Utilization of Analytic Hierarchy Process (AHP) with Linear Regression for Determining Category to Maximize Potential Revenue through View Count

Kevin Joseph de Guzman and Rex Aurelius C. Robielos
School of Industrial Engineering and Engineering Management
Mapua University
Intramuros, Manila, Philippines
kevindeguzman129@gmail.com, racrobielos@mapua.edu.ph

Abstract

This study utilizes the Analytic Hierarchy Process (AHP) in the selection of an optimal niche or category of videos for maximizing view count. Main income from videos are derived from RPM, which is a set amount per a thousand views. A set of criteria was determined from attributes of the dataset that logically contributes to either the videos' SEO or in trend/popularity. The criteria in question was also determined by commonalities in a vast amount of video content platforms, which focuses more on the essential attributes of a video. In order to perform pairwise comparison, weights were derived from coefficients generated from using Linear Regression. Upon creation of the model, we identify the categories that have the most potential for garnering views. Based on results, the study may be performed in another time frame to reflect the major shifts in public interest over time. Thus, importance to its repeatability and degree of usability against datasets from different platforms is emphasized.

Keywords

Analytical Hierarchy Process, Regression Analysis, Video Content Creation

1. Introduction

In contemporary society, social media is changing the way people create, share, and consume information (Mangold & Faulds, 2009). These social media platforms are being driven by content created by, and for, the users of the platform. These types of content are denoted as User Generated Content (UGC). Being a content producer on these platforms is becoming to be a more viable way to earn income in the creative space. The content production through social media allow users to fulfill their information, entertainment, and mood management needs, while its generation (or sharing) allows for self-expression and self-actualization (Shao, 2009). The exponential growth of social media in contemporary society makes them necessary tools for communication, content creation, sharing, and business growth (Kaplan & Haenlein, 2010).

With more and more creators present on a platform, competition can be detrimental to success since viewership share is diluted to more participants. One other factor is the introduction of content from traditional media companies, usually in the form of video clips of shows broadcasted on television networks and cable networks. These types of content are called non-user generated content.

For traditional media companies interested in entering the internet media market, one would have to use one of the many video sharing platforms. They have the capacity and capability to produce content on any topic or category. However, not all content posted on video platforms such as YouTube gets the desired attention and only a fraction can reach a large audience, particularly the videos posted by social media marketers expecting millions of views (Khan, Gohar & Vong, Sokha 2014). Companies want to identify these categories with the most potential for high amount of views on the platform, thereby maximizing profit, while considering parameters affecting quality and timeliness of video production and publishing.

In this paper, we measure the characteristics of videos from different categories, in terms of video duration, view counts, and user engagement, and assess their potential for revenue. By understanding the characteristics of videos with high view counts, the study will help traditional media companies to determine the type of content they would want to produce to ensure returns on investments. From initial observations of the data, we found that the videos from different categories have noticeably different characteristics (Weilong Yang, Zhensong Qian, 2011). This study aims to identify categories with the potential for high view counts, using criteria common to most video sharing platforms. The study also aims to formulate a methodology that can be easily reused against different video sharing platforms, as well as across time periods.

2. Methodology

2.1. Dataset

Data is sourced from video sharing website from 2017 (N~200,000). The features we consider for each video are video length (in seconds), number of views, category names, video resolution, the word count of the title and word count of the tags used. Features common to social media platforms are functions to boost social interaction (Benevenuto et al., 2008), such as, the users' ability to post comments on a video, liking/disliking a video, or share a video to other social network platform such as, Facebook or Twitter. For this study, we will incorporate likes and dislikes and/or its ratio.

The features that were identified for this study can be found in most, if not all, social media platforms, especially for video content. This is to enable the methodology to be used easily across datasets from the different platforms, which in of itself have different priorities. For example, some platforms only permit short form videos while platforms directed to gaming content generally have hour-long videos. Another example is with categories, where there are different methods of categorization per platform.

