DETECTION AND PREVENTION OF COPYWRITING MULTIMEDIA CONTENTS IN YOUTUBE USING MULTI-KEYWORD RANKING ALGORITHM

Dr. R. RAJU [a]
Information Technology.Sri Manakula Vinayagar Engineering College
Puducherry, India
rajupdy@gmail.com

BETTINA O'BRIEN[b]
Information Technology.Sri Manakula Vinayagar Engineering College
Puducherry, India
bettyobrien04@gmail.com

R. KALAISELVI[c]
Information Technology.Sri Manakula Vinayagar Engineering College
Puducherry, India
gayucadbury2013@gmail.com

Abstract*—YouTube, an American video sharing website has become the most preferred site for watching videos online, with millions of content creators. YouTube is often said to be the second largest search engine in the world after Google itself, where people spend hours watching videos, generating billions of views. Popularity dynamics of YouTube videos highly depends on the meta-tags, number of view counts and the social dynamics which indicates the interaction between the content creators (channels) and YouTube users. Social networks like videos spreads and sensitivity of YouTube meta-level features provide an important impact on the popularity of videos. Our dataset includes all the videos from the YouTube site and are imported into a centralized database of our system using EmbedSource. In the context of video analysis, each video is characterized into four attributes: Title, Category, Description, Embedded links. The non-frequent search words used by the users are converted to its equivalent signature using the Digital Signature Algorithm (DSA). We propose a Multi -Keyword Ranking (MKR) algorithm which scales the popular or originally created video using attributes such as view count, number of subscribers, title, description and the comments and ratings generated by users computed using sentimental analysis. The outcome of this proposed work would be an optimized search of the original video that was actually created by the channel.

Keywords*—YouTube, metadata, data mining, sentiment analysis, ranking.

I. INTRODUCTION

YouTube, the best website for sharing videos online has become the most useful and attractive site among users through the years. It is the second largest search engine that is owned by Google. Its popularity has increased rapidly because of its easy use and simplicity to create and share videos. There are many web platforms that are used to share non-textual content such as videos, images, animations that allow users watch and share among other users on the same site. YouTube is probably the most popular of them, which is ranked to be the second largest search engine after Google and is owned by Google itself, with millions of videos uploaded every minute. YouTube allows users to upload, view, rate, share, add to favorites, report, comment on videos, and subscribe to other users. There are many models and methods for predicting the popularity dynamics for videos which mainly includes the view counts, number of subscribers, meta-level features like title and description of the videos. Ranking is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. The internal ranking algorithm used by YouTube is the one defined by Google for ranking of contents and image. The current video ranking algorithm used by YouTube uses crucial factors such as channel keywords, video title, video description, video tags, video quality, user experience metrics, watch time, view count, thumbnails, transcript.

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker in terms of a given service over a website or regarding a particular product. It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. The opinion from the users regarding a searched video is the important input of the ranking algorithm we have proposed. A digital signature is a mathematical technique used to validate the authenticity and integrity of a message, software or digital document. The embed link created for each video in the YouTube search engine would be converted into an equivalent signature and encrypted.

II. RELATED WORK

WILLIAM HOILES ET AL., "ENGANGEMENT DYNAMICS AND SENSITIVITY ANALYSIS OF YOUTUBE VIDEOS.", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE, 2017[1]

In this research, they have analyzed the sensitivity of YouTube videos based on several machine learning algorithm, out of which they have concluded that the Extreme

978-1-5386-2447-0/18/$31.00 ©2018 IEEE

Learning Machine (ELM) has the effective results. The ELM algorithm is more efficient because of its two important computations, Estimation of meta-level features and prediction of view count. To predict the view counts of a particular video that affects its popularity, they have solved a mathematical equation called Sum of Square Derivatives (SSD). The video with the highest SSD value promotes to the effective prediction of view counts. They have also concluded that the meta-level features on the view counts for videos which mostly affect the popularity of the videos are, first day view counts, number of subscribers, thumbnail, google hits, video category, title length. Engagement dynamics refers to the interaction between the content creators (channels) on YouTube and the users who view the content. The causal relationship between the view count and the number of subscribers is shown with the help of a graph using Granger Causality technique. When a video is posted, optimization is carried out by mapping the number of subscribers both before and after 10 to 15 days of posting the video.

