos that potentially address related subjects.

For this process, the graph was submitted to the Label Propagation Algorithm (LPA), an algorithm for community detection on graphs networks that works by exploring the neighborhoods between the vertices [37]. The algorithm uses network structure alone as its guide, and does not require a pre-defined objective function or prior information about the communities. The algorithm sets a unique label for each vertex. At every iteration of propagation, each vertex updates its label to the one that the maximum numbers of its neighbors belong to. Ties are broken uniformly and randomly. The algorithm reaches convergence when each vertex has the majority label of its neighbors. Figure 6 shows the groups of videos identified by the Label Propagation Algorithm. The graph contains 22 groups of videos with very close subjects. The groups are identified by different colors. The vertices representing categories were removed from the graph for a better view.

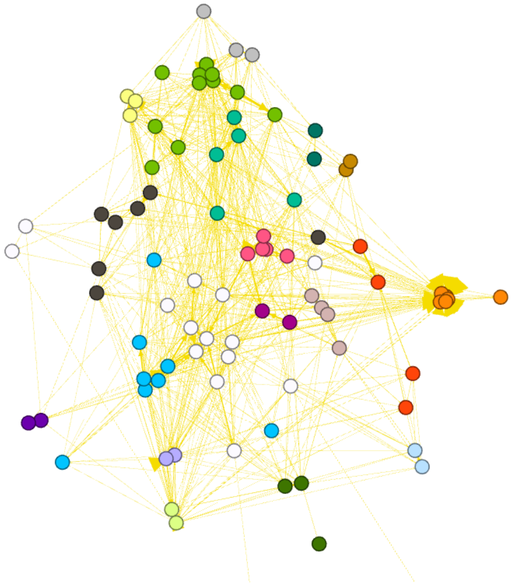


Figure 6. Knowledge graph after the label propagation algorithm

In the second step, we have performed an analysis in the generated groups. For this task, we perform a new search in the graph of categories from DBpedia. Given a set of input resources T, we return a percentage of contribution of each Wikipedia main topic in T. The set of resources T of the group c contains all DBpedia resources of each video automatically annotated of each video in group c. Wikipedia main topics can be seen as end vertices in the DBpedia category graph.

To understand the main topics addressed in each group from the first step, we performed a search in the graph, passing as input the set of resources and walking towards the top of the graph by all the shortest paths between the categories of each resource and the main topics, as proposed in [27]. As a result, a fingerprint is created for each group. A fingerprint is a vector of weights where each dimension represents the weight of that DBpedia main topic. Nowadays, the number of main topics in DBpedia is 19. As an example, the categories used as input for groups 4, 6 and 17 are presented in Table 3. The groups contain 5, 16, and 7 videos, respectively. With the more frequent categories of each group, it is possible to see the difference between the subjects of each group. Group 4 addresses genetics and chemistry topics while Group 6 addresses topics related to sociology, philosophy and law. Group 17 presents subjects focused on artificial intelligence and cryptography.

Group		
4	6	17
Inorganic carbon compounds Alcohols Persistent organic pollutants Genetics by type of organism DNA repair Mitochondrial genetics	Social epistemology Social philosophy Philosophy of education Theories of law Rights	Polynomial-time problems Artificial intelligence Turing tests Cryptographic hardware Computer security software Internet fraud Internet search algorithms

Table 3. Example of input resources of groups 4, 6, and 17

The results of this step can be seen in Figure 7. The topics of technology, society, geography, culture, and history are present in most of the videos. Religion, life, law, and arts are an example of topics that are not addressed in this videos. Some groups have very similar fingerprints, such as Groups 11 and 15. Although the groups contain videos on history, they are distinct groups because of the difference of the semantic annotations in each group. In other words, the groups encompass videos about history, but the groups have few direct categories in common, addressing distinct subjects in this area of knowledge. In fact, Group 11 contains two videos on political theories and Group 15 contains four videos on wars of independence.

It is possible to see the relationship between Tables 3 and 7. For example, when analyzed, the categories for the Group 6 presented in Table 3 are found to be focused in sociology, philosophy and law. In Figure 7, the Group 6 has a greater weight for the topics of philosophy, presenting 35,65 % of relationships with philosophy and 14.78 % with society. The same occurs in the other groups: in the Group 4 the categories encompass biology and chemistry and the main topics classifications are matter (50 %), health (18.75 %) and nature (12.5 %), for the Group 17 the categories are focused on artificial intelligence and cryptography. The main topics are sciences and technologies (36.95 %) and mathematics (27.17 %).