2.2. Linear Regression in Calculating Criterion Weights

In the analytic hierarchy process (AHP), the decision maker makes comparisons between pairs of attributes or alternatives. In real applications the comparisons are subject to judgmental errors. Based on this model we present the formulae for the evaluation of the estimates of the AHP-weights obtained by regression analysis. (Laininen & Hämäläinen, 2003)

For the calculation of weight for each criterion, we utilized Linear Regression to estimate weights from the resulting coefficients. The target variable would be view count as it is the primary driver of assessing rate of income on any video sharing platform. The model is presented in statistical formula as follows:

$$\hat{V} = b_o + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$$

Where:
$\hat{V}$: Views
X_1: Length
X_2: Quality
X_3: Title Length Count
X_4: Tag Word Count
X_5: Like to Dislike Ratio, denoted as Rating

2.3. Linear Regression in Calculating Alternative Weights with regards to each Criterion

Using the same process as with assessing weights for criteria, the researcher performed the same with regards to each category. As with any topics that can be viewed on the platform, not all of them can be produced in such a way that exemplify the attributes identified as primary success factors. In order to perform regression on multiple categorical variables, the categories are coded into its' own matrix of Boolean values. A sample of the model is shown below:

$$\widehat{C_1} = b_o + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + \cdots + b_nX_n$$

Where:
$\widehat{C_1}$: Criterion Coefficient
X_n: Category n (as parsed and identified from the dataset)

This process is repeated for each criterion. With this we can proceed with presenting the AHP model with the weights being provided by the above methods. Then, pairwise comparisons are made and calculated after estimation of weights from applying linear regression against criteria and against the category with regards to the criteria. This was done via Python using Numpy, Scikit-Learn, Pandas, as well as other utility functions.

2.4. Analytic Hierarchy Process (AHP) Model

Figure 1 shows the 4 layers of AHP. The first layer shows the main goal, which is the identification of the category/topic to focus on for maximum potential views. The primary criteria the researched has selected are production quality, search engine optimization, and user engagement. The secondary criteria that the researcher has selected are length, quality, # of words in the video title, # of tags used to describe the video, or by votes, ratings, likes or dislikes.

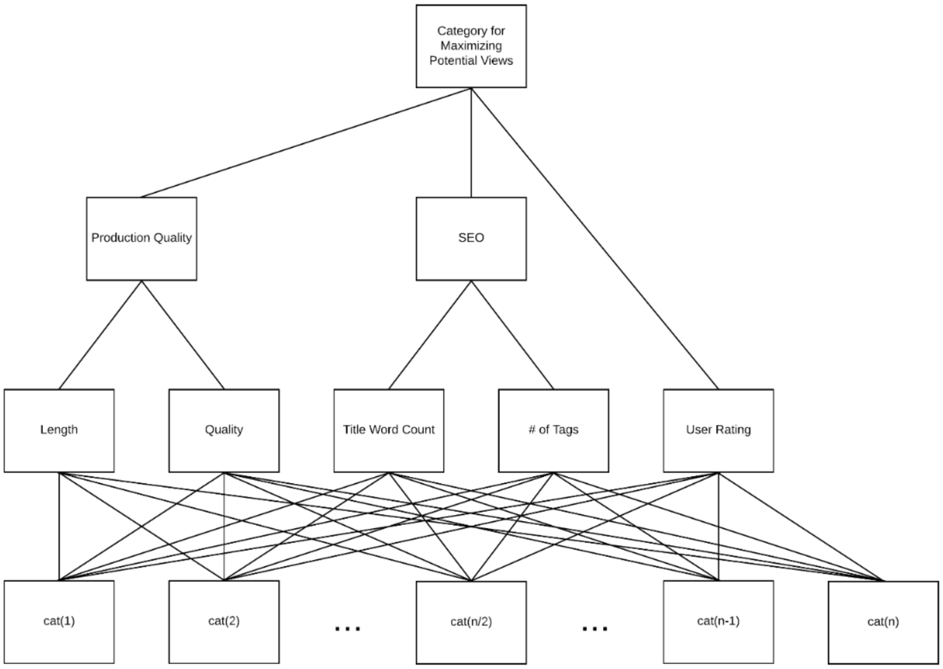


Figure 1. Generalized AHP Model

These criteria are generally the parameters content creators tune during the production process. Length is an important decision to make as it is highly correlated with production time. Content producers must decide on a balance of production time and quality to ensure a consistent upload schedule. Video quality is another primary factor as this pertains to the actual recording quality, as well as production value. Tags improve the videos' visibility in the platform's search functionality. User rating reflects how well the videos are received by viewers, this may depend on the presentation (positive/negative, conservative/controversial) of a specific issue or topic within the general category.

Also, videos with higher ratio of likes per view, higher negative sentiment in comments, and higher view count are more likely to be watched (Park, Minsu, Naaman, Mor, Berger, Jonah, 2014). Higher user engagement in any kind of way is more likely to be shared, in either to continue the discussion or to have their social network have a say in the content of the video.