LIQIANG NIE ET AL., "PERPETUAL ATTRIBUTES OPTIMIZATION FOR MULTIVIDEO SUMMARIZATION", IEEE TRANSACTIONS ON CYBERNETICS, IEEE, 2015[2]

This article proposes perceptual multi-attribute optimization which jointly refines multiple perpetual attributes in a multi-video summarization process. Video summarization is a useful technique in many computer vision and multimedia applications. Summarizing multiple handheld video is a challenging task because they have different styles and semantics and also, different degree of shakiness. Some attributes regarding videos are video aesthetics, coherence, stability. The problem arises when different gadgets are used for capturing videos, where stabilization and summarization of the video with all the supporting attributes has to be achieved. For each frame in the video, the most semantically important regions are discovered using a weakly supervised learning framework. A manifold embedding algorithm is used to discover the characteristics of videos by leveraging the semantics of the video tags. To reconstruct the summary or semantics of the entire video, by selecting the representative key frames, an active learning algorithm has been used. Based on the proposed video summarization model, they have concluded that a collection of very long videos can be condensed into a significantly and semantically representative shorter video clip, which in turn increases efficiency in analyzing the resultant video.

CEDRIC RICHIER ET AL., "PREDICTING POPULARITY DYNAMICS OF ONLINE CONTENTS USING DATA FILTERING METHODS", FIFTY-FOURTH ANNUAL ALLERTON CONFERENCE, ALLERTON HOUSE, UIUC, SEPT. 27 TO 30, 2016 [3]

This article proposes a new prediction technique to predict the popularity evolution of YouTube videos. Initially, to predict the view count that tremendously affects the popularity of YouTube videos, a study on the classification of videos were done. The important features or factors that contribute to the popularity dynamics of videos were identified through classification, based on different categories was considered to be a filtering method to identify the factors responsible for popularity dynamics. The proposed prediction method includes three ways of processing. First, the classification of video are based on the view count patterns using seven mathematical models. Second, groups the videos by range of popularities. And finally, the last method includes the combination of the first and second model, where both the classification and grouping are carried out as one model. Then the proposed prediction model is compared with the previous models like Simple Huberman(S-H) and Multivariant Linear (ML) and it provides a conclusion that this classification increases the popularity accuracy over the other two models. Furthermore, evaluation has been done, whether to add popularity criteria in the classification that would result in more accurate predictions. The evaluated prediction method reduces the average prediction error but issue related to variance occurs.

TOMASZ TRZCINSKI ET AL., "PREDICTING POPULARITY OF ONLINE VIDEOS USING SUPPORT VECTOR REGRESSION", IEEE TRANSACTIONS ON MULTIMEDIA, IEEE, 2017. [4]

In this work, a regression method to predict the popularity of an online video based on the number of view count has been proposed. The regression technique called Support Vector Regression (SVR) provides a stable prediction result. It uses a dataset of 24,000 sample videos both from Facebook and YouTube, and prediction is done using the Gaussian Radial Basis function of the Support Vector Regression technique. The dynamics' metric of the visual features regarding a video can be useful for popularity prediction before content publication. The improved version of the prediction technique uses a combination of the early distribution patterns on the number of view count with the social and visual features for accuracy. The visual features are those features computed using seven computer vision algorithms applied on raw video data. This technique follows a method where each video has been divided into frames and the visual features of each frame has been recognized and captured, from which the optimization of searching a particular feature based on the visual factor has been done effectively. Implementation of support vector regression algorithm has been done by their own code in python, using Scikitlearn package. The results suggest that using only visual features computed before the publication of the video can be helpful to predict future video popularity. The output of using SVR has provided the best popularity prediction technique when the social features, view count and the visual features have been combined.

C´EDRIC RICHIER et al., "FORECASTING ONLINE CONTENTS' POPULARITY", RESEARCHGATE PUBLICATION, 277603649, MAY 2015 [5]

978-1-5386-2447-0/18/$31.00 ©2018 IEEE

This article capitalizes the behavior and the functionality of view counts of YouTube videos which likely adds advantage in increasing the popularity of videos. They have used six bio-inspired mathematical models, in which 90% of videos are associated with one of the models with mean error rate less than five. Their previous study suggests that the view count is the most important metric for a videos' popularity, but further study proves that the correlation between other metrics such as comments, adding to playlists or favorites, ratings and the view count also contribute a major portion to popularity. They use three methods for filtering the metrics, out of which filtering by method provides a better result than ML baseline. But filtering based on popularity provides best results only for certain days after which the popularity fades. Hence they have concluded by combining both the filtering technique, with method and popularity, which would give the atmost outcome of predicting the popularity based on a videos' metrics.

