Categorization of YouTube Videos

Nitesh Gupta

Computer Science

Iowa State University

nkgupta@iastate.edu

Tanmay Gore

Computer Science

Iowa State University

tgore03@iastate.edu

**Abstract**

Categorization of web-based videos is an important task in video search and advertisement targeted applications. Unfortunately, the notion of a category varies depending upon the purpose, domain and level of abstraction. We therefore use unsupervised learning to cluster similar videos together based on their description and metadata such as video statistics. The proposed method also focuses on improving the intra-cluster similarity and reducing the training time. Various preprocessing algorithms such as TF-IDF and PCA are implemented to improve the quality of data. Isomap is later used to visualize the structure of this data. Now with reasonable dimensions, we cluster the data using K Means and Hierarchical Clustering. Since traditional accuracy measures cannot be applied to unsupervised learning, we determine the quality of clusters using Silhouette Coefficient.

**Keywords:** TF-IDF, PCA, YouTube, Isomap, K Means, Agglomerative Hierarchical Clustering, Clustering, Dimensionality Reduction.

1. **Introduction**

Categorization of videos is an increasingly prominent area of research, rising with the increasing number of videos shared through online platforms such as YouTube. Its applications are of paramount importance to video recommendation and Advertisement targeted applications. Some recent applications are identifying fake, influential or misleading videos. However, the categorization of videos to poses a great challenge. One major concern is that categories vary based on their purpose, domain and level of abstraction. Also, videos can belong to multiple categories making it difficult to define a fixed cluster size.

In this paper, we try to obtain a very good clustering model inclined towards resolving above mentioned problems where the quality of a model is determined by the training time and similarity within clusters.

This project focuses on four main aspects – Data Preprocessing, Data Visualization, Clustering and evaluating quality of Clustering. The dataset obtained from Kaggle about trending youtube videos contains various features such as title, description, tags, and metdata such as likes, dislikes, comments and views.

In the first step we preprocess the data by converting the textual information to TF-IDF. This greatly increases the dimension, hence to perform machine learning in reasonable amount of time, we reduce the dimensions using PCA. The visualization step enables us to identify the structure of the data. We use Isomap to further reduce the dimensions to 2 so as to visualize the structure of the data. Based on this structure, we decided to use K Means and Agglomerative Hierarchical Clustering to obtain the clusters. Once the clusters were obtained, we validated its quality using Silhouette Coefficient.

* 1. **TF-IDF**

TF-IDF stands for *term frequency-inverse document frequency*, and the TF-IDF weight is often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance of a word is proportional to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

The following are the equations for TF and IDF respectively:

… (1)

… (2)

The TF-IDF is the product of equation (1) and (2).

* 1. **Principle Component Analysis**

Principal Component Analysis (PCA). PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that first principal component has the largest possible variance and each succeeding component has the highest possible variance under the constraint that it is orthogonal to the preceding components. A data point can have thousands of features. The running time for any machine learning algorithm depends on the number of data points in the dataset and the dimensionality of a data point. Through PCA it is possible to achieve the faster running time for the dataset by reducing the dimensions of a data point while preserving most of its value. Thus, PCA plays an important role in the situation when there is a need to perform various algorithms on a dataset to obtain the comparison results.

* 1. **Isomap**

Isomap is a nonlinear dimensionality reduction method. It is one of several widely used low-dimensional embedding methods. Isomap is used for computing a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. The algorithm provides a simple method for estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point’s neighbors on the manifold. Isomap is highly efficient and generally applicable to a broad range of data sources and dimensionalities.

* 1. **K Means**

It is an unsupervised learning technique (i.e. no labels for the data are provided) that is used for clustering similar data into groups. The number of groups made depend on the value k that we choose.

* 1. **Agglomerative Hierarchical Clustering**

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Agglomerative is a bottom up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Decision to merge cluster happens in greedy fashion and depends upon the metrics used to calculate distance between clusters. The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations. Single-linkage clustering happens in a bottom up fashion where two clusters that contain the closest pair of elements not yet belonging to the same cluster are merged. In complete-linkage clustering, the link between two clusters contains all element pairs, and the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other. The shortest of these links that remains at any step causes the fusion of the two clusters whose elements are involved. The method is also known as farthest neighbor clustering. In ward linkage, the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. If the objective function is sum of square errors, then it is called ward’s minimum variance method.

* 1. **Silhouette Coefficient**

line Euclidean distance. Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

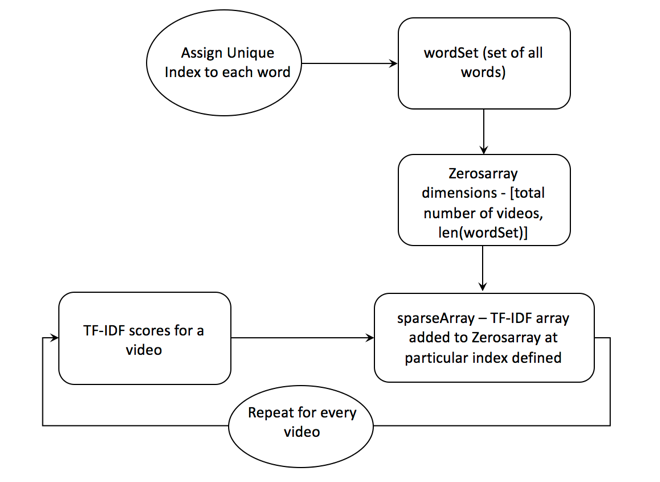
1. Methodologies

To categorize the videos, we perform three major steps: (1) Preprocess Data (2) Clustering (3) Measure the quality of clusters. However, after preprocessing, we can visualize the data to better understand the structure. Each of the three steps focuses on getting a good quality clustering in terms of running time and intra-cluster similarity.