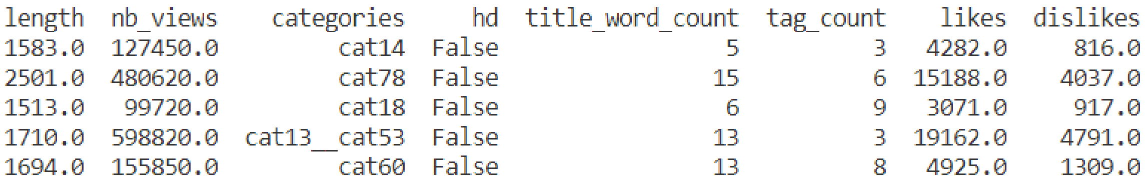
Videos from different categories have different statistics on video duration, popularity, user engagement and so on (Weilong Yang, Zhensong Qian, 2011), and thus each category have different priorities in which to maximize views. Video platforms permits users to share different categories of videos with different groups of people (Lange, 2007). Thus, we believe that different types or categories of video (e.g., music, comedy, drama, and animation) may affect view count differently.

3. Results

3.1. Dataset Exploration and Pre-requisites for Linear Regression

From the dataset, alternatives were generated from the unique categories of the data used, which resulted into 84 categories. Categorical data is converted into binary columns before performing linear regression. Data was processed in Python using standard statistical libraries.

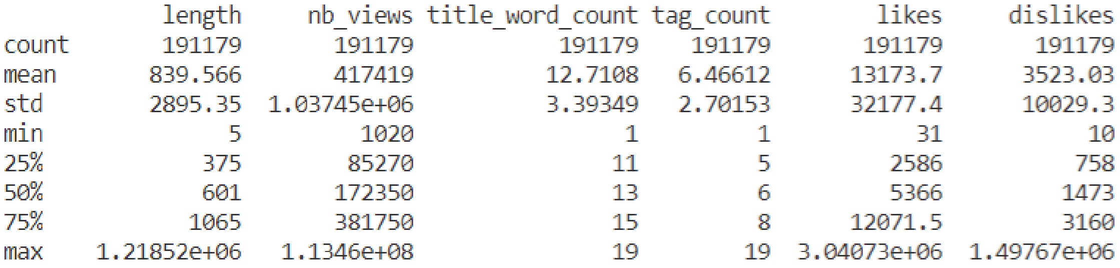
Table 1. Sample of Dataset



```
length  nb_views    categories     hd  title_word_count  tag_count    likes  dislikes
1583.0  127450.0         cat14  False                 5          3   4282.0     816.0
2501.0  480620.0         cat78  False                15          6  15188.0    4037.0
1513.0   99720.0         cat18  False                 6          9   3071.0     917.0
1710.0  598820.0  cat13__cat53  False                13          3  19162.0    4791.0
1694.0  155850.0         cat60  False                13          8   4925.0    1309.0
```

For the features in the dataset, shown from a sample set in Table 1, we have video length in seconds, number of views, number of likes and dislikes, the number of tags, and the word count of the title. We observe that there can be cases where two (or more) categories can be tied to a video. This will be considered upon creation of the model through the means of coding. Aside from this we also have the quality variable denoted as “hd” in the dataset. Further inspection of the likes and dislikes, we can simplify into a singular value as a ratio between the two. This is to remove multi-collinearity between likes and dislikes

Table 2. Summary Statistics of Numerical Data in Dataset



```
           length      nb_views  title_word_count  tag_count         likes      dislikes
count      191179        191179            191179     191179        191179        191179
mean      839.566        417419           12.7108    6.46612       13173.7       3523.03
std       2895.35   1.03745e+06           3.39349    2.70153       32177.4       10029.3
min             5          1020                 1          1            31            10
25%           375         85270                11          5          2586           758
50%           601        172350                13          6          5366          1473
75%          1065        381750                15          8       12071.5          3160
max   1.21852e+06    1.1346e+08                19         19   3.04073e+06   1.49767e+06
```

Looking at the summary statistics of the dataset in Table 2, we see that the average video length is about 15 minutes, with the minimum of 5 seconds. This may be due to having videos from short duration video platforms line Vine or TikTok, re-published to other video sharing websites. Videos longer that 15 minutes tend to be educational in nature, or in the form of radio shows and/or podcasts on a myriad of topics. An appealing study on user generated content illustrated the result of difference video popularity and length between user generated content and non-user generated content (Cha, et al., 2007). A mean view count of around 420k shows us that most of the videos on the dataset are somewhat popular in nature. Title and word count seem to be in close correlation, as tags can be derived from keywords used in video titles. Some such practices are called keyword brainstorming in the SEO space.