YOUMNA BORGHOL et al., "CHARACTERIZATION AND MODELING POPULARITY OF USER GENERATED VIDEOS" [6]

This article proposes a model that records the important properties of a video that contributes to the popularity dynamics. It develops a framework to study the popularity dynamics and a characterization of it. The analysis of the sampling from recently uploaded videos provides a dataset which is relatively unbiased. Popularity dynamics is measured based on the views rated to a video over the time period before and after the video was uploaded. The relative properties of the videos promoting popularity are dynamic in nature and seem non-stationary. The proposed model in this article captures the popularity dynamics of recently uploaded videos as they evolve in time, including their key measures such as statistics and the increase in view counts and total view distribution. Sampling's use concerns with the biggest challenge for researchers to study on large volumes of content, but sampling may provide datasets biased towards content related to either long term or short term popularity. They have concluded that there is a substantial churn in the properties that affect the popularity and thus the future popularity of videos are not to be relied on the current dynamics. Their future works includes analysis of large scale tests on this model and from various social media's contents. The extension of their proposed work would be a study on the factors which directly or indirectly affect the popularity's characterization.

AHMED ABDELSADEK et al., "DISTRIBUTED INDEX FOR MATCHING MULTIMEDIA OBJECTS", SIMON FRASER UNIVERSITY, 2014 [7]

This article demonstrates the design and assessment of DIMO, Distributed Index for Matching Multimedia contents. The distributed multimedia objects are those contents that are distributed across different special addresses. It provides a function of finding the nearest neighbors on a large scale index. It also paves a way to define specific operations relating to applications to further process the computed neighbors. The MapReduce programming model-supported infrastructure over a distributed environment provides a novel method for searching, partitioning and storing high scale datasets. In this article, they have evaluated it on Amazon clusters by implementing it on 128 machines. DIMO produces high precision data points extracted from images and utilize various computing resources. Their results have proved that DIMO outperforms RankReduce in terms of precision of the computed nearest neighbors.

Y. DING, et al., "BROADCAST YOURSELF: UNDERSTANDING YOUTUBE UPLOADERS", in PROC. OR THE ACM SIGCOMM CONFERENCE ON INTERNET MEASUREMENT. NEW YORK, NY, USA: ACM, 2011, pp. 361–370. [8]

In this paper, they provide a comprehensive study on YouTube uploaders, the central agents in the YouTube phenomenon. They conduct extensive measurement and analysis and obtain an in-depth understanding of YouTube uploaders. They estimate YouTube scale and examine the uploading behavior of YouTube users. Furthermore, they have examined whether YouTube users are really broadcasting themselves, via characterizing and classifying user generated videos and user copied videos. Moreover, they demonstrate the positive reinforcement between on-line social behavior and uploading behavior. The number of uploaded videos clearly follows a Zipf distribution. The maximum, mean, and median of uploaded videos are reported based on the analyzed statistics. Among these uploaders, the most active 20% of the uploaders contribute 72.5% of the videos, which largely follows the famous 80-20 rule. Perhaps the most surprising result in the paper is the discovery that much of the content in YouTube is not user generated. They have found that 63% of the most popular uploaders are primarily uploading UCC (User Copied Content), and that UCC uploaders on average upload many more videos than UGC (User Generated Content) uploaders. The results and observations have been used as the first step towards an automatic algorithm for classifying UGC and UCC content.