* 1. Preprocessing Data

The YouTube data contains various features such as title, description, tags, channel name, comments, views, likes and dislikes. To process the textual features, we convert them to numerical representation using TFIDF. We first assign a score for every word in video tags. This would enable us to have a vector representation of every feature that was obtained. Once the scores were obtained, a set containing every word was made. Let’s call this set “*wordSet*”. Every word in *wordSet* was assigned a unique index (*index word*). Next, a sparse zeros array containing dimensions [number of videos, length of *wordSet*] was generated. We iterated through the scores for every word for each video and added them to the specific index as determined by *index word*. This was done for every video in the sample space. Thus, the array contained the TF-IDF vectors for each video.

One benefit of TFIDF is that it retains the contextual information by assigning weighted scores to each word where higher weightage is given to unique, meaningful words and lower weightage to stop words. However, this significantly increases the dimensions (50000 in our case). We therefore apply dimensionality reduction algorithm called PCA to reduce the dimensions. Since the matrix obtained after TFIDF is a sparse array matrix, PCA can greatly reduce the dimensions (3000 in our case) with 95% of the variance retained.



**Figure 1 Process of TFIDF Algorithm**

* 1. **Visualizing the Data**

Given the unknown nature of the data, it is crucial to visualize the data so as better determine the clustering algorithm to be used. However, the reduced dimensions (3000 in our case) are still too much for humans to visualize. We therefore apply another dimensionality reduction algorithm called Isomap to visualize the structure of the data. Isomap further reduces the dimensions to 2 by estimating the structural information of data.

* 1. **Clustering the Data**

Next K Means and Agglomerative Hierarchical Clustering is applied to cluster the data. K Means is applied for different values of clusters ranging from 1 to 25, 50, 100, 300, 500, 700 and 1000. For each of the clusters, the model is trained 20 times independently with random seeds to overcome local minima.

* 1. **Measure Quality of Clusters**

Due to lack of labeled data, traditional accuracy measures cannot be applied to clustering. Therefore, we measure the quality of clusters obtained using Silhouette Coefficient. It determines the relationship between clusters and measures the similarity of data points within the cluster. For each of the model instance we determine the silhouette coefficient and plot a silhouette plot which depicts the density of each cluster and misclassification of labels.

1. Results and Discussions

Since the entire process is divided into 4 logical steps, namely Data Preprocessing, Data Visualization, Clustering and Quality Evaluation we will be discussing the results obtained in each of the 4 steps.

* 1. Data Preprocessing

To convert the textual features to numerical format, TFIDF was applied which converted the original 20 dimensions to 50000 dimensions. Since IDF was applied during the conversion, contextual meaning of data was retained using weighted scores.

To reduce the training time of machine learning algorithms, dimensionality reduction algorithm called PCA was applied. Due to the sparse data obtained from TFIDF, PCA reduced the dimensions from 50000 to 3000 with 95% of the variance retained.

* 1. **Data Visualization**

To better understand the structure of data, another dimensionality reduction algorithm called Isomap was applied. Since Isomap reduces the dimensions by retaining the structural information of the data, we could visualize the data in 2 or 3 dimensions.

By Plotting the PCA reduced data and non PCA reduced data using Isomap we could see that PCA helped to normalize the data ensuring better clustering.

A screenshot of a cell phone

Description generated with high confidence

**Figure 2 : Structure of Data when PCA is applied**

A screenshot of a cell phone

Description generated with very high confidence

**Figure 3 : Structure of Data when PCA is not applied**

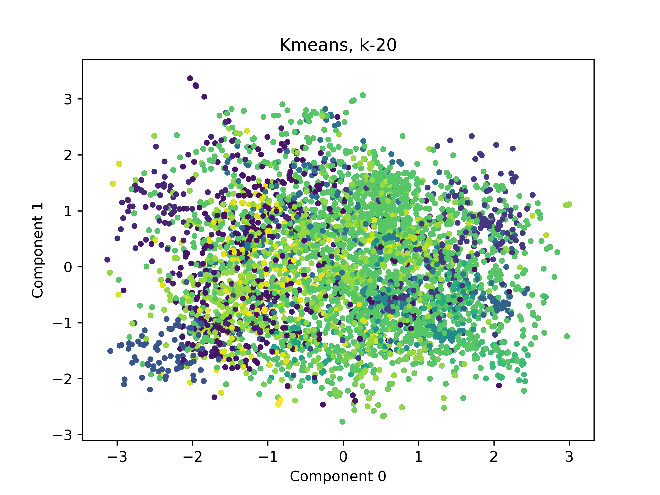
* 1. **Clustering**

We performed clustering using K Means and Agglomerative Hierarchical Clustering. Both the algorithms were performed for different values of clusters and the model was trained for 20 epochs for each cluster value. The plots we got for cluster value 4 and 20 using K Means are as follows.

A picture containing text

Description generated with high confidence

**Figure 4 : Clusters for K=4**



**Figure 5 :Clusters for K=20**

Plots we got using Hierarchical Clustering were as follows.

* 1. **Quality Evaluation**

Since classical accuracy measures do not work on unsupervised learning due to lack of labels, we used Silhouette Coefficient. The Silhouette value obtained for different clusters using K Means and its training time were as follows.

|  |  |  |
| --- | --- | --- |
| Clusters | Silhouette Coefficient | Training Time |
| 2 | -0.006 |  |
| 5 | 0.002 |  |
| 10 | 0.009 |  |
| 15 | 0.015 |  |
| 20 | 0.022 |  |
| 50 |  |  |
| 100 |  |  |
| 300 |  |  |
| 500 |  |  |
| 700 |  |  |
| 1000 |  |  |