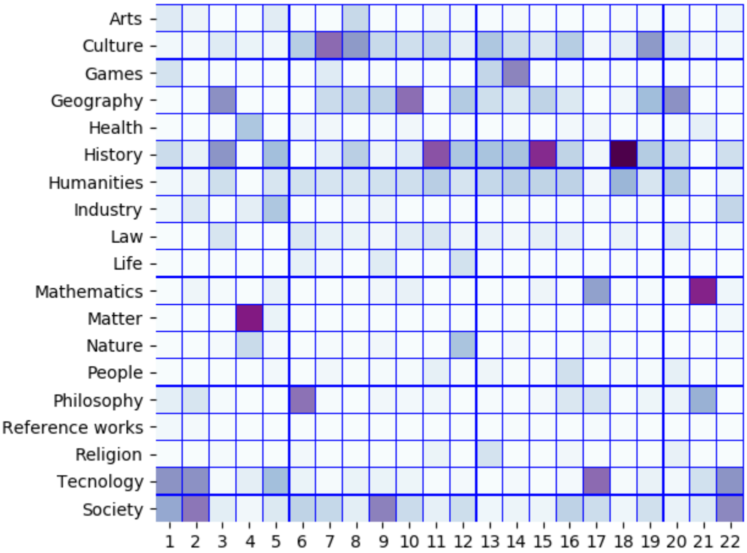


Figure 7. Fingerprint of the groups. The X-axis contains the group numbers and the y-axis contains the Wikipedia main topics.

It is possible to understand the information of the groups automatically, and it can be used to recommend to the user a set of videos that are related in a higher level. Although some topics are classified out of line, as is the case of Group 17 that presents weight in culture (2.17 %), nature (3.2 %) and philosophy (9.78 %), a threshold can be used to filter the results by discarding the lower values.

5 CONCLUDING REMARKS

In this paper, we proposed a framework for knowledge discovering in video lectures repositories. This framework is composed of three main stages: content processing through ASR, context association using semantic annotation and the construction of a knowledge graph through a walk in the DBpedia ontology and similarity calculations. Each part of the framework was evaluated separately and, in the end, the final result of the complete process was used to demonstrate the applicability of the framework in a real repository of video lectures.

As the main contribution of this work, we highlight the applicability of our proposal in video lectures repositories where there is a lack of metadata to describe the videos. As explained throughout the text, this lack of metadata hampers indexing systems, which makes search and recommendation in these repositories very ineffective. Our proposal can automatically extract knowledge from video lectures and can improve several applications in large repositories, such as searching, recommendation, and advertising systems. The advances in automatic speech recognition systems [8, 2, 23] allow developers to easily reproduce these results using well-known ASR tools. The extracted metadata can be used for indexing or clustering, which are everyday tasks in information retrieval and recommendation systems. Our experiments have shown that discovering knowledge allows determining the subjects and relationships of video lectures. Other contributions were the analyses made at each stage of the framework, where it was possible to raise the main points, challenges and solutions for the development of our proposal.

Future work includes analyzing the scope of other knowledge bases in the framework. Other knowledge bases make it possible to find different, more specific resources if a domain base is used. We also intend to study a cross-domain approach with multiple knowledge bases simultaneously. Finally, we also envisage the creation of an interface for navigation in the information found in the repository.

6 ACKNOWLEDGEMENTS

We thank RNP (Brazilian National Education and Research Network) and CAPES (Coordination for the Improvement of Higher Education Personnel) for the financing of this work.

REFERENCES

[1] Asghar, M. N.—Hussain, F.—Manton, R.: Video Indexing: A Survey. International Journal of Computer and Information Technology, Vol. 3, 2014, No. 1, pp. 148–169.

[2] Bahar, P.—Bieschke, T.—Ney, H.: A Comparative Study on End-to-End Speech to Text Translation. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 2019. arXiv preprint arXiv:1911.08870.

[3] Che, X.—Luo, S.—Yang, H.—Meinel, C.: Automatic Lecture Subtitle Generation and How It Helps. 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT 2017), 2017, pp. 34–38, doi: 10.1109/ICALT.2017.11.

[4] Chen, Z.—Cao, J.—Song, Y.—Zhang, Y.—Li, J.: Web Video Categorization Based on Wikipedia Categories and Content-Duplicated Open Resources. Proceedings of the 18th ACM International Conference on Multimedia (MM '10), 2010, pp. 1107–1110, doi: 10.1145/1873951.1874162.

