Pertanika J. Sci. & Technol. 26 (4): 1571 - 1590 (2018)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Automatic Annotation of Educational Videos for Enhancing Information Retrieval

Poornima, N.* and Saleena, B.

School of Computing Science and Engineering, VIT Chennai, Tamil Nadu - 600 127, India

ABSTRACT

Educational videos are one of the best means of imparting knowledge to the users/learners. Videos can convey information in an effective and interesting manner. These videos can be accessed through online or from stored repositories using queries. Search queries play important role in the retrieval. Whenever a user gives an ambiguous query, the search engine may produce irrelevant results. Thus a lot of time is being spent by the users in retrieving the relevant videos. In order to improve the probability of retrieving relevant results, semantic web technologies are applied. This paper aims to extract keywords from the videos and to find the association between the extracted terms. The associated terms are arranged based on their frequency of occurrences. These terms are used to annotate the video automatically, which in turn improves the retrieval of more relevant videos. An ontology is created by experts based on the e-learning video domain. Videos are grouped based on the keywords and on domain ontology, which also helps in enhancing the retrieval results. Videos containing text are only considered for processing.

Keywords: Annotation, e-learning, ontology, semantics, Term Frequency Inverse Document Frequency (TF-IDF), video retrieval

INTRODUCTION

Use of technology and internet in the modern educational system makes learning more interactive and interesting. Nowadays universities and educational institutions follow various techniques to improve the student's learning skills, for example: flipped classroom, collaborative learning, differential learning and virtual classrooms. In this modern educational system, study

ARTICLE INFO

Article history:
Received: 07 August 2017
Accepted: 27 June 2018
Published: October 2018

E-mail addresses:
poornima.n2014@vit.ac.in (Poornima, N.)
saleena.b@vit.ac.in (Saleena, B.)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

materials are delivered to the users in various modes such as document, presentation slides, audios or videos. These materials can be delivered directly to the users or it can be stored in a repository, so that users can access materials through querying.

After the invention of video sharing websites, educational videos have become popular among learners. A survey showed that there was tremendous increase in the amount of educational videos uploaded in YouTube (Che & Lin, 2015). YouTube has introduced a separate channel YouTube EDU for learners and educators. Large number of educational videos makes the search process tedious and time consuming. In case of text document search, entire document is analysed and results are displayed based on the user query. Unlike text document search, videos are retrieved only based on the annotation given to the video without any analysis on the video contents. In general, videos are annotated manually by the authors of video. Manual annotation consumes more time and there is also a possibility of giving improper keyword annotation, just to attract the users.

The central goal of this research is to automate the annotation process by analysing the video contents and to provide semantic meaning to the keywords to enhance the information retrieval. Text from the videos are captured using Tesseract Optical Character Recognition (OCR) which helps in analysis of the video contents. Using WordNet, semantic meanings and relationships of extracted words are found. Term frequencies for all the extracted words from video and WordNet are identified and arranged based on frequent occurrence. In addition, domain ontologies are created for all the categories of videos available in database. According to domain ontologies, highly occurring words are grouped along with its videos. If the user query is related to any of the video group, then the entire group will be retrieved.

Major contributions of the paper include:

- Automatic video content Extraction: Visual analysis techniques such as keyframe extraction and text extraction from video helps in analysing the video contents. Textual features are identified and extracted using Video OCR technique.
- Filling semantic gap between query and retrieved videos: Keywords are identified from the videos. The related terms of the keywords are retrieved using WordNet. The keywords along with the related terms and the videos were clustered based on the domain ontology, which was helpful in improving the retrieval results.

Section 2 discusses some of the open source coursewares and its major issues during retrieval. Section 3 discusses the related work carried out in this field. Section 4 describes the methodology for information retrieval using semantic web technologies. Section 5 discusses experimentation setup, results and performance evaluation metrics. Finally section 6 concludes with future directions for this research.

Learning Issues in E-Learning Coursewares

Many open source coursewares by major universities allow students to watch and download course materials anytime and anywhere from the world. Massachusetts Institute of Technology: MIT (MITOPENCOURSEWARE), Harvard University (Harvard Open Learning Initiative) and University of California (UC Irvine, OpenCourseWare) are some of the open coursewares provided by popular universities.

Population of engineering students in India outnumbers ("NPTEL Frequently Asked Questions", 2018) every other country. NPTEL is a curriculum building exercise aims to create open source contents for major science and engineering courses. NPTEL projects are funded by Government of India and it is used by most of the Indian engineering students. NPTEL website does not allow keyword based search. Search on NPTEL website can be done only based on the course name and professor name. NPTEL courses can be filtered by discipline, content type (like subject) and institutions (Figure 1). Searching a topic "Properties of transaction" from NPTEL responds with no result (Figure 1). Database Design course video contains content on "Transaction Properties" (Figure 2). Even Database Design course syllabus contains topic "Properties of transactions" (Figure 2).

In case of MITOPENCOURSEWARE, users can browse through a topic by selecting a topic, subtopic and specialty (Figure 3). After choosing from the list of topic, sub topic and specialty, results are given. Results for the topic "Data mining" from MITOPENCOURSEWARE are shown in figure (Figure 3).