3.2. Pre-Requisites for Linear Regression

One of the prerequisites in utilizing linear regression is to verify if the independent variables follow a normal distribution. A normal distribution is a probability distribution of outcomes that is symmetrical or forms a bell curve. In a normal distribution 68% of the results fall within one standard deviation and 95% fall within two standard deviations. In order to visualize this, we need to plot the logarithmic values of the independent variables individually.

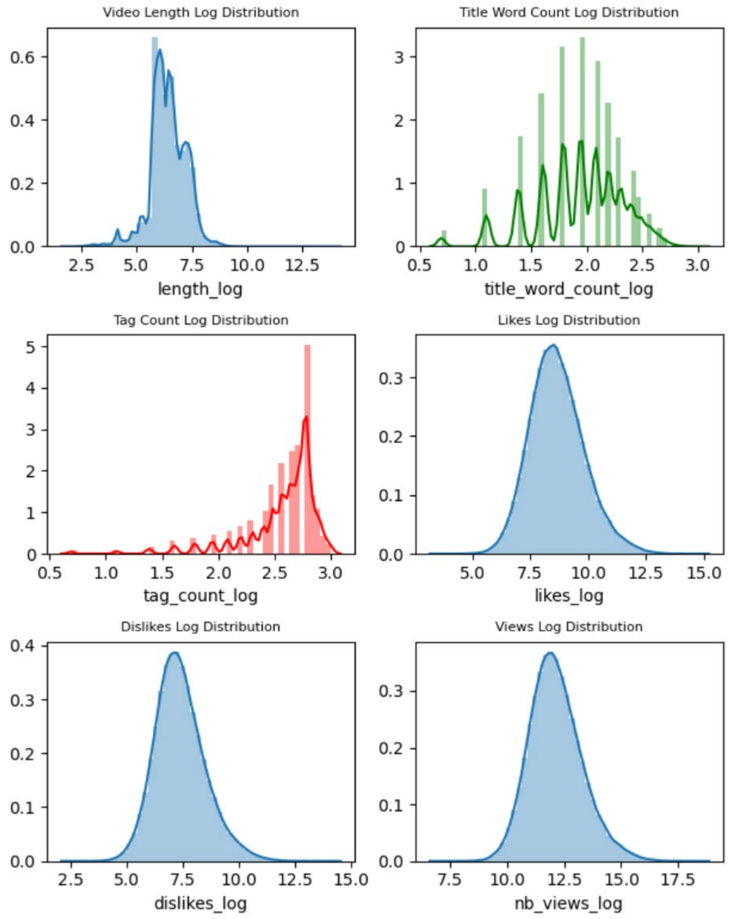


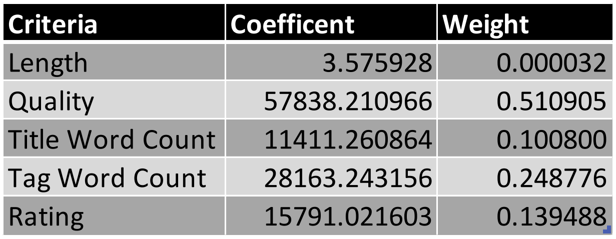
Figure 2. Log Distribution of Primary Independent Variables

Inspecting the log density of the numerical variables in Figure 2, we could see that almost all, apart from tag count, can be considered a normal distribution. This shows that most variables are appropriate to be used in linear regression in order to get weights from the coefficients.

3.3. Linear Regression

From the results, we see that that the R-score for the linear regression model to predict view count from the independent variables (length, quality, title/tag word count, and like to dislike ratio), would be around 0.02, which is quite low. Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. As such, we can still draw important conclusions about how changes in the predictor values are associated with changes in the response value.

Table 3. Identified Criteria with Calculated Weights

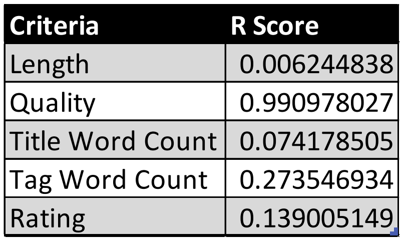


Criteria	Coefficent	Weight
Length	3.575928	0.000032
Quality	57838.210966	0.510905
Title Word Count	11411.260864	0.100800
Tag Word Count	28163.243156	0.248776
Rating	15791.021603	0.139488

For the weighs of alternatives shown in Table 3, we can see that quality has the highest R score from the criteria, followed by tag word count, rating, title word count, with length having the smallest score. With average network throughput increasing year-by-year, it is apparent that more and more users demand content produced in higher fidelity. Improvements in screen resolution for both mobile devices and home systems affect viewing experience negatively when consuming lower quality videos.