SHUMEET BALUJA et al., "VIDEO SUGGESTION AND DISCOVERY FOR YOUTUBE: TAKING RANDOM WALKS THROUGH THE VIEW GRAPH", GOOGLE, Inc. MOUNTAIN VIEW, CA, USA. [9]

In this paper, they have presented a novel method based upon the analysis of the entire user–video graph to provide personalized video suggestions for users. The resulting algorithm, termed Adsorption, provides a simple method to efficiently propagate preference information through a variety of graphs. They have extensively tested the results of the recommendations on a three month snapshot of live data from YouTube. By using the adsorption algorithm, we were able to improve the expected efficacy of suggestions in YouTube. The most commonly used heuristics, recommending the overall most popular videos and/or the most co-watched videos did not perform as well. Because of

978-1-5386-2447-0/18/$31.00 ©2018 IEEE

the short half-life of videos on YouTube, the large number of uploads, and the exposure that users have to new and popular videos, recommending anything other than commonly viewed videos was not guaranteed to provide improvement. However, the fact that trends can be found provides strong evidence not only in favor of a graph-based algorithm, but that there is indeed interesting usage information to be mined from YouTube beyond the casual viewing of popular videos. In addition to this, they have presented a method to backtest recommendation systems through historical log-analysis. They have also summarized the performance of the proposed algorithm on the data set harvested from YouTube logs. They have executed each algorithm (the reference algorithms and the two variants of the adsorption algorithm) to produce a list of related videos for each video node, and for each user, up to 100 recommendations were made, ensuring that the algorithms never recommended videos already watched during the training period.

YIPENG ZHOU et al., "VIDEO POPULARITY DYNAMICS AND ITS IMPLICATION FOR REPLICATION", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 17, NO. 8, AUGUST, 2015. [10]

In this work, based on the big data available from a real-world VoD system, they have studied video popularity as a dynamic system. They have found that as a percentage of total views, replays of online videos are insignificant. So one way to think of video popularity can be based on view count, or the number of users reached eventually. At any particular time, however, the dynamics of video popularity depends on the age of the video, following a pattern for each type of video. According to these observations, they have proposed a mixed strategy to determine the videos cached on CDN servers. Since the mixed strategy takes both age-sensitive videos and popularity stable videos into account, most popular videos can be captured to achieve high hit rate. The study of popularity is not restricted to videos. Entropy is an effective method to evaluate the frequency how users replay a certain video. Their work focuses on the user behavior, dynamic changes of video popularity and their implications for video cache replacement strategies.

III. PROPOSED METHODOLOGY

We propose a Multi -Keyword Ranking (MKR) algorithm which scales the popular or originally created video using attributes such as view count, number of subscribers, title, description. Study on the factors that affect the ranking of YouTube videos branches into:

- Internal factor
- External factor

The internal factors includes filename in which the video has been saved before uploading it into the YouTube server, video title, video description, number of view counts, number of likes, number of dislikes, number of subscribers, number of add to playlists or favorites, number of comments, time watched which is the time a user stays online to watch a particular video, subscription driven, transcript which includes the ability to add text to the video. The external factors that affects the popularity of videos are social network signals or social bookmarks which includes comments or likes given for a particular channel's video on different social media like, Fb, G+, Ln, Twitter and Embeds or Back Links which includes posting that video on the user's blogs or any other community websites that they are involved. We have reduced these factors to a total of five:

- Watch time
- Keyword relevance
- Meta- tags (description, title)
- Reactions of users (comments)
- Google hits

There are four modules involved:

SERVER CONSTRUCTION

Server construction requires the same layout as such as the YouTube site, where the options of uploading or removing videos by the admin are included.

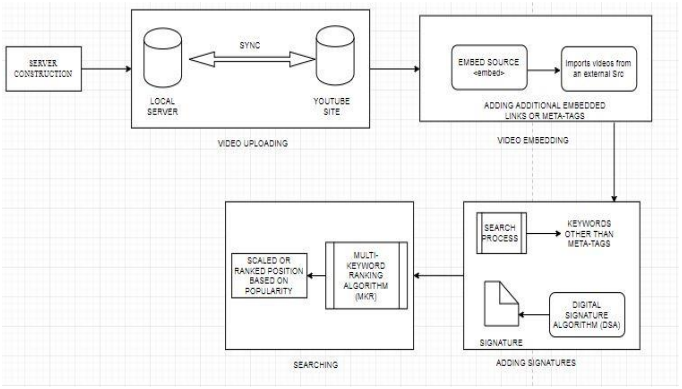


Fig.1 System Architecture

VIDEO UPLOADING and EMBEDDING

Video Uploading is the process of uploading all videos from the YouTube site directly into the local server of our website, using a technique called Embed source, which is an external source used to link with the local server. Embed source is a simple html tag, where the attribute 'src' specifies the address of the external file to embed. Video Embedding is a way in which the various attributes or meta-tags or other external links required to refer the video are embedded along with it.