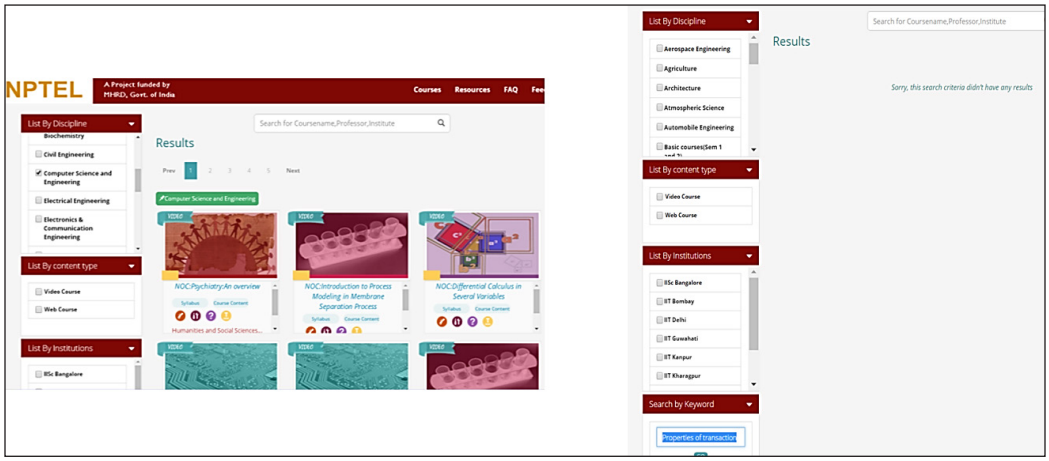

Figure 1. Screenshot showing different search facilities available in NPTEL (Selection of Discipline -> Civil Engineering, Computer Science and Engineering, Electronics & Communication Engineering and so on, content type -> Video Course and Web Course and Institutions -> IISC Bangalore, IIT Kanpur, IIT Madras and so on) and Result of searching topic "Properties of Transaction" through Keyword Search option in NPTEL

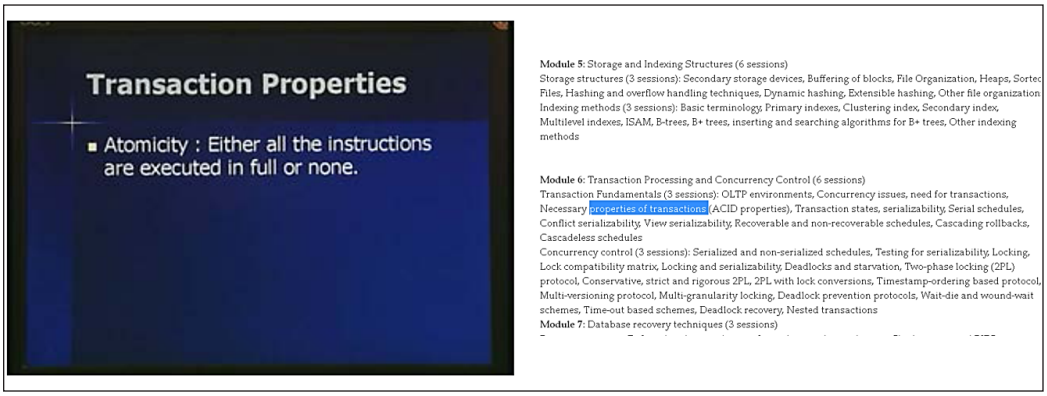

Figure 2. Syllabus and Video frame of NPTEL course 'Database Design' showing 'Transaction Properties' in it

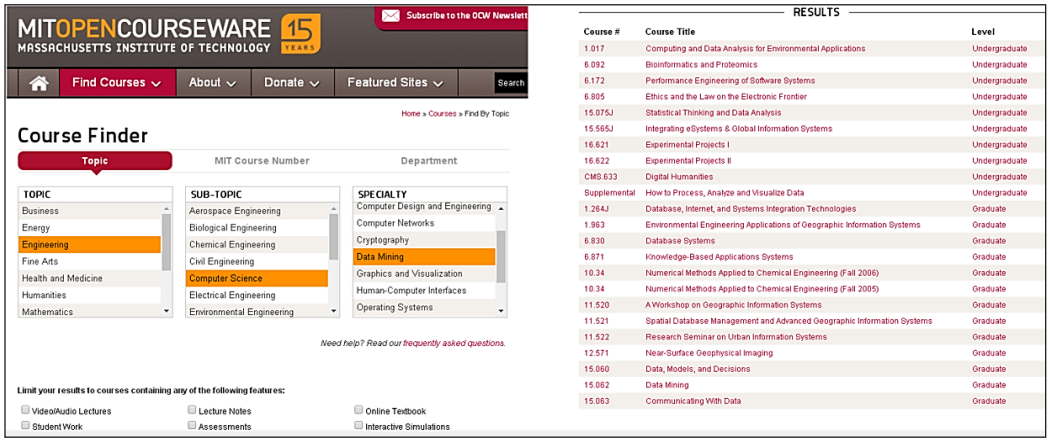

Figure 3. Course Finder of MITOPENCOURSEWARE (showing selection of Topic, Sub-Topic and Speciality) and Search result of course 'Data Mining' in MITCOURSEWARE provides irrelevant course results

Similar challenges are also encountered in Harvard Open Learning Initiative and UC Irvine open coursewares. Search by instructor, keyword and course are the possible ways for searching in Harvard Open Learning Initiative. Keyword search results are not more relevant to the search query (Figure 4). Even though University of California (UC Irvine, OpenCourseWare) allows you to search the content through keywords, search results are completely irrelevant to the user query (Figure 5). Search through type and categories are also possible in UC Irvine OpenCourseWare.

In the above open coursewares, retrievals are inefficient because videos are not analysed based on the contents. Learner can fetch the materials only based on limited options such as selection through department/school, category/subject, topic/subtopic and so on. Videos are mostly annotated with random keywords which are totally inappropriate and leads

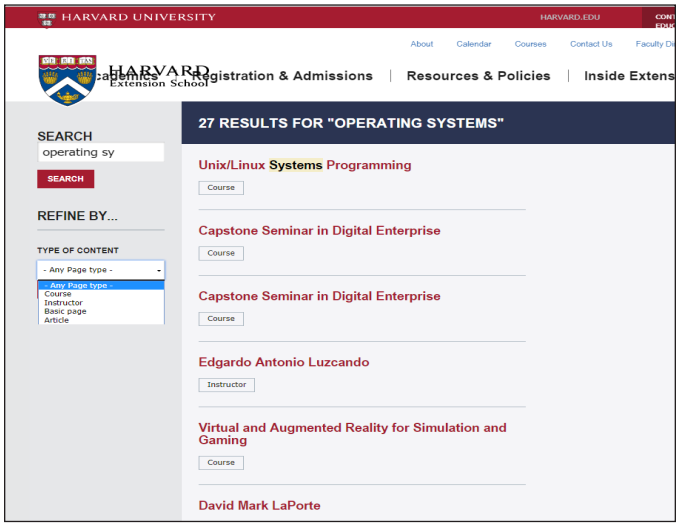

Figure 4. Search result of course 'Operating Systems' in Harvard Open Learning Initiative showing irrelevant course results