Table 4 presents the results after creation of the Linear Regression models to verify how well the metrics affect the primary criteria. It is apparent that video quality holds the most weight in determining popularity of a video. Tag Count also has a significant effect in views as it positively affects the videos' visibility in search results as well as being correctly identified in profiling algorithms. It will be more likely to be suggested to viewers who also watched videos that have similar tags.

Table 4. Table of R Scores for model with regards to each criterion



Criteria	R Score
Length	0.006244838
Quality	0.990978027
Title Word Count	0.074178505
Tag Word Count	0.273546934
Rating	0.139005149

User rating follows tag count in the comparisons of R scores. Videos suggested from the profiling of viewers' interests will also have a higher likelihood of being liked, commented by, and shared by those users. Viewers who are trying to find content outside their usual interests may be attracted to a video with a higher level of user engagement.
Title word count has a slightly less significant score than that of tag count. It may be due to the limitations of how much information to be able to include in such a small space. At most times, titles are used to entice viewers to misleading to the actual content, aptly named "click bait" titles.

Using more tags that describe the video in order to increase visibility in search engines also proves to be a good indicator that it will have more views. Keywords taken from either the video's title, or determined through keyword brainstorming, can be an effective way in capturing more viewer share.

Scores from user engagement, either through a voting or a like/dislike system also proves to be a good metric in determining whether a specific video gets more views. As more and more active viewers engage with the content through provided means, it is more likely to be shared across their own social networks, thus increasing view count even further.

Title word count has a significantly lower weight when compared to tags count, as tags don't have the limitation of character or word length. Titles must be able to convey the essence of the video in a such a short space that some context may be lost. In certain videos, titles and tags are used in complement, having a catchy title not exactly describing what the video entails while having tags correctly identify content for target viewers.

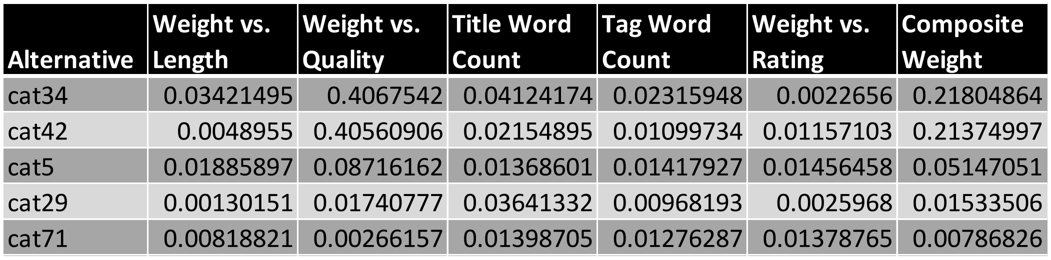
The lowest weight was video length. As seen from the summary statistics, videos come in many lengths and forms, from 5 seconds to a couple of hours. As far as the results show, video length does not entirely matter in terms of potential view counts. Shorter videos can be consumed easily while longer videos tend to have more potential for revenue outside of view counts, such as advertisements that are part of the production, or advertisements inserted by the platform in the middle of videos.

With these primary weights, we can proceed in calculating the weights for the alternatives for each of the criteria.

3.4. AHP Weight Calculation

The data in Table 5 shows the calculated weights for the top 5 categories, with weights against each criterion, with the final column as the final composite weight. As the previous section observed the weights regarding views, this section will discuss observations made to the distribution of the weights for each category.

Table 5. Top Categories from the Utilized Methodology



Alternative	Weight vs. Length	Weight vs. Quality	Title Word Count	Tag Word Count	Weight vs. Rating	Composite Weight
cat34	0.03421495	0.4067542	0.04124174	0.02315948	0.0022656	0.21804864
cat42	0.0048955	0.40560906	0.02154895	0.01099734	0.01157103	0.21374997
cat5	0.01885897	0.08716162	0.01368601	0.01417927	0.01456458	0.05147051
cat29	0.00130151	0.01740777	0.03641332	0.00968193	0.0025968	0.01533506
cat71	0.00818821	0.00266157	0.01398705	0.01276287	0.01378765	0.00786826

We can see that the top 2 categories focused on video/production quality. However, the 2nd top category's composite weight (cat42) holds close to the top category even though it has significantly lower weight on the length criteria. This may prove to show that video length does not entirely matter if the video production quality is high.