ADDING SIGNATURES

Adding Signatures is an important module for ranking. The various keywords used by the users to search a particular video, are encrypted into a secure format and added as a meta-data to the video. This promotes popularity. Keywords are converted into signatures using a Digital Signature Algorithm (DSA). This algorithm translates the search keywords into an encrypted form, thus provides security in dynamically adding these signatures into the channel's source page. The external link of the video from YouTube which is used through the Embed source is

978-1-5386-2447-0/18/$31.00 ©2018 IEEE

encrypted into a signature and stored in the database for security. Any recreated content would be detected, whereby duplication would be prevented

SEARCHING

Searching is an action performed by the end user, where a particular keyword is selected for search. The background work of searching performs two functions:

- Polarity consistency check
- Ranking

Polarity consistency check is performed on the comments and ratings provided by the users for various videos. Using sentimental analysis, the positive and negative root words are found based on matching the words with a "bag of words" containing all possible commenting words. Ranking is the way in which the video that is searched is optimized to the least index in the website (scaling the position in which the video should be displayed)

PROPOSED RANKING ALGORITHM

Algorithm ***Multi-KeywordRanking*** *L(tag, category, description, Google_hits, keywords)*

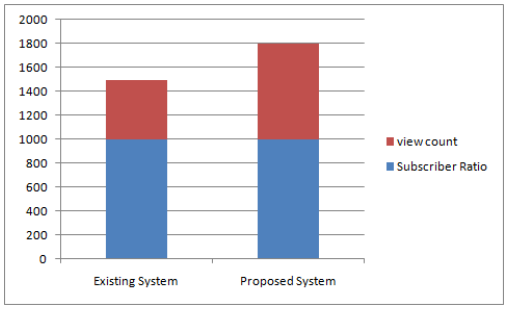
{

Initialize keyword 'k'
F= max(L)
Input C : comments from users viewing the video
C= comments: J(positive, negative)
Initialize C, J
Analyze C : verify(positive, negative)
Result (J)
Actual result:
If F > Result(J)
Search keyword
Else
Search views and ratings (Result (J))

}

The proposed algorithm Multi-Keyword ranking is named so because we consider an important feature of the video which is the keyword used by users to search videos. The various combination of search word that a user attempts to search would be considered. The keyword that the user uses to search a video is initialized as *'k'*. The parameters considered for the algorithm are *tag*, *category*, *description*, G*oogle hits* and *keywords*. The count for each parameter is taken and the maximum among these is stored in the variable *'F'*. The comments from users for a particular video is initialized as *'C'*. The comments are an input to rank the video to its optimized position. The comments are categorized into two ranges, '*positive'* and '*negative'*. The function that classifies the comments into its types is *'J'*. The comments are analyzed using the function '*verify(positive, negative)'*. The result of analysis is stored in '*Result(J)'*. The actual result is the change in position of the video based on the user's search. Comparison is done against the '*max(L)'* and '*Result(J)'*. If the attributes have a maximum value compared to the comments, then the *'search keyword'* would be used to rank the video else the *'comments'* and the *'view count'* would be used to used to rank the video.

IV. RESULT ANALYSIS

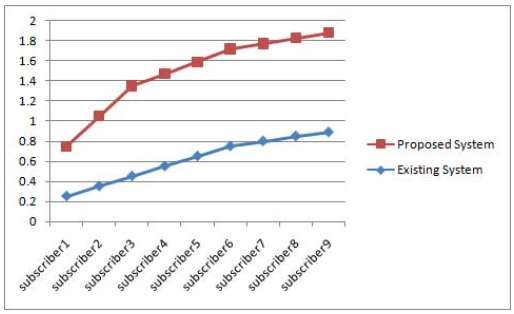


(a) Fig 2. Number of subscribers and view count

TABLE I.

NUMBER OF SUBSCRIBERS & VIEW COUNT

Category	Existing	Proposed
View count	450	800
Subscriber ratio	1000	1000



(b) Fig 3. Impact of view count based on the number of subscribers

TABLE II.