[5] Dasiopoulou, S.—Giannakidou, E.—Litos, G.—Malasioti, P.—Kompatsiaris, Y.: A Survey of Semantic Image and Video Annotation Tools. In: Paliouras, G., Spyropoulos, C. D., Tsatsaronis, G. (Eds.): Knowledge-Driven Multimedia Information Extraction and Ontology Evolution. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 6050, 2011, pp. 196–239, doi: 10.1007/978-3-642-20795-2_8.

[6] Davis, D.—Chen, G.—Hauff, C.—Houben, G.-J.: Gauging Mooc Learners' Adherence to the Designed Learning Path. Proceedings of the 9th International Conference on Educational Data Mining (EDM), 2016, pp. 54–61.

[7] De Medio, C.—Limongelli, C.—Sciarrone, F.—Temperini, M.: MoodleREC: A Recommendation System for Creating Courses Using the Moodle E-Learning Platform. Computers in Human Behavior, Vol. 104, 2020, pp. 106–168, doi: 10.1016/j.chb.2019.106168.

[8] Drugman, T.—Pylkkönen, J.—Kneser, R.: Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models. Interspeech 2016, ISCA, 2016, pp. 2318–2322, arXiv Preprint arXiv:1903.02852, doi: 10.21437/Interspeech.2016-1382.

[9] Farrell, R. G.—Liburd, S. D.—Thomas, J. C.: Dynamic Assembly of Learning Objects. Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers and Posters, 2004, ACM, pp. 162–169, doi: 10.1145/1013367.1013394.

[10] Franzoni, V.—Tasso, S.—Pallottelli, S.—Perri, D.: Sharing Linkable Learning Objects with the Use of Metadata and a Taxonomy Assistant for Categorization. In: Misra, S. et al. (Eds.): Computational Science and Its Applications – ICCSA 2019. Springer, Cham, Lecture Notes in Computer Science, Vol. 11620, 2019, pp. 336–348, doi: 10.1007/978-3-030-24296-1_28.

[11] Galanopoulos, D.—Mezaris, V.: Temporal Lecture Video Fragmentation Using Word Embeddings. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W. H., Vrochidis, S. (Eds.): MultiMedia Modeling (MMM 2019). Springer, Cham, Lecture Notes in Computer Science, Vol. 11296, 2019, pp. 254–265, doi: 10.1007/978-3-030-05716-9_21.

[12] Gaona-García, P. A.—Martin-Moncunill, D.—Montenegro-Marin, C. E.: Trends and Challenges of Visual Search Interfaces in Digital Libraries and Repositories. The Electronic Library, Vol. 35, 2017, No. 1, pp. 69–98, doi: 10.1108/EL-03-2015-0046.

[13] González-Pérez, L. I.—Ramírez-Montoya, M.-S.—García-Peñalvo, F. J.: User Experience in Institutional Repositories: A Systematic Literature Review. International Journal of Human Capital and Information Technology Professionals (IJHC-ITP), Vol. 9, 2018, No. 1, pp. 70–86, doi: 10.4018/IJHCITP.2018010105.

[14] Gruhn, R. E.—Minker, W.—Nakamura, S.: Statistical Pronunciation Modeling for Non-Native Speech Processing. Springer Science and Business Media, 2011, doi: 10.1007/978-3-642-19586-0.

[15] Gupta, Y.—Saini, A.—Saxena, A. K.: A New Fuzzy Logic Based Ranking Function for Efficient Information Retrieval System. Expert Systems with Applications, Vol. 42, 2015, No. 3, pp. 1223–1234, doi: 10.1016/j.eswa.2014.09.009.

[16] Gurbuz, F.: Students' Views on Distance Learning in Turkey: An Example of Anadolu University Open Education Faculty. Turkish Online Journal of Distance Education, Vol. 15, 2014, No. 2, pp. 239–250, doi: 10.17718/tojde.54964.

[17] Habibian, A.—Mensink, T.—Snoek, C. G. M.: Discovering Semantic Vocabularies for Cross-Media Retrieval. Proceedings of the 5th ACM Interna-

tional Conference on Multimedia Retrieval (ICMR '15), 2015, pp. 131–138, doi: 10.1145/2671188.2749403.

[18] Hadian, H.—Sameti, H.—Povey, D.—Khudanpur, S.: Flat-Start Single-Stage Discriminatively Trained HMM-Based Models for ASR. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 26, 2018, pp. 1949–1961, doi: 10.1109/TASLP.2018.2848701.

