VIDEO CLASSIFICATION USING TEXT ANALYTICS

Aarti Raghani
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Aditi Sawant
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Kajol Chawla
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Revati Pathak
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

Mrs. Lifna C.S.
Assistant Professor
Dept. of Computer Engineering
V.E.S.I.T
Chembur, Mumbai, India

***Abstract*—With the advent of the Information Age, there has been an exponential growth in the number of videos available for any given topic. Hence, internet users are relentlessly in search for foolproof Video Classification Algorithms. The existing state-of-the-art techniques do not provide a satisfactory summary for ranking videos. The objective of the paper is to rank videos using Text Summarization techniques. This application will provide the users with the most relevant videos and their summaries, thereby saving time and increasing efficiency.**

Keywords—Text Summarization; Video Ranking; Video Classification.

I. Introduction

In today's digital era, the internet is the reliable source for obtaining videos. There has been an enormous growth in videos available for any given topic. A query for any topic yields a heap of results and, it is not easy to manually discard irrelevant results. Hence, it is essential to automatically summarize the videos to get a gist of their content.

The available video analytics systems analyze videos by using Image Processing techniques. The efficiency of these Image-based Summarization techniques heavily depends on the quality of the videos. Hence, these techniques have the following drawbacks: (1)Blurred Images; (2) Poor Illumination Effect; (3) Need for large bandwidth; (4) Storage requirements; (5) Output is also in the form of video. These constraints degrade the quality of the summaries generated by the systems. This led us to consider Text-based Summarization techniques for summarizing the videos and evaluating the efficiency.

In short, our proposal can be summarized as follows: (1) Search for topmost videos using input keyword; (2) Extract audio from the videos; (3) Convert the audio files to text; (4) Generate document summaries; and (5) Re-rank the videos based on the summaries.

II. Literature Survey

Towards finalizing Text-based summarization techniques, a thorough survey was performed. As a part of the survey, the following conclusions were extracted from the respective papers.

Paper [1] discusses two kinds of Text Summarization approaches; (1) Extractive Summarization - extract the most relevant sentences from the given text document and include them in the summary. (2) Abstractive Summarization - original sentence from the given document is replaced with a sentence with similar meaning. While abstractive summaries are closer to the summaries generated by a domain expert, extractive summarization techniques are faster and more efficient for summarizing documents with multiple sentences. Since an exhaustive study is required to validate summaries of large documents, extractive summarization approach was selected and further studied.

In paper [2], an Extractive Text Summary is generated from the list of top-ranked sentences by using Sentence Scoring Method. For Sentence scoring, the following factors are considered: (1) Word frequency; (2) Sentence Position; (3) Cue words; (4) Title Similarity; (5) Sentence length; (6) Proper noun; and (7) Sentence reduction. After each sentence is scored, the sentences are arranged in the descending order of their score value. From the top-ranked sentences, a number of sentences are chosen to be included in the summary.

In paper [3], lexical chains are used as source representation for summarization. For this, WordNet is used to find relatedness among words. Words of the same category are linked through semantic relations like synonymy and hyponymy. The algorithm differs from other approaches as it creates chains for words using relatedness criteria and merges different segments using strong criterion, i.e. chains will be merged if they contain a common word with the same sense. The problem of similarity between sentences can be solved by

978-1-5386-4304-4/18/$31.00 ©2018 IEEE.

using the chain representation approach. The final summary can be extracted based on different heuristics functions i.e. Heuristic 1- For each chain in the summary representation, choose the sentence that contains the first appearance of a chain member in the text. Heuristic 2-The sentence containing the first appearance of a representative chain member is chosen. Heuristic 3- A frequently discussed topic may have its chain spread all over the document.

In this paper [4], an indexing structure based on the context of the document is proposed. The WordNet database helps in identifying similarities between different sentences in the document. Later, TextRank Algorithm is used to find a context-sensitive indexing weight of each term. On generating a graph of a document, each edge gives the lexical association between the terms corresponding to the vertices.The similarity between words is found using context-based indexing weights. Then this similarity between words is used to find similarity between sentences. Thus, for each sentence in the document, the sentence vector is built. Later sentences are extracted based on higher sentence score. So, this approach can be used to retrieve results within a short span.

