# Fuzzy Clustering of Lecture Videos Based on Topic Modeling

Subhasree Basu*, Yi Yu†, Roger Zimmermann*

*National University of Singapore, Singapore 117417, {sbasu, rogerz}@comp.nus.edu.sg
†National Institute of Informatics, Tokyo 101-8430, Japan, yiyu@nii.ac.jp

***Abstract*—Lecture videos constitute an important part of the e-learning paradigm. These online video-lectures contain multimedia materials aimed at explaining complex concepts in a more effective way. The videos are mostly grouped by their subjects. However, often there are overlaps between the subjects, *e.g.*Mathematics and Statistics. Hence, educational content-wise, some of the lecture videos can belong to more than one subject. When they are labeled by only one subject, students searching for the content of the lecture might miss some of these videos. To solve this problem, we aim to provide a clustering of these lecture videos based on their educational content rather than their titles so that such lectures will not be missed out based on the subject labels. Our novel algorithm uses topic modeling on video transcripts generated by automatic captions to extract the contents of these videos. We choose representative text documents for each of the clusters from the Wikipedia. Then we calculate a similarity between the topics extracted from the videos and those of the representative documents of the clusters. Finally we apply fuzzy clustering based on these similarity values and provide a lecture-content based clustering for these lecture videos. The initial results are plausible and confirm the effectiveness of the proposed scheme.**

## I. Introduction

One of the faculties that have benefited from the various advancements in computer science and communication technologies is the Technology enhanced learning (TEL) and among the various aspects of TEL, advancements in distance education is note-worthy [1]. TEL is mainly concerned with the design and development of various socio-technical innovations for various kinds of learning and education, involving technologies for individual learners as well as for groups and organizations. It is therefore an application domain that generally covers technologies that support all forms of teaching and learning activities [2].

E-learning, a new approach to distance learning, augments learning experiences by integrating multimedia and network technologies and providing instant availability of various types of relevant study materials to the students. As an integrated part of e-learning, the lecture videos captured in classrooms contain a substantial portion of the instructional content [3]. These videos are hosted on numerous e-learning sites like Coursera, Khan Academy, VideoLectures.net, *etc.*, and uploaded to video-sharing sites like YouTube, Slideshare, *etc.* Leading universities like MIT and Stanford have made their lectures available online for distance learning. Thus the popularity and the importance of online lecture videos are increasing at a substantial rate.

The lecture videos available online have title, metadata (subject, description, keywords, *etc.*), multimedia visual content and also textual transcripts generated by closed caption (CC) feature. The videos are generally grouped together by their subjects or titles rather than their educational material. However, often there are overlaps among the course contents between subjects *e.g.*Electrical Engineering and Electronics and Communication. Lectures on ‘Diode’ can be found separately grouped under the subjects like Electrical Engineering, Electronics and Communications as well as Physics. Hence, educational content-wise a lecture video can be part of more than one group when grouped by subjects alone.

In this paper, we attempt to cluster the lecture videos based on the course-content covered in these videos, not just the subject alone. Our algorithm will yield the clusters of the videos along with the degree of membership of the videos for each of the clusters. We leverage Latent Dirichlet Allocation (LDA) to find the topics in a particular lecture video. The input files to our system are the transcripts generated from the the videos and we use LDA to find latent topics from them. These topics provide better representation of the educational content of the lecture video compared to video’s title or metadata. We need a representative for each cluster and we use the Wikipedia pages (wiki-pages), having the same subject as their Wikipedia Category, as the cluster-label for such representation *e.g.*for the representative for the cluster for Mathematics, we use the wiki-pages having the Wikipedia Category as Mathematics. The topic distribution of the collection of these pages represent the topic distribution for the particular cluster. Based on these distributions, we find the similarity score between the representative documents for each cluster and the videos in our repository. Next we cluster the videos using a fuzzy clustering technique with the similarity scores as the input. In summary, our specific contribution, in this paper are as follows:

- Use Fuzzy Clustering technique to cluster the videos based on their educational content and finding latent topics from them.
- Effectively use topic modeling on video transcripts for lecture video clustering.