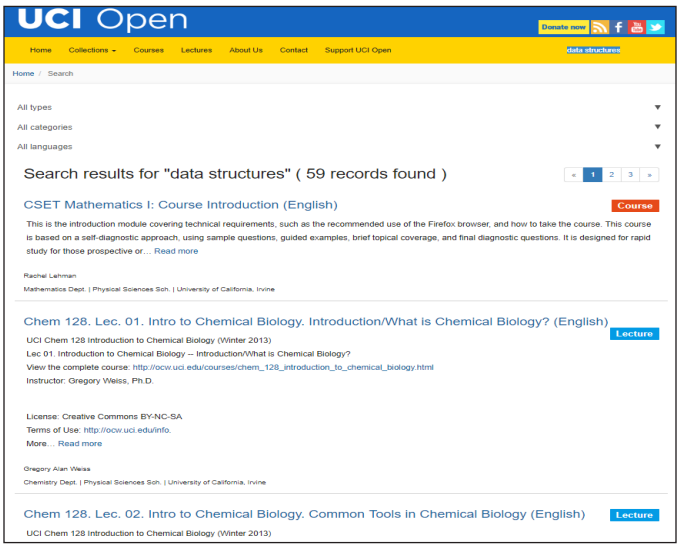

Figure 5. Search result of course 'Data Structures' in University of California (UC Irvine, OpenCourseWare) showing irrelevant course results

to irrelevant results. Some of the reputed journal papers (Balasubramanian et al., 2016; Muralikumar et al., 2016) have mentioned several open standard coursewares. They have clearly mentioned that the coursewares largely dependents on the tags, annotations and limited user-provided data for video retrieval. Lectures on these coursewares are presented with topic based segments, however, the structure and the organization of lectures is a result of manual processing. The search supported by most of these systems is mainly

occurrence based or tag-based, where the occurring search terms are highlighted in the transcripts. Some systems allow for navigating directly to the place where the search terms occur. Since annotations are given by users, there is a possibility for random keywords.

Annotating video with appropriate keywords from the video visuals will improve the accuracy which in turn improves precision. The search results can be improved by annotating the videos with the relevant keywords, which in turn will improve the search results.

Related Work

This section discusses about the various video retrieval systems using the video features and different methods for semantic video annotation.

Videos are segmented into shots. From each shot, shot level objects are selected by the user (El-Khoury et al., 2013). To identify objects in the shots, concept detectors are trained using classification algorithms such as k-nearest neighbour (k-NN), Support Vector Machine (SVM) or decision tree. Object features are extracted using Scale invariant feature transform (SIFT) descriptors. User chosen object are tracked by feature extraction methods. Annotation for each shot is done by concept detectors. Ten key frames are selected in algorithm using k-means clustering (Ravinder & Venugopal, 2016). Texture, edge and motion features are combined from all the ten key frames to represent feature vector. Feature vectors of query video and videos from video repository are compared for finding relationship between them using Euclidean distance measure. Resultant videos with less Euclidean distance are retrieved. However there is no defined methodology for filling the semantic gap between the content and the retrieved results.

Character regions are identified using Maximally Stable Extremal Regions (MSERs) based on the stroke width, letter height and width, character spaces (Wattanarachothai & Patanukhom, 2015). Text candidates are then classified according to their height and width. Tesseract OCR is applied to recognize the characters. Super-Fast Event Recognition system combines features of static visual descriptor, motion descriptor and audio descriptor (Jiang et al., 2015). All the feature descriptors are converted into bag of words representation. Then SVM kernel classifier is used to classify the events based on each feature. Middle level representation is created in Li's system to bridge the semantic gap between the low level features and high level features from the videos (Li et al., 2015). Middle level representations are built using Latent Dirichlet Allocation (LDA). To improve the system further and to reduce the computational cost, SIFT descriptors from LDA and fisher vectors are combined.

The work presented by Viana and Pinto (2017) proposed a video content annotation that used the concepts of crowdsourcing and gamification to collect metadata. To enhance the search and access, metadata were linked to specific time stamps. Semantic concepts

were identified through crowdsource tag-based dictionaries instead of standard dictionaries and thesaurus. Such semantic concepts lacks validation since system depends upon the external resources.

CourseMapper (Chatti et al., 2016) is an annotation platform that enables learners to collaborate and interact with video lecture using visual learning. Annotation editor allows user to add annotation to the viewed videos. These annotations are used for Visual learning. Visual learning methods are based on Annotation Maps and Heatmaps. AnnotationMap overlays stacks of annotation windows within the given timeline. To minimize the user distractions and to simplify the visual seeking for annotations, the cue points are marked in yellow color. This notifies the user that this portion of the video timeline has a larger number of annotations and most likely contains interesting information. Heatmap highlights the most viewed parts of video with warm colours such as orange and red, and less viewed parts are usually highlighted with cold purple and blue colours. Using this user can easily find the most interesting part of the video. Heatmap also records and displays the view count. This approach depends on the manual annotation of the authors.

In Kravvaris system, speech transcript for each video is collected from the repository (Kravvaris et al., 2015). These transcripts and the search query are converted into vectors. Cosine similarity between these two vectors is found. In addition to that, social weight to each video is added. That is likes and dislikes given by the registered YouTube users are taken into account. Based on the cosine similarity and the (like + dislike) values, videos are ranked.

Semantic annotation platform used by Xu & Mei (2015) enabled user to semantically annotate videos using vocabularies defined by traffic events ontologies. At first video annotation ontology is designed by following the traffic law which is machine understandable data. Description for video resources are given by annotator using those traffic field vocabularies. Semantic relationship between the annotations are used for management of annotated videos. However here annotation is carried out manually by the authors. Correlated Naive Bayes CNB classifier combines the methods of correlation and naïve bayes to retrieve relevant videos using the visual contents (Poornima & Saleena, 2018).