This is also a graph-based unsupervised algorithm for extractive summarization [5]. It uses TF-IDF (Term Frequency- Inverse Document Frequency) to calculate how important a word is in multiple documents. A vector is defined for every sentence where the entry in the vector for every word is the value of its frequency in the sentence multiplied by its idf. The dimensionality of this vector is equal to the number of all words in the targeted language. An idf-modified-cosine function is defined to compute the similarity between two sentences. A cosine similarity matrix is formed with entries for the similarity calculated above. A threshold value is chosen to choose only the significantly similar sentences and discard the rest.

In paper [6], a recursive TF-ISF based method that takes into account the local context of a sentence is proposed. The context is defined as the previous and next sentence of the current sentence. On comparing this method to the TF-ISF baseline, statistically significant improvements in the results were found.

In this paper [7], TF-IDF method is used to generate a summary. Here, features of the document are extracted by obtaining the scores for the sentences in the text document based on their importance with a value between zero and one. Thirdly, it includes sentence selection and assembly, where the sentences are stored in descending order of the rank, and the highest ranked sentences are considered for the summary. TF-IDF finds the importance of a word in the document and to control this value, the frequency of the word in the document is considered. The frequency term is the number of occurrences of a term in a document. Inverse Document Frequency is calculated by dividing the total number of documents by the number of documents in which the term occurs.

TABLE I. SUMMARY OF LITERATURE SURVEY

	Pros	**Cons**
[2]	Sentence Scoring method is used so accuracy is much higher than when considering only Tf-Idf or sentence ranking. Any number of extensions for scoring techniques can easily be added.	Query based summarization is not supported. It involves summarization with title which involves that if title word is present in sentence, then that sentence will get higher priority rather than considering the relevance of title with sentence.
[3]	The problem of similarity between sentences produced in the summary is overcome. Also, the sense of the word based on its position in the sentence is considered which is done using lexical chains.	Large sentences are more likely to be a part of summary.This approach does not concern the length and detail of the summary produced, which may act as noise.
[4]	It builds the index by considering context of the document. This is different compared to earlier methods that consider the terms for building the index. This approach is found to give better results than previously used approaches.	It fails to consider related terms i.e verbs, adjectives of words. It uses a linear function to calculate similarity between two sentences. It is found that a cosine function produces better results.
[5]	Sentences subsuming the information of other sentences get higher scores than individual sentences; thus a compact summary is formed. Prevents unnatural boosting of sentence score by an irrelevant topic.	Basic LexRank uses threshold to establish links between the sentences. Thus, improper threshold values may result into information loss or inclusion of irrelevant details in the summary making it bigger.
[6]	Summary is generated by considering relevance between previous and next sentence. It is achieved by defining recursive ranking function.	It does not consider factors for sentence scoring. Recursive TF-ISF algorithm is purely query based approach. This algorithm not considers factors such as Combination of word frequency in document, Sentence positional value ,Sentence length and Proper noun etc
[7]	The result of this research produces 67% accuracy with three data samples which are higher compared to the other online summarizers.	Relevance with the title factor is not considered while generating summaries. There is a need for involving more respondents to evaluate the system by determining the number of correct, wrong, or missed sentences within the summary.

III. Proposed System

The proposed system has the following modules.(1) Search Module; (2) Audio Extraction Module; (3) Speech-to-Text Conversion Module; (4) Text Summarization Module and (5) Re-ranking Module.

A. Search Module

The query input by user is searched for in the database. If found, relevant video links and generated summaries are displayed to the user. Else, the query is redirected to YouTube.

B. Audio Extraction Module

The video for which summary is to be generated is the input and the audio is extracted from it.

C. Speech-to-Text Conversion Module

The audio extracted in the preceding step is converted to textual format. This text document forms the input to the Summarization Module.

D. Text Summarization Module

The text document obtained is summarized using extractive summarization technique. The factors that we consider for sentence scoring are: (1) Position of sentence; (2) Length of Sentence; (3) Similarity with the Title; (4) Proper Nouns; (5) Lexical Chains; (6) Term Frequency-Inverse Sentence Frequency. These scores are combined and the best scored sentences are ranked better.