The 3rd category (cat5) might describe average-length videos but with a lower video quality. Its composite weight is being raised by higher quality or number of tags, as well as its user rating, which has the highest weight of the top 3 categories.

4. Conclusion

The utilization of Linear Regression for estimating weights in AHP proved to be a good approach for processing data generated from high-traffic, social media platforms. It can remove biases that can come from small sample sizes such as surveys directed to a handful of executives. The more data-driven research becomes, it is inherently more reliable, and can be easily implemented in other industries or subject matter.

The independent variables in the study were chosen with the importance of being able to apply the methodology across different video platforms, as well as be easy to repeat in periodic time frames. This is to easily capture shifts in trends, changes in categorization, as well as changes in how videos are measured.

From this study, we have found out, for the dataset used, video length does not matter as much to view count. Also, we have found out that while some attributes have weights consistent across most categories, like search engine optimization related attributes (title, tags), some categories value other attributes more than that of other categories. This can be useful to media companies, or other individuals who peruse the methods in this study, to selectively control how a production should be made for the categories they have selected.

It can also be noted that the study did not select a singular category. Instead, the study presents the top categories by its composite weight. This is to further express that different categories have characteristics that may be better or worse than that of its peers, composite weights being relatively equal. This would mean that the individuals who may use this model can have more control on what the final decision of selecting a category according to their priority over certain video attributes.

References

Benevenuto, Fabrício & Duarte, Fernando & Rodrigues, Tiago & Almeida, Virgilio & Almeida, Jussara & Ross, Keith. (2008). Understanding video interactions in youtube. MM'08 - Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops. 761-764. 10.1145/1459359.1459480.

Kaplan, Andreas & Haenlein, Michael. (2010). Users of the World, Unite! The Challenges and Opportunities of Social Media. Business Horizons. 53. 59-68. 10.1016/j.bushor.2009.09.003.

Khan, Gohar & Vong, Sokha. (2014). Virality over YouTube: an empirical analysis. Internet Research. 24. 629-647. 10.1108/IntR-05-2013-0085.

Laininen, Pertti & Hämäläinen, Raimo. (2003). Analyzing AHP-matrices by Regression. European Journal of Operational Research. 148. 514-524. 10.1016/S0377-2217(02)00430-7.

Lange, P.G. (2007), Publicly Private and Privately Public: Social Networking on YouTube. Journal of Computer-Mediated Communication, 13: 361-380. doi:10.1111/j.1083-6101.2007.00400.x

Mangold, W. & Faulds, David. (2009). Social media: The new hybrid element of the promotion mix. Business Horizons. 52. 357-365. 10.1016/j.bushor.2009.03.002.

Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. 2007. I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC ’07). Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/1298306.1298309

Park, Minsu, Naaman, Mor, AND Berger, Jonah. "A Data-Driven Study of View Duration on YouTube" International AAAI Conference on Web and Social Media (2016): n. pag. Web. 24 Jul. 2020

Shao, G. (2009), "Understanding the appeal of user-generated media: a uses and gratification perspective", Internet Research, Vol. 19 No. 1, pp. 7-25.

Yang, Weilong, and Zhensong Qian. "Understanding the Characteristics of Category-Specific YouTube Videos." Entstanden im Rahmen eines Informatikseminars an der kanadischen Simon Fraser University (2011).

Biographies

Rex Aurelius C. Robielos is the Dean of the School of Industrial Engineering and Engineering Management at Mapua University. Before joining Mapua, he was Section Manager of Operations Research Group, Analog Devices General Trias. He has a BS in Applied Mathematics from the University of the Philippines Los Baños, and a Diploma and MS in Industrial Engineering from the University of the Philippines Diliman. He is pursuing Ph.D in Industrial Management (candidate) at National Taiwan University of Science and Technology in Taiwan. He is the current Secretary of Human Factors and Ergonomics Society of the Philippines and Director of the Philippine Institute of Industrial Engineers and Operations Research Society of the Philippines.

Kevin Joseph S. De Guzman is a Cloud and Automation Software Engineer at DXC Technology. He has a BS in Information of Technology from the Mapua Institute of Technology (now known as Mapua University). He is currently pursuing master’s in Business Analytics at Mapua University.