The rest of the paper is organized as follows. In Section II

978-1-4673-8695-1/16/$31.00 ©2016 IEEE

we provide a brief summary of the related works. In Section III we give a description of the model of our system and Section IV summarizes our experiments and the results. We conclude with the direction of our future work in Section V.

## II. Related Work

Huang et al., have tried to cluster text documents based on wiki-pages [4]. However, they suggest mapping documents to wiki-pages by only leveraging the anchor texts of the pages. Vidal et al., have tried to mine Wikipedia semantically to design a system for keyword-extraction based on Wikipedia categories and metadata [5]. Again, in none of these methods the latent topics of the wiki-pages are explored. Lee et al., have developed a clustering algorithm based on LDA in iVisClustering [6]. iVisClustering presents a visualization tool for the user to interactively cluster documents. Each of these clustering methods has designed systems for text documents and not for multimedia documents. In comparison, we propose to use LDA in fuzzy clustering to successfully cluster multimedia documents like lecture videos. One of the attempts to clustering lecture videos was to cluster them based on keywords useful to users [7] where the keywords are matched with the query words used in video-search by the users.

It is also necessary to detect and recognize the content structures of instructional videos. One kind of the structures in these videos is the change of different presentation formats, also called narrative elements. Dorai et al. [8] provide a decision tree model and use color moments to classify in lecture videos narrative elements such as narrator frame sequence, web text frames, and slide text frames. Specially for videos of electronic slide (such as PowerPoint) presentations, there are previous research work [9]–[11] that detect the changes of slides and relate slide content to video segments, enabling further content query, event detection, and audio synchronization. However, only detecting narrative elements and slide changes is not enough for effective content retrieval. For example, in one lecture, an instructor explains one topic using a combination of electronic slide, blackboard, and narration, while in another lecture, he may explain several topics only using the board. In both cases, a semantic structure based on instructional topics is more meaningful for indexing and retrieval than the narrative elements are. One method is provided to recognize semantic structure in blackboard videos based on spatial and temporal grouping of teaching content [12]. However, in none of the previously attempted problems, the authors have tried to automatically cluster videos based on the academic content covered in the lectures. Hence our attempt to clustering the videos by discovering the latent topics in them and also to assigning each video to more than one cluster by assigning fuzzy membership to them is novel in nature.

## III. System Model for Video Clustering

In our system as described pictorially in Fig. 1, we will use the transcripts of the lecture videos to represent the content of lectures delivered in the videos. To find the representative content for each of the clusters, we use the wiki-pages having the same subject as their Wikipedia Category as the cluster-name for such representation.

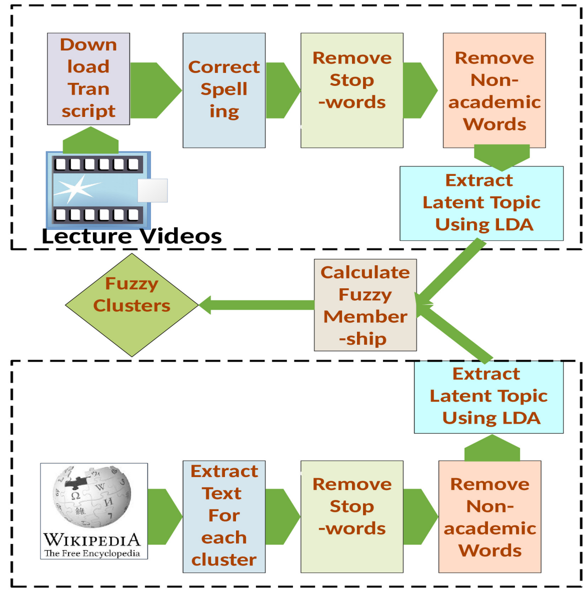



Fig. 1: System model for video clustering.