E. Re-ranking Module

After summaries are generated for the videos, they are re-ranked on the basis of their summaries' relevance to the query.

The diagram shown in Fig. 1.describes the proposed system. The user inputs the search query over the internet, query is searched for in the database and if the corresponding query is accessible in the database, the results, that is, the relevant re-ranked videos, along with the videos links and the summary of those videos is displayed to user. But if query is not available, it is redirected to YouTube and the listed videos are pre-processed by initially converting them to text document, summarizing their content and finally re-ranking them based on the video content and the input query.

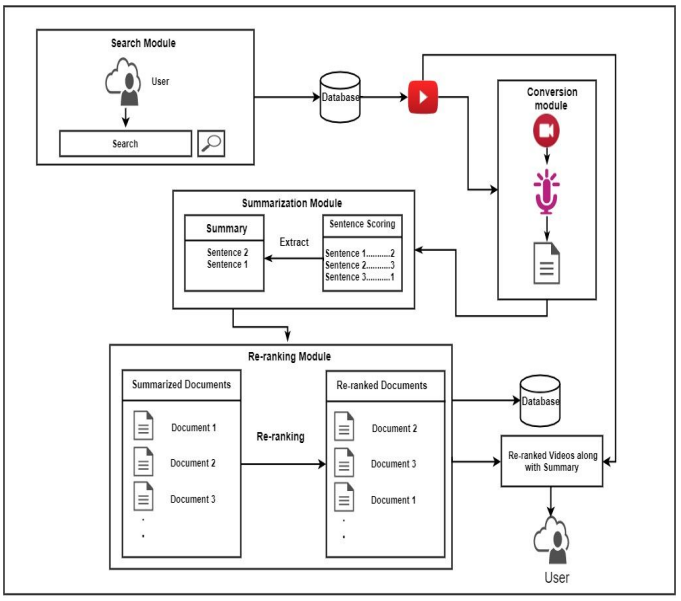

Fig. 1. Proposed System Diagram

IV. Conclusions

Our system includes scoring sentences in the document based on the sentence position, similarity with the title, sentence length, proper nouns, etc. The approach proposed in a paper [1] doesn't take into account the similarity between different sentences in a document. This drawback can be overcome, by using lexical chains approach, which gives scores to sentence based on chains form- extra strong chains, strong chains, medium strong chains.

Instead of considering the frequency of words and cue words, proposed system consider the local context of a sentence using recursive TF-ISF. This method improves upon the approach proposed in paper [1] and gives a more efficient summary.

The system can be tuned further as an application for Cyber Security Cell to address Terrorism by incorporating multilingual module

Acknowledgment

We are sincerely thankful to Mrs. Sujata Khedkar, Associate Professor, Dept. of Computer Engineering, V.E.S.I.T. for her valuable insight and guidance in developing this approach.

References

[1] Shimpikar, Sheetal, and Sharvari Govilkar. "A Survey of Text Summarization Techniques for Indian Regional Languages." *International Journal of Computer Applications* 165.11 (2017).

[2] Raju, T. Sri Rama, and Bhargav Allarpu. "Text Summarization using Sentence Scoring Method." (2017).

[3] Barzilay, Regina, and Michael Elhadad. "Using lexical chains for text summarization." *Advances in automatic text summarization* (1999): 111-121.

[4] Pawar, Dipti D., M. S. Bewoor, and S. H. Patil. "Text Rank: A novel concept for extraction based text summarization." *International Journal of Computer Science and Information Technologies* 5.3 (2014): 3301-3304.

[5] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of Artificial Intelligence Research* 22 (2004): 457-479.

[6] Doko, Alen, Maja Stula, and Darko Stipanicev. "A recursive TF-ISF Based Sentence Retrieval Method with Local Context." *International Journal of Machine Learning and Computing* 3.2 (2013): 195.

[7] Christian, Hans, Mikhael Pramodana Agus, and Derwin Suhartono. "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)." *ComTech: Computer*