978-1-5386-2447-0/18/$31.00 ©2018 IEEE

IMPACT OF VIEW COUNTS BASED ON NUMBER OF SUBSCRIBERS

Subscriber Ratio	Existing	Proposed
S1	0.2	0.6
S2	0.3	0.65
S3	0.35	0.69
S4	0.4	0.73
S5	0.48	0.78
S6	0.5	0.88
S7	0.6	0.92
S8	0.75	0.97

(a) Search keyword Vs. Google hits

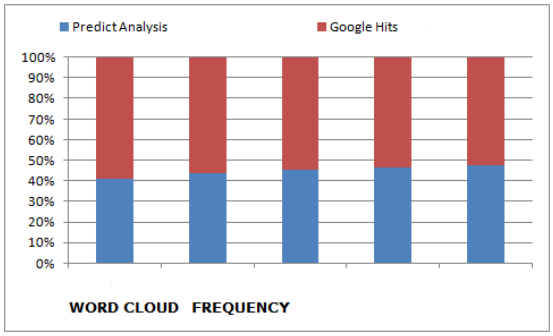


Fig 4. Search keyword vs. Google hits

V. CONCLUSION AND FUTURE WORK

The outcome of this proposed work is an optimized search where the searched video based on the converted signature is searched from the local host and ranked using the comprised attributes among the various internal and external factors. The result of the search provides an optimized position in which the video would be dynamically ranked and the recreated or duplicated content would be prevented because the input take in is from the YouTube site through the Embed Source and would detect the repeating content, whereby duplication would be prevented. Also, the embed link for each video would be encrypted into its relevant signature for security. The extension of this proposed work would be the inclusion of other social Medias data.

VI. REFERENCES

[1] William Hoiles, Anup Aprem and Vikram Krishnamurthy, "Engagement Dynamics and Sensitivity Analysis of YouTube Videos", IEEE Transactions on Knowledge and Data Engineering, IEEE, 2017

[2] Liqiang nie, Richang Hong, Luming Zhang, Yingjie Xia, Dacheng Tao and Nicu Sebe, "Perpetual attributes optimization for multivideo summarization", IEEE Transactions on Cybernetics, IEEE, 2015

[3] Cedric Richier, Rachid Elazouzi, Tania Jimenez, Eitan Altmany and Georges Linares "Predicting Popularity Dynamics of Online Contents Using Data Filtering Methods", Fifty-Fourth Annual Allerton Conference, Allerton House, UIUC, Sept. 27 to 30, 2016

[4] Tomasz Trzcinski and Przemysław Rokita, "Predicting Popularity of Online Videos Using Support Vector Regression", IEEE Transactions on Multimedia, IEEE, 2017

[5] C´Edric Richier, Rachid Elazouzi, Tania Jimenez, Eitan Altmany, and Georges Linares "Forecasting Online Contents' Popularity", Research Gate Publication, 277603649, May 2015

[6] Youmna Borghol , Youmna Borghol, Siddharth Mitra, Sebastien Ardon, Niklas Carlsson, Derek Eager, Anirban Mahanti, "Characterization and Modeling Popularity of User generated Videos"

[7] Ahmed Abdelsadek, "Distributed Index For Matching Multimedia Objects", Simon Fraser University, 2014

[8] Yuan Ding, Yuan Du, Yingkai Hu, Zhengye Liu, Luqin Wang, Keith W. Ross and Anindya Ghose, "Broadcast Yourself: Undestanding Youtube Uploaders", in Proc. or the ACM Sigcomm Conference on Internet Measurement. New York, NY, USA: ACM, 2011, pp. 361–370

[9] Shumeet Baluja, Rohan Seth, D. Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran and Mohamed Aly, "Video Suggestion and Discovery for Youtube: Taking Random Walks Through the View Graph", Google, Inc. Mountain View, CA, USA.

[10] Yipeng Zhou, Liang Chen, Chunfeng Yang and Dah Ming Chiu "Video Popularity Dynamics and Its Implication For Replication", IEEE Transactions on Multimedia, Vol. 17, No. 8, August 2015

[11] https://searchengineland.com/video-optimization-not-underestimate-poweryoutube

[12] https://www.shoutmeloud.com/youtube-seo-tips.html

[13] http://selfmadesuccess.com/rank-youtube-videos-seo

[14] https://www.searchenginepeople.com/blog/16071-youtube-ranking-algorithm.html

978-1-5386-2447-0/18/$31.00 ©2018 IEEE