### A. Dataset

We collected the videos from the YouTube channel for National Programme on Technology Enhanced Learning (NPTEL) (http://www.nptel.ac.in/) and extracted the transcripts uploaded for some of the videos in this channel. These transcripts are generated from the YouTube platform using the closed caption feature of YouTube. On top of that we also used lecture videos uploaded on the VideoLectures.net and used the subtitle files available along with the videos. All the videos have English as the medium of instruction. Hence the transcripts generated consist of words in English. The text available after removing the timing information from the subtitle files, provided the text for the transcripts of these videos. We extract the subjects for each of the videos from the metadata of the videos. Our video repository now consists of 3,000 videos covering the following subjects: Humanities, Metallurgical and Material Science, Electrical Engineering, Chemistry, Mathematics, Electronics and Management. We have manually annotated the videos with the probable subjects they can be grouped under the above seven subjects. These subjects form the labels for the clusters of our system as well and we compare the accuracy of our method by comparing with both state-of-the-art methods and the ground truth.

### B. Pre-Processing Transcripts

These transcripts extracted from the videos are not 100% accurate and necessitate correcting the transcript by checking with a dictionary. We matched the words against a dictionary provided by **WordNet** [13]. We queried WordNet with the incorrect word for a substitute word. We used the first element of the list of words returned by WordNet as the substitute for the incorrect words. Next we remove the stop-words from the remaining words in the transcripts. This leaves us with the main bag of words of the documents. The videos are educational in nature and hence the representative words in these transcripts should come from an academic vocabulary. Thus, we remove the non-academic words from the bag of words by matching them against a set of 103 million academic words obtained from Academic Vocabulary List (AVL) [14]. The remaining bag of words now represent the videos and we applied LDA on them to find the topics.

### C. Obtaining Wikipedia Texts

Wikipedia has been gaining popularity as a referential educational material for students. Thus, we leverage the content of wiki-pages to form the main educational content that should be covered in each cluster.The videos in the repository are labeled by subjects. As mentioned in Section I, using these subject labels we navigate to the wiki-pages which have the same Wikipedia Category names as the subject-names, *e.g.*, we navigate to all the wiki-pages having Mathematics as the Category name to prepare the representative document for the subject Mathematics and so on. Next, we extract contents from wiki-pages for each of the clusters, collected as mentioned above, and process them as described in Section III-B. This provides us with the bag of words to be used as representative for each cluster.

### D. Topic Modeling

The transcripts and the bag-of-words are generated for each video in the video repository stored on the server. We use LDA to find the latent academic topics present in the transcripts of each of the videos [15]. In case of LDA, each document is represented as a distribution of topics and each topic in turn is represented as a multinomial distribution over words. Let the distribution of latent topics present in video, $Vid_k$, be denoted as the $\{P_{Vid_k}(z_i)\}$, which are returned by LDA, where $z_i$ are the latent topics present in $Vid_k$ and let the distribution of latent topics present in the representative document for cluster $C_m$ be $\{P_{Wiki_{C_m}}(z_j)\}$, where $z_j$ are the latent topics present in $Wiki_{C_m}$, the collection of wiki-pages for cluster $C_m$.

### E. Similarity Calculation

In the probability theory and information theory, the Kullback-Leibler divergence (KL Div) is a non-symmetric measure of the difference between two probability distributions $P$ and $Q$ [16]. Specifically, the KL Div of $Q$ from $P$, denoted $D_{KL}(P||Q)$, is a measure of the information loss when $Q$ is used to approximate $P$. KL Div being a non-symmetric metric, we calculate the KL Div of both $P$ from $Q$ and $Q$ from $P$ and then take the average of the two. The lower the value of the KL Div between two distributions, the closer they are in terms of similarity. While clustering the videos, we use KL Div between the topic distribution of the videos and the representative topic distribution for each of the clusters generated by the wiki-pages belonging to each clusters. The similarity measure of a video with a cluster is thus defined as:

$$\begin{aligned} Sim(Wiki_{C_m}, Vid_k) = \\ [\frac{1}{2}(D_{KL}(\{P_{Vid_k}(z_i)\} \| \{P_{Wiki_{C_m}}(z_j)\}) \\ + D_{KL}(\{P_{Wiki_{C_m}}(z_j)\} \| \{P_{Vid_k}(z_i)\}))]^{-1} \end{aligned} \tag{1}$$

### F. Fuzzy Clustering

Fuzzy clustering is a soft clustering technique in which every data point has a degree of belonging to each of the clusters instead of belonging only to a particular cluster as a whole. We use fuzzy clustering to cluster our videos because often the content of the lecture videos contains materials belonging to more than one subject, *e.g.*, there are some videos explaining Probability for Electrical Engineering, Electronics, Computer Science in addition to Mathematics. Hence, these videos when clustered according to subjects alone, will be grouped under one of the subjects and will not be retrieved if other subjects are used as the query.