[19] Halavais, A.—Lackaff, D.: An Analysis of Topical Coverage of Wikipedia. Journal of Computer-Mediated Communication, Vol. 13, 2008, No. 2, pp. 429–440, doi: 10.1111/j.1083-6101.2008.00403.x.

[20] Harasim, L.: Shift Happens: Online Education as a New Paradigm in Learning. The Internet and Higher Education, Vol. 3, 2000, No. 1-2, pp. 41–61, doi: 10.1016/S1096-7516(00)00032-4.

[21] Hauptmann, A. G.—Jin, R.—Ng, T. D.: Video Retrieval Using Speech and Image Information. Storage and Retrieval for Media Databases 2003, Proceedings of the SPIE, Vol. 5021, 2003, pp. 148–160, doi: 10.1117/12.479747.

[22] Heath, T.—Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 1, 2011, No. 1, pp. 1–136, doi: 10.2200/S00334ED1V01Y201102WBE001.

[23] Herchonvicz, A. L.—Franco, C. R.—Jasinski, M. G.: A Comparison of Cloud-Based Speech Recognition Engines. Proceedings of the Computer on the Beach, 2019, pp. 366–375.

[24] Hinton, G.—Deng, L.—Yu, D.—Dahl, G. E.—Mohamed, A.-R.—Jaitly, N.—Senior, A.—Vanhoucke, V.—Nguyen, P.—Sainath, T. N.—Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, Vol. 29, 2012, No. 6, pp. 82–97, doi: 10.1109/MSP.2012.2205597.

[25] Kota, B. U.—Davila, K.—Stone, A.—Setlur, S.—Govindaraju, V.: Generalized Framework for Summarization of Fixed-Camera Lecture Videos by Detecting and Binarizing Handwritten Content. International Journal on Document Analysis and Recognition (IJDAR), Vol. 22, 2019, pp. 221–233, doi: 10.1007/s10032-019-00327-y.

[26] Lei, W.—Qing, F.—Zhou, J.: Improved Personalized Recommendation Based on Causal Association Rule and Collaborative Filtering. International Journal of Distance Education Technologies (IJDET), Vol. 14, 2016, No. 3, pp. 21–33, doi: 10.4018/IJDET.2016070102.

[27] Medeiros, J. F.—Nunes, B. P.—Siqueira, S. W. M.—Leme, L. A. P. P.: TagTheWeb: Using Wikipedia Categories to Automatically Categorize Resources on the Web. In: Gangemi, A. et al. (Ed.): The Semantic Web: ESWC 2018 Satellite Events (ESWC 2018). Springer, Cham, Lecture Notes in Computer Science, Vol. 11155. 2018, pp. 153–157, doi: 10.1007/978-3-319-98192-5_29.

[28] Mendes, P. N.—Jakob, M.—García-Silva, A.—Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11), ACM, 2011, pp. 1–8, doi: 10.1145/2063518.2063519.

[29] Mohan, P.—Brooks, C.: Learning Objects on the Semantic Web. Proceedings of the 3$^{\text{rd}}$ IEEE International Conference on Advanced Learning Technologies, 2003, pp. 195–199, doi: 10.1109/ICALT.2003.1215055.

[30] Neves, D. E.—Brandão, W. C.—Ishitani, L.: Automatic Content Recommendation and Aggregation According to SCORM. Informatics in Education, Vol. 16, 2017, No. 2, pp. 225–256, doi: 10.15388/infedu.2017.12.

[31] Ney, H.—Essen, U.: On Smoothing Techniques for Bigram-Based Natural Language Modelling. 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP '91), IEEE, 1991, pp. 825–828, doi: 10.1109/ICASSP.1991.150464.

[32] Oliveira, A. L. C.—Silva, E. S.—Macedo, H. T.—Matos, L. N.: Brazilian Portuguese Speech-Driven Answering System. Proceedings of the 6$^{\text{th}}$ Euro American Conference on Telematics and Mation Systems (EATIS '12), ACM, 2012, pp. 277–284, doi: 10.1145/2261605.2261647.

[33] Omheni, N.—Kalboussi, A.—Mazhoud, O.—Kacem, A. H.: Recognition of Learner's Personality Traits Through Digital Annotations in Distance Learning. International Journal of Distance Education Technologies (IJDET), Vol. 15, 2017, No. 1, pp. 28–51, doi: 10.4018/IJDET.2017010103.

[34] Piedra, N.—Chicaiza, J. A.—López, J.—Tovar, E.: An Architecture Based on Linked Data Technologies for the Integration and Reuse of OER in MOOCs Context. Open Praxis, Vol. 6, 2014, No. 2, pp. 171–187, doi: 10.5944/openpraxis.6.2.122.