In this article, we define an approach that extracts visual contents of lecture video for automatic and semantic annotation. In addition, proposed approach does not rely on external resources such as social media, crowdsourcing, etc. for semantic video retrieval.

Educational Semantic Content Video Retrieval

Educational videos contain images, colourful illustrations, audio and many more. Most of the educational videos contain text as major feature (Balasubramanian et al., 2016; Li et al., 2015; Yang & Meinel, 2014). Analysing those text data and using it for annotation will

improve the search results. This automatic annotation will reduce the problem of author annotation and the time consumed for it. From the user's perspective, the user may not have clear idea on how to search a topic with proper query. So the user queries will be ambiguous and it will not match with the subjective contents of the video. This problem will be solved by giving meaning and relationships in the contents of video using WordNet ontology (http://wordnet-rdf.princeton.edu/ontology). WordNet ontology act as a thesaurus which groups English words based on synonyms. WordNet represents number of relationships between the members of WordNet. Workflow of the proposed work is shown in Figure 6. To annotate a video automatically requires complete analysis of the video. Analysis of the video needs image processing operations to extract the content and then data mining operations are needed for further processing of the content.

The process of semantic and automatic annotation based on video content involves the following two steps:

(i) Extraction of video contents

(ii) Generation and grouping of semantic keywords.

First step (Extraction of video contents) of the proposed work carries out image processing operations (Algorithm steps 1-3) such as selection of keyframe from video and text extraction from keyframe. Second step (Generation and grouping of semantic keywords) carries out datamining operations (Algorithm steps 4-7) such as detection of most frequent words, identification of semantic relationships and grouping of semantically related words & videos.

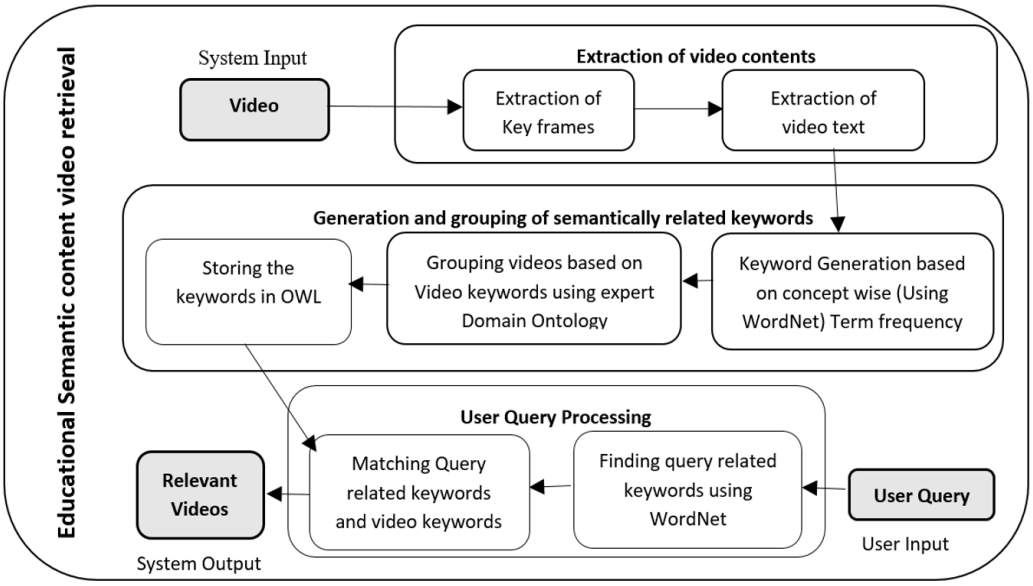

Figure 6. Educational Semantic video retrieval – Block Diagram

Extraction of Video Contents

Analysing a video content involves breaking the video into frames which is the basic element of a video. Unlike other videos, educational video contains text as major source of information. For the experimental setup, as of now only JPEG video formats are taken, but it can be extended for all other video formats. Extracting text contents from an educational video requires (i) Extraction of key frames and (ii) Extraction of text from videos.

Extraction of Key Frames

Video is made up of collection of frames. Frames are arranged in a temporal order to get a sequential flow. Key frames are the collection of frames which represents all the major elements of video. Key frames gives summary/abstraction of a video. Checking the transition change is the major task in finding the key frame. Several methods have been discussed in the literature for choosing the key frames, they are

(i) Reference frame: Reference frame is generated manually and then each video frame is compared with the reference frame to find the key frame (Ferman & Tekalp, 2003). Accuracy of the keyframe selection solely depends on the accuracy of reference frame selection.

(ii) Sequential comparison: Current frame and previous frame are compared based on pixel value. If there is much dissimilarity then the current frame will be taken as next key frame (Zhang et al., 2003). This method is very simple but there may be a repetition of same key frame since key frame analysis is carried out only based on local properties.

(iii) Clustering: Frames are clustered into groups and the frame which is nearer to the cluster center is taken as key frame. Accuracy of key frame selection depends on accuracy of clustering method used and its results (Yu et al., 2004). And also setting the number of key frames/cluster for grouping is difficult. Advantage over other methods is that it reflects global features of video.

Reference frame and sequential frame comparison methods uses pixel difference as key factor. Hence keyframe extraction techniques can be grouped into two approaches, pixel differencing approach or clustering approach.

Pixel comparison consists of finding the distance between the pixel values of consecutive frames. Pixel based methods compare a specific pixel in one frame with a corresponding pixel in a successive frame. Frames are converted into grayscale/black and white images before the comparison. Euclidean distance is defined in equation 1.

$$d\ (frame1, frame2) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\ \left(frame1(i,j) + frame2(i,j)\right)^2} \quad [1]$$

where d is the distance measure. Salt and pepper noise in the video may affect the keyframe selection accuracy while using pixel level comparison (Yang & Meinel, 2014) method.

Histogram based methods are alternative to pixel-based methods (Janwe & Bhoyar, 2016). Histogram gives the color distribution of the image. Successive similar frames will contain approximately the same color information and will have a similar histogram. Histogram difference between two frames is calculated using equation 2.

$$HFD = \{Histogram\ of\ 1^{st}\ frame - Histogram\ of\ 2^{nd}\ frame\} * Number\ of\ Gray\ levels \quad [2]$$

If HFD of particular frame is more than the threshold value, then that frame will be taken as key frame. Threshold can be calculated using equation 3. Drawback is that images with similar histograms may have different visual appearance.

$$Threshold = Mean\ Deviation + (a * Standard\ Deviation) \quad [3]$$

where a is a constant.