We demonstrated above, the need for assigning the same video to more than one clusters. To this end, we will need a method to model uncertainty and thus choose the Fuzzy C-Means clustering technique [17]. Given a set of objects, $X = x_1, \ldots, x_n$, a fuzzy set $S$ is a subset of $X$ that allows each object in $X$ to have a membership degree between 0 and 1. Formally, a fuzzy set, $S$, can be modeled as a function, $F_S : X \to [0, 1]$.

This fuzzy set idea is applied on clusters and given a set of objects, a cluster, which is a fuzzy set of objects, is obtained. Such a cluster is called a fuzzy cluster and consequently, a fuzzy clustering contains multiple fuzzy clusters. Formally, given a set of objects, $o_1, \ldots, o_n$, a fuzzy clustering of $k$ fuzzy clusters, $C_1, \ldots, C_k$, can be represented using a partition matrix, $M = [\mu_{ij}]$ $(1 \leq i \leq n, 1 \leq j \leq k)$, where $\mu_{ij}$ is the membership degree of $o_i$ in fuzzy cluster $C_j$. The partition matrix should satisfy the following three requirements:

- For each object, $o_i$, and each cluster, $C_j$, $0 \leq \mu_{ij} \leq 1$. This requirement enforces that a fuzzy cluster is a fuzzy set.
- For each object, $o_i$, $\sum_{j=1}^{k} \mu_{ij} = 1$. This requirement ensures that every object participates in the clustering equivalently.
- For each cluster, $C_j$, $0 < \sum_{i=1}^{n} \mu_{ij} < n$. This requirement ensures that for every cluster, there is at least one object for which the membership value is nonzero.

Following the definition above, we assign each video, the membership value calculated by the following formula:

$$\mu(Vid_k, C_j) = \frac{Sim(Wiki_{C_j}, Vid_k)}{\sum_j Sim(Wiki_{C_j}, Vid_k)} \tag{2}$$

We generate the word frequency vectors for each of the videos and their membership values for each cluster is calculated as above. These word frequencies and the membership values are input to the fuzzy C-means clustering algorithm and the clusters are generated according to Algorithm 1.

### G. Clustering Algorithm

The algorithm used to cluster the videos is given in Algorithm 1.

**Algorithm 1** Fuzzy clustering based on topic modeling.

**procedure** PREPAREVIDEOTOPICMODELS
  $V$ = Set of videos in the Video Repository
  **for** $Vid_k \in V$ **do**
    Extract the content of $Vid_k$ as $Vid_k^{'}$
    Remove stop-words from $Vid_k^{'}$
    Check the words in $Vid_k^{'}$ against a dictionary to correct spelling mistakes, remove non-academic words
    Use LDA to find the latent topics in the $Vid_k^{'}$ and return $\{P_{Vid_k}(z_i)\}$
  **end for**
**end procedure**

**procedure** CLUSTERVIDEO($V$, $k$)
  PREPAREVIDEOTOPICMODELS()
  **for** each cluster $C_j$ **do**
    Prepare Cluster Representative $Wiki_{C_j}$
  **end for**
  **for** each $Wiki_{C_j}$ **do**
    **for** each $Vid_k^{'} \in V$ **do**
      Calculate $Sim(Wiki_{C_j}, Vid_k^{'})$
    **end for**
  **end for**
  **for** $Vid_i \in V$ and each cluster $C_j$ **do**
    $\mu(Vid_i, C_j)$ = FINDMEMBERSHIP($Vid_i^{'}$, $C_j$)
  **end for**
  Prepare Word Frequency feature for each each $Vid_j \in V$
  Run Fuzzy C-Means Clustering with Word Frequency and $\mu(Vid_i, C_j)$
  Return the $k$ clusters
**end procedure**