[35] Purarjomandlangrudi, A.—Chen, D.—Nguyen, A.: Investigating the Drivers of Student Interaction and Engagement in Online Courses: A Study of State-of-the-Art. Informatics in Education, Vol. 15, 2016, No. 2, pp. 269–286, doi: 10.15388/infedu.2016.14.

[36] Qazi, A.—Goudar, R. H.: Emerging Trends in Reducing Semantic Gap Towards Multimedia Access: A Comprehensive Survey. Indian Journal of Science and Technology, Vol. 9, 2016, No. 30, doi: 10.17485/ijst/2016/v9i30/99072.

[37] Raghavan, U. N.—Albert, R.—Kumara, S.: Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. Physical Review E, Vol. 76, 2007, No. 3, Art. No. 036106, pp. 36–106, doi: 10.1103/PhysRevE.76.036106.

[38] Raimond, Y.—Lowis, C.: Automated Interlinking of Speech Radio Archives. Linked Data on the Web (LDOW 2012) Workshop at WWW 2012, CEUR Workshop Proceedings, Vol. 937, 2012.

[39] Rao, K.—Peng, F.—Sak, H.—Beaufays, F.: Grapheme-to-Phoneme Conversion Using Long Short-Term Memory Recurrent Neural Networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4225–4229, doi: 10.1109/ICASSP.2015.7178767.

[40] Rojas-Contreras, M.—Portilla-Jaimes, O.: Integration of Learning Objects for Adaptative Learning. IOP Conference Series: Materials Science and Engineering, IOP Publishing, Vol. 519, 2019, Art. No. 012030, doi: 10.1088/1757-899X/519/1/012030.

[41] Sad, S. N.—Goktas, O.—Bayrak, I.: A Comparison of Student Views on Web-Based and Face-to-Face Higher Education. Turkish Online Journal of Distance Education, Vol. 15, 2014, No. 2, pp. 209–226, doi: 10.17718/tojde.02246.

[42] van Merriënboer, J. J. G.—Ayres, P.: Research on Cognitive Load Theory and Its Design Implications for E-Learning. Educational Technology Research and Development, Vol. 53, 2005, No. 3, pp. 5–13, doi: 10.1007/BF02504793.

[43] Vanajakshi, P.—Mathivanan, M.: A Detailed Survey on Large Vocabulary Continuous Speech Recognition Techniques. 2017 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2017, 7 pp., doi: 10.1109/ICCCI.2017.8117755.

[44] Yang, H.—Meinel, C.: Content Based Lecture Video Retrieval Using Speech and Video Text Information. IEEE Transactions on Learning Technologies, Vol. 7, 2014, No. 2, pp. 142–154, doi: 10.1109/TLT.2014.2307305.

[45] Zhao, B.—Xu, S.—Lin, S.—Luo, X.—Duan, L.: A New Visual Navigation System for Exploring Biomedical Open Educational Resource (OER) Videos. Journal of the American Medical Informatics Association, Vol. 23, 2016, No. e1, pp. e34–e41, doi: 10.1093/jamia/ocv123.

Jorão Gomes Jr. holds a degree in exact sciences from Federal University of Juiz de Fora (UFJF) and is currently pursuing a master's degree from UFJF. He has experience in the area of computer science, with emphasis on information retrieval and natural language processing.

Laura Lima Dias holds her Master's degree in computer science from Federal University of Juiz de Fora (UFJF) (2017). She is researcher at the Laboratory of Applications and Innovation in Computing (LApIC). She has experience in multimedia systems, information retrieval, semantic annotation and informatics in education.

Eduardo Rocha Soares holds his Bachelor's degree in computer science and is currently pursuing his Master's degree from UFJF, also in computer science. He is interested in the area of multimedia systems, with an emphasis on content and natural language processing.

Eduardo Barrere is Professor at Federal University of Juiz de Fora (UFJF), Brazil. He has Ph.D. degree in systems engineering and computing since 2007. He coordinates the LApIC Laboratory (Applications and Innovation in Computing).

Jairo Francisco de Souza is Assistant Professor at the Department of Computer Science, Federal University of the Juiz de Fora (UFJF), Brazil, where he teaches courses in databases and information retrieval. He holds his M.Sc. (2007) from the Federal University of Rio de Janeiro (UFRJ) and his Ph.D. from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil, both in computer science area. His research interests include knowledge representation, natural language processing, information integration, and semantic models.