Lecture video contents such as text lines, images, and tables, can be taken as connected components of an image (Yang & Meinel, 2014). So connected components method is used in this proposed work to identify the key frames (Figures 7 - 8). Two pixels are said to be connected, if there is a path from one pixel to the other i.e., both pixels share same intensity value. Connected components can be identified by analysing the image from left to right and top to bottom. Connected component can be easily determined by giving a pixel particular label value. For example: Labelling of pixel p can be done through following information:

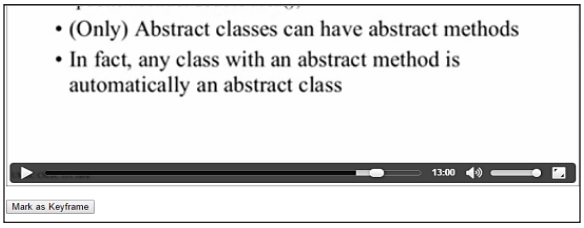

Figure 7. Selection of Key Frame from any online video

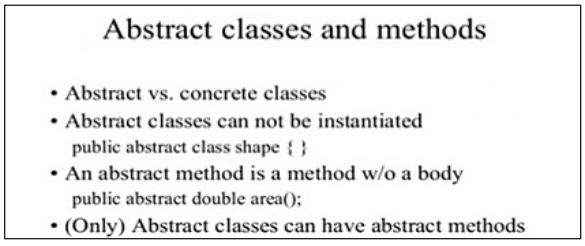

Figure 8. Selected Key frame taken as image for further text extraction

Step 1: If all four neighbors of a pixel p are 0, assign a new label to pixel p, else
Step2: If only one neighbor of a pixel p has V={1}, assign its label to p, else
Step3: If more than one of the neighbors have V = {1}, assign one of the labels to p and make a note of the equivalences.

Extraction of Text from Videos

Lecture contents closely depends on the text in lecture slide. These texts help in retrieval task for automation. Text extraction from video includes text detection and text recognition. Text detection refers to the presence of text in the frames. Text recognition is the process of converting text present in images into machine readable data.

Naturally, text has some properties such size, similar pattern, more interest point, high contrast than the background, connectedness and many. Text detection algorithms focus on these properties to detect text. Presence of text in frames is identified using connected components methods. Text lines are the major content of the video frame which are used for finding the connectedness. Tesseract is an open source OCR used to extract the video text and are converted into edited format for using it in next step. Key frames are converted into black and white images as a pre-processing for Tesseract. The procedure for conversion is as follows: step 1. Convert the colour images into binary/black and white image, step 2. Identify blobs and character lines, step 3. Match the character lines with pre-trained character set to find the character. For implementation, Java Tesseract (Figure 9) is used to recognize the text characters from the image frame. Then the extracted text are stored in an editable format (i.e., doc or txt) for further processing. Accuracy of Tesseract OCR are enhanced by using dictionaries.

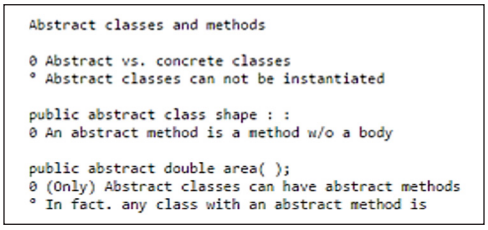

Figure 9. Extracted text in editable format from the selected video key frame

To extract information from a bulk data, which will lead to some meaningful pattern or knowledge, pre-processing plays very important role. Pre-processing includes case folding, stopwords removal, tokenization, Parts-Of-Speech (PoS) tagging, stemming and lemmatization. Varieties of capitalization may affect the processing, most common approach is to reduce all the words into lower cases (case folding). Most of the words in the sentence are the connecting parts rather than showing the subjects, objects or intent. Those words can be removed by comparing it with list of stopwords. Some examples of stopwords are

'of', 'the', 'an'. Tokenization describes splitting paragraphs into sentences, or sentences into individual words. Sentences can be split into individual words and punctuation through splitting across white spaces.

A Part-Of-Speech Tagger (POS Tagger) reads text and assigns parts of speech to each word, such as noun, verb, and adjective. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. However, the two words differ in their flavour. Stemming refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. For example: The stemmed form of analysis is: analysis and the lemmatized form of leaves is: analysis.

Generation and Grouping of Semantically Related Keywords and Videos

Relationship between the text extracted from the videos is found using WordNet. Finally frequency of terms is calculated for arranging the videos.

Relationship between Extracted Texts Using Wordnet

Each word from the extracted text is given meaning and relationships using WordNet. WordNet is a large lexical database of English words that are connected together by their semantic relationship. WordNet acts as both dictionary and thesaurus. Using WordNet ontology (Figure 10), meaning and relationships between the words in the video contents are found. In general, WordNet ontology has some flaws. If the dictionary is domain based, then it may be exactly suiting to our needs.

Algorithm:

Input: User Query and Video Database

Output: Set of relevant lecture videos

Step 1: Let us consider a video database D, which is a collection of lecture videos.

$$D = \{ V_1, V_2, \ldots, V_D, \ldots, V_n \} \quad [4]$$

where $V_1, V_2, \ldots V_n$ are the individual videos in the collection and n is the number of videos.

Step 2: Every individual video contains n number of frames in it. Let us consider V_D that has K number of keyframes.

$$V_D = \{V_D^1, V_D^2, \ldots, V_D^F, \ldots, V_D^K\} \quad [5]$$

where $V_{D,}^{1}$ V_{D}^{2}, .. are number of frames in D^{th} video and K represents total number of keyframes.