**procedure** FINDMEMBERSHIP($Vid_i^{'}$, $C_j$)
  **for** each cluster $C_j$ $j$ **do**
    $\mu(Vid_i^{'}, cl_j) = \frac{Sim(Wiki_{C_j}, Vid_i^{'})}{\sum_j Sim(Wiki_{C_j}, Vid_i^{'})}$
  **end for**
  Return $\{\mu(Vid_i^{'}, C_j)\}$
**end procedure**

As described in the procedure PREPAREVIDEOTOPICMODELS, our system extracts the transcripts of the videos, removes the stop-words from these transcripts and checks the remaining words against a dictionary to eliminate spelling mistakes, remove non-academic words, *etc.* The details are given in Section III-B. Then, LDA is applied on these transcripts to extract the topic distribution of these videos. Let the content of the video $Vid_k$ after checking against a dictionary be $Vid_k^{'}$. LDA will return a set of topics $\{P_{Vid_k}(z_i)\}$ along with the probability distribution of the topics. We next calculate the topic distribution of the representative documents for each of the clusters and denote them as $\{P_{Wiki_{C_m}}(z_j)\}$. We calculate the membership values of each video by the formula given in Equation 2 and described in FINDMEMBERSHIP. Next we extract word frequency as the feature vectors for each of the videos. The word frequency and membership values of the videos, $\mu(Vid_k, C_j)$, provide the input to the Fuzzy C-Means clustering algorithm.

## IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

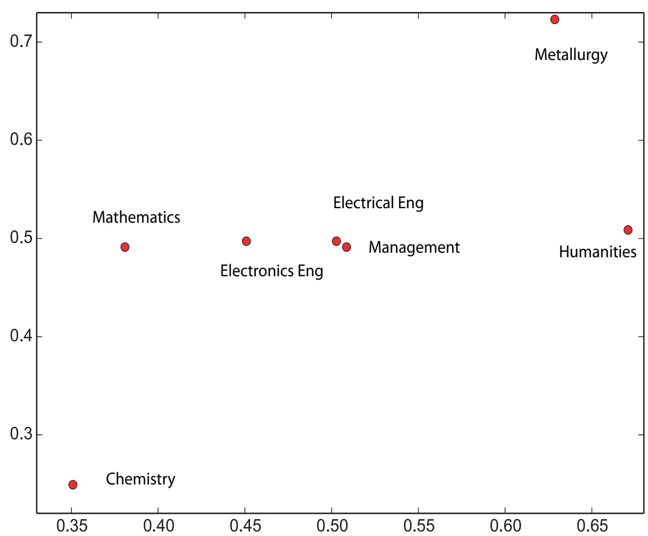
In this section we compare our Algorithm presented in Section III-G with two other clustering techniques. As discussed in Section III-A, in our dataset we have lectures belonging to subjects which are close to each other conceptually, *e.g.*Electrical Engineering and Electronics, Chemistry and Metallurgy, *etc.*To address the semantic overlap between the lecture videos from different subjects we employed the soft clustering techniques of Fuzzy C-Means clustering. In our system we use topic modeling methods to find latent topics from a collection of videos by leveraging LDA. We have generated 10 latent topics from the content of each cluster since in our previous work of lecture video recommendation, we had analyzed and discovered that 10 topics gives the best performance [18]. The topic distribution for the clusters are shown in Table I. As is evident from the table, the subjects which are bound to have overlaps in content like Chemistry and Metallurgy and Material Sciences or Electrical Engineering and Electronics Engineering or Humanities and Management have similar distributions with respect to relative proportions of the topics. However the subjects Mathematics and Metallurgy and Material Sciences also show a similarity in the distribution which should not be the case ideally. The reason for this maybe the inability of our system to deal with equations which are intrinsic parts of any academic material on Mathematics or science-based subjects . We cluster these videos based on the content generated in the form of topic distribution. We apply KL Div to deduce the similarity between the probability distribution of the topics. We use these KL Div values to calculate the fuzzy membership values for each of the videos as defined in Section III-F.