Step 3: Let us consider a video frame V_{D}^{F} that contains the textual features. The number of keywords generated from the frame V_{D}^{F} is,

$$W = \{ w_1, w_2, \ldots, w_m \} \quad [6]$$

where V_{D}^{F} is the F^{th} frame in the D^{th} video, w_1, w_2, . . are the keywords generated from the key frame $V_D{}^F$ and m is the total number of keywords extracted from the key frame V_{D}^{F}. Extraction of the keywords from the key frames follows the OCR technology, which uses the tesseract classifier.

Step 4: Each keyword carries some set of semantic words. Semantic words are the words which has similar synonym.

$$WS_1 = \{ ws_1^1, ws_1^2, \ldots ws_1^t \} \quad [7]$$

where $ws_1{}^1$, $ws_1{}^2$, . . , $ws_1{}^n$ are the semantic words extracted from the word w_1 and $ws_1{}^t$ represent the t^{th} semantic word generated from w_1 word. These semantic words are extracted from the keywords using WordNet Ontology.

Step 5: Term Frequency-Inverse Document Frequency (TF-IDF) is used to measure the importance of a keyword. tf(t,d) is the frequency of term t in document d.

$$idf(t) = log_2 \left(\frac{D}{df(t)} \right) \quad [8]$$

where df(t) is the document frequency and D is total number of documents in the domain corpus. TF−IDF for term t in document d is

$$tf - idf_{(t,st),d} = tf_{(t,st),d} * idf_t \quad [9]$$

where t is term and st is semantic term. TF-IDF helps in differentiating domain-specific terms and highly generic terms.

Step 6: Based on the words, semantic words and frequency, words are clustered into group. Grouping depends on the minimum distance between the clusters.

Step 7: When query arrives, the classifier matches query with clusters. Cluster with the maximum probability gets retrieved.

```
> findAssocs(dtm, "data", corlimit=0.6)
          data
analysi   1.00
base      1.00
becom     1.00
call      1.00
comput    1.00
consist   1.00
follow    1.00
ieee      1.00
issu      1.00
messag    1.00
micro     1.00
part      1.00
predict   1.00
process   1.00
tool      1.00
use       1.00
veri      1.00
engine    0.98
feature   0.98
repress   0.98
extract   0.98
word      0.98
classify  0.97
financi   0.97
```

Figure 10. Identifying associated words

Term Frequency Calculation

Frequency of the terms taken from the ontology for every video is computed. Number of terms is calculated along with the semantic terms. If words appear frequently in a document, then they will be considered as important words. These words will be given a high score. But if a word appears in many documents, that word will get low score since it is not a unique identifier. Term frequency of documents are calculated using equation 9. Based on the frequency of terms and semantic terms (Figures 11-12), videos are ordered according to user query. Key frame selection and text extraction can be carried out at the initial stage itself even before the user starts querying. Association and semantic relationship meanings are found among the contents based on the user query information, so it can be carried out after the user gives the query.

	Terms								
Docs	recent	refer	relate	report	repres	resolv	result	science	sens
Text1.txt	4	9	2	2	1	3	2	2	2
Text2.txt	1	3	2	2	2	1	1	5	1
Text3.txt	2	1	3	1	1	4	1	4	1
	Terms								
Docs	site	social	special	support	system	technolog	time	train	univers
Text1.txt	1	1	4	3	13	1	8	8	4
Text2.txt	8	8	1	2	10	1	12	12	6
Text3.txt	6	5	1	6	8	3	1	1	6
	Terms								
Docs	updat	use	whole	word	work				
Text1.txt	1	2	3	17	18				
Text2.txt	4	13	1	4	1				
Text3.txt	1	36	1	7	1				

Figure 11. Document-wise term frequency

> freq

abstract	accord	all	analyz	base	better	build	case	center	confus
5	6	12	4	12	5	5	4	8	6
construct	content	data	develop	differ	each	effect	emot	example	exist
14	3	35	19	9	9	10	12	4	3
Find	first	focus	found	general	group	help	implement	import	improv
8	5	5	4	4	12	12	8	8	7
Inform	issu	journal	keyword	know	knowledg	main	make	mean	model
36	15	12	3	7	11	8	14	4	36
Network	object	one	onlin	other	out	own	paper	perform	practic
17	4	8	19	15	10	4	14	5	3
Present	prevent	problem	provid	public	purpos	recent	refer	relat	report
3	14	16	12	7	5	4	5	7	4
Repress	resolv	result	scienc	sens	site	social	special	support	system
14	4	8	10	3	15	15	6	11	31
Technolog	time	tool	train	univers	updat	use	whole	word	work
5	21	11	10	16	6	51	5	24	20

Figure 12. Overall term frequency

Grouping of Semantically Related Videos

Video database consists of video from four different domains (categories) as sample which includes agriculture, India, quantum optics and datamining. Domain ontology (Figure 13-14) is a formal model that serves as system's structure. Domain expert explores specific knowledge, analyse the most relevant entities and organises them into concepts and relationships. The skeleton of ontology consists of a hierarchy of generalized and specialized concepts. Domain ontology is created for each category in video database. Keywords of each video will be compared with the domain ontology. Videos are grouped on the basis of domain ontology and video keywords. When query matches with the cluster, all the videos in the cluster will be given as results.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:video="http://vit.ac.in#">
 <rdf:Description rdf:about="http://vit.ac.in#Video3">
  <video:course>DataMining Clustering Algorithms</video:course>
 </rdf:Description>
 <rdf:Description rdf:about="http://vit.ac.in#Video1">
  <video:course>DataMining Basics</video:course>
 </rdf:Description>
 <rdf:Description rdf:about="http://vit.ac.in#Video2">
  <video:course>DataMining Classification Algorithms</video:course>
 </rdf:Description>
</rdf:RDF>
```

Figure 13. Sample RDF used for retrieval using Jena

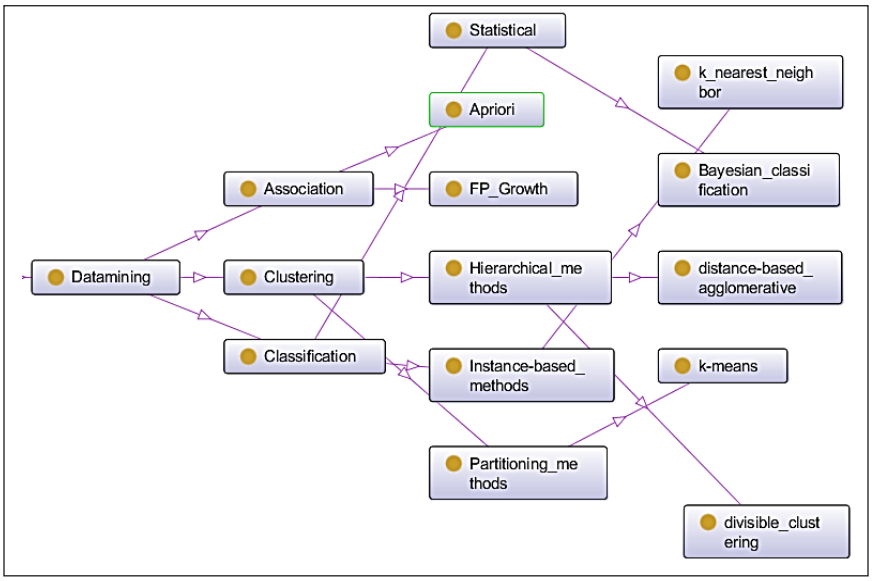

Figure 14. Graphical representation of relationship between entities

RESULTS AND DISCUSSION

Key frame extraction, text extraction, term frequency calculation and finding semantic meaning and relationships are the major components taken for result analysis. Key frames are selected randomly and from the key frames text are extracted through Tesseract, an open source API. WordNet, an open source electronic lexical database used for finding the semantic meaning and relationships of terms in document and in query. Term frequencies are calculated using Term Frequency and Inverse Document Frequency.

This research work includes various video samples from the fields of agriculture, India, quantum optics and datamining for the lecture video retrieval. Queries includes all the four categories of videos in the database. In total 50 video samples are taken for experimentation. Duration of each video is around one hour. Number of frames on each video depends on the quality of the video. Figure 15 explains the retrieved video contents for query. The performance of the proposed method is compared with the existing Correlation incorporated Naive Bayes (CNB) lecture video retrieval (Poornima & Saleena, 2018), since this is the most recent work done among all other related work. Ground truth was generated by experts who have a knowledge of the topic with an understanding of topics to follow. For a chosen subset of documents in these sets, ground truth was determined by generating potential results for sample documents that fall under the different categories of relatedness. To compare our results against the ground truth, we use precision (no. of relevant results/no. of results obtained), recall (no. of relevant results/no. of expected results) and f-measure metrics. Average of precision, recall and f-measures for queries of each category is calculated and included for those calculations.

Figure 15. Query and relevant videos retrieved

Precision is the ratio of the number of relevant videos retrieved to the total number of irrelevant and relevant videos retrieved. Figure (Figure 16) shows the precision rate of semantic information retrieval for query with CNB. Precision rate of semantic information retrieval achieves 0.9205 whereas precision rate for CNB is 0.72, which is low compared to semantic information retrieval. Thus, the values conclude that the precision rate is better for semantic information retrieval when compared to CNB.

Recall is the ratio of the number of relevant videos retrieved to the total number of relevant videos in the database. Figure (Figure 17) shows the recall rate of semantic information retrieval with CNB. Recall rate of semantic information retrieval achieves 0.9335 whereas recall rate for CNB method is 0.78, which is low compared to semantic information retrieval. Thus, the values conclude that the recall rate is better for semantic information retrieval when compared to CNB method. Comparison between precision and recall is shown in the figure (See Figure 18).

F-measure is the harmonic mean of precision and recall, that is, it is a combination of precision and recall. F-measure is intended to combine these two into a single measure of search effectiveness. Traditional equation for F-measure is given in equation 12:

$$F = 2.\frac{Precision.Recall}{Precision+Recall} \qquad [12]$$

Figure (Figure 19) shows the f-measure of semantic information retrieval for query with CNB. F-measure of semantic information retrieval achieves 0.925 whereas f-measure for CNB method is 0.75, which is low compared to semantic information retrieval. Thus, the values conclude that the f-measure is better for semantic information retrieval when compared to CNB.