The input to the Fuzzy C-Means algorithm are the word frequency feature vectors for the videos and the membership values of the videos generated by $\mu(Vid_i, C_j)$. Fig. 2 shows the cluster centers after fuzzy C-means clustering. We had assigned each of the videos to a 2-dimensional coordinate system and the starting centers were randomly chosen using a random number generator. We performed the recommended 20 iterations of the clustering algorithm using the Python package Peach (http://code.google.com/p/peach/). The centers are plotted on the diagram with respect to this 2-dimensional coordinate system. As expected the cluster centers of Electrical Engineering and Electronics Engineering are in close proximity as there will be a substantial amount of videos from both the subjects that are close to each other semantically. Similarly the proximity of the centers of Mathematics and the Electrical Engineering as well as Electronics proves the semantic overlap among the lecture content of the three subjects. However, the center for the cluster of Management is very close to that of Electronics rather than being close to Humanities which should not be the case and needs to be explored further as a part of

TABLE I: Topic distribution in each cluster.

| Cluster | Topic 0 | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Chemistry** | 0.0576 | 0.1679 | 0.1499 | 0.0802 | 0.0601 | 0.1726 | 0.0386 | 0.1129 | 0.0772 | 0.0828 |
| **Electrical Engineering** | 0.095 | 0.1425 | 0.0559 | 0.1015 | 0.0917 | 0.1196 | 0.0714 | 0.0997 | 0.0872 | 0.1352 |
| **Electronics Engineering** | 0.1050 | 0.1129 | 0.0673 | 0.0221 | 0.1550 | 0.0666 | 0.1602 | 0.2208 | 0.0678 | 0.0221 |
| **Humanities** | 0.0746 | 0.1406 | 0.1152 | 0.0809 | 0.0685 | 0.0784 | 0.0865 | 0.1153 | 0.1241 | 0.1159 |
| **Management** | 0.0885 | 0.0851 | 0.0913 | 0.1063 | 0.1007 | 0.1044 | 0.0965 | 0.1166 | 0.0866 | 0.1239 |
| **Mathematics** | 0.0701 | 0.1344 | 0.0896 | 0.0572 | 0.1068 | 0.0656 | 0.1108 | 0.1385 | 0.1116 | 0.1154 |
| **Metallurgy and Material Sciences** | 0.0881 | 0.1996 | 0.0568 | 0.0868 | 0.0746 | 0.0840 | 0.0252 | 0.1514 | 0.0837 | 0.1499 |

future work. Next we compare our clustering algorithm with the following two state-of-the-art clustering techniques:




| Cluster | Cluster Center |
|---|---|
| **Chemistry** | (0.35087, 0.24910) |
| **Electrical Engineering** | (0.50277, 0.49723) |
| **Electronics Engineering** | (0.45088, 0.49723) |
| **Humanities** | (0.67088, 0.50872) |
| **Management** | (0.50862, 0.49123) |
| **Mathematics** | (0.38087, 0.49127) |
| **Metallurgy & Material Sciences** | (0.62862, 0.72319) |

Fig. 2: Cluster centres after fuzzy clustering.

- k-Means clustering technique as a representative of the hard-clustering techniques.
- Probabilistic clustering technique using Probabilistic Latent Semantic Analysis (PLSA) as another soft clustering technique.

### A. k-Means Clustering Algorithm

This is a partition based hard clustering technique. Suppose a data set, $D$, contains $n$ objects in the Euclidean space. Partitioning methods distribute the objects in $D$ into $k$ non-overlapping clusters, $C_1, \ldots, C_k$, that is, $C_i \subset D$, $\bigcup_{i=1}^{k} C_i = D$, and $C_i \cap C_j = \phi$ for $(1 \leq i, j \leq k, i \neq j)$. A centroid-based partitioning technique uses the centroid of a cluster, $C$, to represent that cluster. In case of k-means clustering, the centroid is the mean. The difference between an object $p \in C_i$ and $c_i$ , the representative of the cluster, is measured by $dist(p, c_i)$, where $dist(x, y)$ is the Euclidean distance between two points $x$ and $y$. As in the previous case, the input to the k-means algorithm is the word frequency feature vector of the videos. That means that each object is represented by their word frequency feature vector and we minimize the distance between the frequency vector of each object from that of the center.

### B. Probabilistic Clustering Algorithm

As we mentioned before, we compare our method with one more soft clustering technique. In this case we use Probabilistic Latent Semantic Analysis (PLSA). PLSA is a technique to statistically analyze the co-occurrence of words and documents [19]. We find the word distribution of the clusters using PLSA. Let $Z_{PLSA}(Wiki_{C_j}) = \{P(z_i)\}$, where $z_i$ are the latent distributions of topics according to PLSA, present in the Wikipedia representation of cluster $C_j$. Next we find the latent distributions of topics according to PLSA for each of the videos and denote it by $Z_{PLSA}(Vid_m) = \{P(z_k)\}$, where $z_k$ are the latent topics present in $Vid_m$. Then we find the KL Divergence of the topic distribution of the videos with each of the cluster representatives using the formula:

$$\begin{aligned} Sim_{PLSA}(Wiki_{C_j}, Vid_m) = \\ [\frac{1}{2}(D_{KL}(Z_{PLSA}(Wiki_{C_j}) \| Z_{PLSA}(Vid_m)) \\ + D_{KL}(Z_{PLSA}(Vid_m) \| Z_{PLSA}(Wiki_{C_j})))]^{-1} \end{aligned} \quad (3)$$

$Vid_m$ is assigned to the cluster $C_j$ having the maximum value of $Sim_{PLSA}(Wiki_{C_j}, Vid_m)$.

### C. Discussion of Results

We analyze the accuracy of our technique using two metrics: Precision and Sum of Squared Error (SSE). The precision values are calculated by the following formula:

$$Precision = \frac{\text{Total Number Of Correctly Retrieved Videos}}{\text{Total Number of Videos}} \quad (4)$$

The number of correctly retrieved videos will be the number of videos that match the subject of the cluster of the video and that of the ground truth for the video. In case of fuzzy clustering while calculating precision, we consider the subject of the cluster having the highest membership value as the cluster for the corresponding video. The precision values for our algorithm of fuzzy clustering and the two baselines are reported in Table IIIa. As can be seen, our method of fuzzy C-means clustering has a precision value higher than both the baselines.

Fuzzy clustering being a soft clustering technique, we can not assign a particular video to a single cluster alone. Hence calculation of the precision value for this clustering technique does not give an accurate measure of the effectiveness of the clustering technique, as there is a membership for belonging to each cluster. As a measure of accuracy for the clustering, we thus calculate the SSE for the Fuzzy Clustering as well as the baseline methods using the formula given below.

$$SSE(C) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{k} \mu(Vid_i, C_j) dist(Vid_i, C_j)^2 \quad (5)$$

where $n$ represents the total number of videos in the repository, $k$ is the total number of clusters, $\mu(Vid_i, C_j)$ represents the membership value of $Vid_i$ in cluster $C_j$ and $c_j$ is the center of $C_j$.

From Table IIIb, we can see that the error factor is considerably lower in case of our Fuzzy C-Means based method compared to the k-means method as well as the clustering based on PLSA. Hence we can see that the videos when clustered by the Fuzzy C-Means method are closer to the centroids than when clustered by k-means or PLSA. Thus our method of soft clustering of the videos produce tighter clusters than the two state-of-the-art clustering techniques used as the baselines as well as having higher precision value. It also captures the semantic overlap between the lecture videos from different subjects and can thus be effectively used to cluster lecture videos in a repository.

TABLE III: Evaluation of clustering.