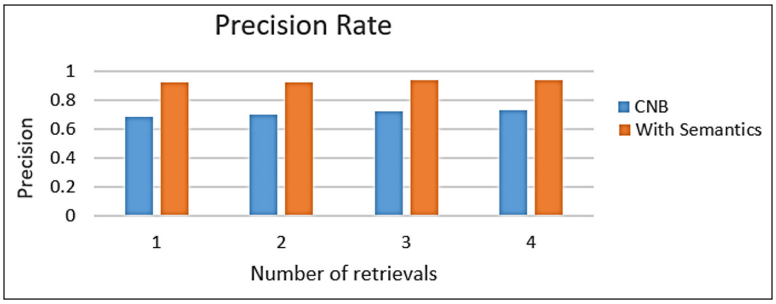

Figure 16. Precision rate

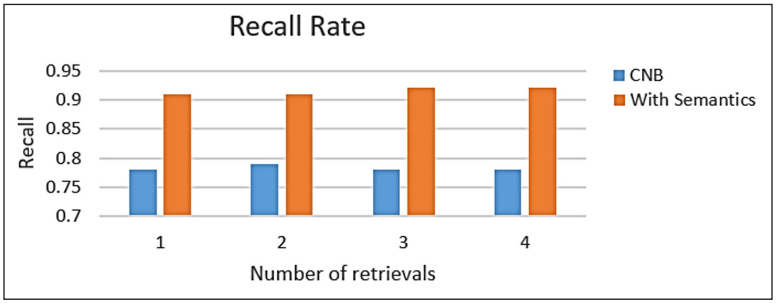

Figure 17. Recall rate

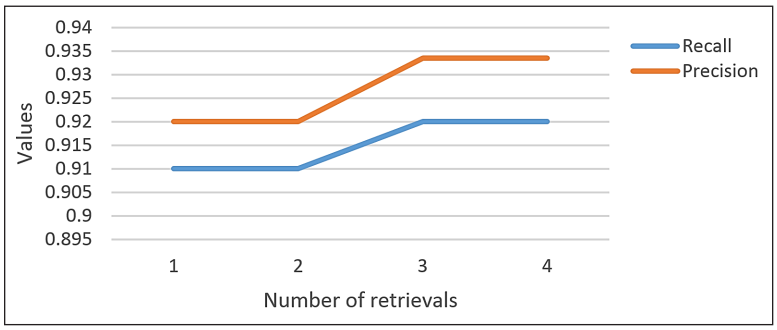

Figure 18. Precision and recall curve

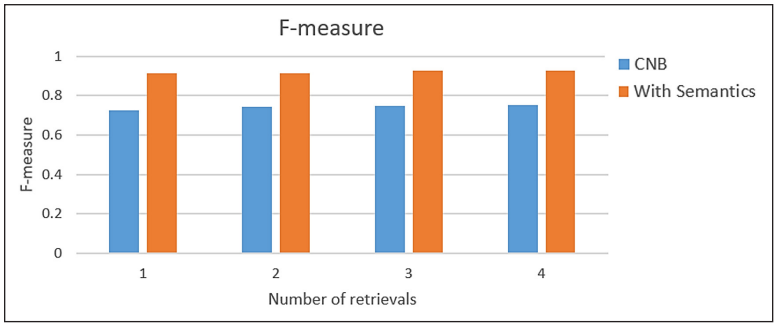

Figure 19. F-measure

CONCLUSION

Our proposed work has addressed the issues of extracting content descriptive annotations for the purpose of supporting content based video lecture retrieval. This paper has discussed a two phase methodology for capturing semantically related keywords from the video contents (visuals). First phase covers the extraction of text features from the video using connected components method and optical character recognition technique for keyframe detection and text extraction respectively. Second phase captures semantically related words for the frequently occurring words from the first phase. Videos are clustered by comparing semantically related words with domain ontology. The effectiveness of the proposed approach has been proven based on the experimentations done on that actual video set. Future work includes further experiments to extract keywords from all kinds of videos other than educational videos.

REFERENCES

Balasubramanian, V., Doraisamy, S. G., & Kanakarajan, N. K. (2016). A multimodal approach for extracting content descriptive metadata from lecture videos. *Journal of Intelligent Information Systems, 46*(1), 121-145.

Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Yousef A. M. F., & Schroeder, U. (2016). Video annotation and analytics in CourseMapper. *Smart Learning Environments, Springer, 3*(10), 1-21.

Che, X., & Lin, L. (2015). A Survey of Current YouTube Video Characteristics. *IEEE Transaction on Multimedia, 22*(2), 56-63.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370-383.

El-Khoury, V., Jergler, M., Bayou, G. A., Coquil, D., & Kosch, H. (2013). Fine-granularity semantic video annotation. *International Journal of Pervasive Computing and Communications, 9*(3), 243-269.

Ferman, A. M., & Tekalp, A. M. (2003). Two-stage hierarchical video summary extraction to match low- level user browsing preferences. *IEEE Transactions on Multimedia, 5*(2), 244-256.

Janwe, N. J., & Bhoyar, K. K., (2016). Video Key-Frame Extraction using Unsupervised Clustering and Mutual Comparison. *International Journal of Image Processing (IJIP), 10*(2), 73-84.

Jiang, Y. G., Dai, Q., Mei, T., Rui, Y., & Chang, S. F., (2015). Super Fast Event Recognition in Internet Videos. *IEEE Transactions on Multimedia, 17*(8), 1174-1186.

Kravvaris, D., Kermanindis, K. L., & Chorianopoulos, K. (2015, May). Ranking Educational Videos: The Impact of Social Presence. In *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)* (pp. 342-350). Athens, Greece.

Li, H., Liu, L., Sun, F., Bao, Y., & Liu, C. (2015). Multi-level feature representations for video semantic concept detection. *Neurocomputing, 172*, 64-70.

Li, K., Wang, J., Wang, H., & Dai, Q. (2015). Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(5), 1233-1246.

Muralikumar, J., Seelan, S. A., Vijayakumar, N., & Balasubramanian, V. (2016). A statistical approach for modeling inter-document semantic relationships in digital libraries. *Journal of Intelligent Information Systems, 48*(3), 477–498.

NPTEL Frequently Asked Questions. (2018, March 05). Retrieved July 8, 2017, from http://nptel.ac.in/faq.php

Poornima, N., & Saleena, B. (2018). Multi-modal features and correlation incorporated Naive Bayes classifier for a semantic-enriched lecture video retrieval system. *The Imaging Science Journal, 66*(5), 263-277.

Ravinder, M., & Venugopal, T. (2016). Content-Based Video Indexing and Retrieval using Key frames Texture, Edge and Motion Features. *International Journal of Current Engineering and Technology, 6*(2), 672-676.

Viana, P., & Pinto, J. P. (2017). A collaborative approach for semantic time-based video annotation using gamification. *Human-Centric Computing and Information Sciences, Springer, 7*(13), 1-21.

Wattanarachothai, W., & Patanukhom, K. (2015). Key Frame Extraction for Text Based Video Retrieval Using Maximally Stable Extremal Regions. In *2015 IEEE 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)* (pp. 29-37). Tokyo, Japan.

Xu, Z., Mei, L., Liu, Y., Zhang, H., & Hu, C. (2015). Crowd Sensing Based Semantic Annotation of Surveillance Videos. *International Journal of Distributed Sensor Networks, 11*(6), 1-9.

Yang, H., & Meinel, C. (2014). Content Based Lecture Video Retrieval Using Speech and Video Text Information. *IEEE Transactions on Learning Technologies, 7*(2), 142-154.

Yu, X., Wang, L., Tian, Q., & Xue, P. (2004). Multilevel video representation with application to keyframe extraction. In *2004 IEEE 10th International Multimedia Modelling Conference* (pp. 117-123). Brisbane, Queensland, Australia.

Zhang, X., Liu, T., Lo, K., & Feng, J. (2003). Dynamic selection and effective compression of key frames for video abstraction. *Pattern Recognition Letters, 24*(9–10), 1523-1532.