(a) Precision of clustering

| Method | Precision |
|---|---|
| **Fuzzy C-Means** | 0.453 |
| **PLSA** | 0.214 |
| **k-Means** | 0.346 |

(b) SSE

| Method | SSE |
|---|---|
| **Fuzzy C-Means** | 0.2635 |
| **PLSA** | 0.385 |
| **k-Means** | 0.8189 |

## V. Conclusions

We designed and implemented a novel technique for clustering online lecture videos based on their content. In this attempt, we addressed the scenario where a lecture video can semantically belong to more than one cluster and hence be represented by a degree of membership to each cluster. We tested our clustering technique on the online lecture videos and we could perform the Fuzzy C-Means clustering as the number of clusters was known to us from the ground truth. To the best of our knowledge, this is the first attempt at clustering lecture videos based on the educational content present in the videos. We can extend this method to work on videos other than online lecture videos if the number of clusters is known at the onset and there is semantic overlap between the clusters. This clustering method can also be used in future for storing and indexing multimedia documents if there is a way to retrieve the semantic content of these documents.


## Acknowledgments

This research has been supported in part by Singapore's Ministry of Education (MOE) Academic Research Fund Tier 1, grant number T1 251RES1415, and also partially by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 15H06829.


## References


[1] M. M. Moore, A. Tait, P. Resta, G. Rumble, and Y. Zaparovanny, *Open and Distance Learning: Trends, Policy and Strategy Considerations*, 2002.

[2] H. Drachsler, K. Verbert, O. C. Santos, and N. Manouselis, "Panorama of Recommender Systems to Support Learning," in *Recommender systems handbook*, 2015, pp. 421–451.

[3] T. Liu and J. R. Kender, "Lecture Videos for E-learning: Current Research and Challenges," in *Proc. of the IEEE Sixth International Symposium on Multimedia Software Engineering*, 2004, pp. 574–578.

[4] A. Huang, D. Milne, E. Frank, and I. H. Witten, "Clustering Documents with Active Learning Using Wikipedia," in *Proc of the Eighth IEEE International Conference on Data Mining*, 2008, pp. 839–844.

[5] M. Vidal, G. V. Menezes, K. Berlt, E. S. de Moura, K. Okada, N. Ziviani, D. Fernandes, and M. Cristo, "Selecting Keywords to Represent Web Pages Using Wikipedia Information," in *Proc. of the 18th Brazilian Symposium on Multimedia and the Web*, 2012, pp. 375–382.

[6] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park, "iVisClustering: An Interactive Visual Document Clustering via Topic Modeling," in *Computer Graphics Forum*, vol. 31, no. 3pt3, 2012, pp. 1155–1164.

[7] S. Repp and C. Meinel, "Semantic Indexing for Recorded Educational Lecture Videos," in *Proc. of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops*, 2006, pp. 5pp – 245.

[8] C. Dorai, V. Oria, and V. Neelavalli, "Structuralizing Educational Videos Based on Presentation Content," in *Proc. of International Conference on Image Processing*, 2003, pp. II–1029.

[9] T. F. Syeda-Mahmood, "Indexing for Topics in Videos Using Foils," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 312–319.

[10] C.-W. Ngo, F. Wang, and T.-C. Pong, "Structuring Lecture Videos for Distance Learning Applications," in *Proc. of the Fifth International Symposium on Multimedia Software Engineering*, 2003, pp. 215–222.

[11] T. Liu, R. Hjelsvold, and J. R. Kender, "Analysis and Enhancement of Videos of Electronic Slide Presentations," in *Proc. of IEEE International Conference on Multimedia and Expo*, 2002, pp. 77–80.

[12] L. Tiecheng and J. R. Kender, "Spatial-temporal Semantic Grouping of Instructional Video Content," in *Image and Video Retrieval*, 2003, pp. 362–372.

[13] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[14] D. Gardner and M. Davies, "A New Academic Vocabulary List," *Applied Linguistics*, 2013.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[16] A. Huang, "Similarity Measures for Text Document Clustering," in *Proc. of the sixth New Zealand computer science research student conference*, 2008, pp. 49–56.

[17] J. Han, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[18] S. Basu, Y. Yu, V. K. Singh, and R. Zimmermann, "Videopedia: Lecture Video Recommendation for Educational Blogs Using Topic Modeling," in *Proceedings of the 22nd International Conference on MultiMedia Modeling*, 2016, pp. 238–250.

[19] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.