Research Article

A MOOC Video Viewing Behavior Analysis Algorithm

Yong Luo,1,2 Guochang Zhou,1 Jianping Li,1 and Xiao Xiao3

1College of Science, National University of Defense Technology, Changsha, Hunan, China
2Department of Information Technology, Hunan Police Academy, Changsha, Hunan, China
3Chinese University MOOC, Beijing, China

Correspondence should be addressed to Yong Luo; yngluo@163.com

Received 4 June 2018; Accepted 24 September 2018; Published 16 October 2018

Academic Editor: Alessandro De Luca

Copyright © 2018 Yong Luo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

MOOCs (massive open online courses) are developing rapidly, but they also face many problems. As the MOOC’s most important resource, the course videos have a very important influence on the learning. This article defines the ratio \( R = \frac{Average \text{ viewing duration}}{Video \text{ length}} \), which reflects the popularity of the video. By analyzing the relationship between the video length, release time, and \( R \), we found a significant negative linear correlation between video length and \( R \) and video release time and \( R \). However, when the number of videos is less than the threshold, the release time has less influence on \( R \). This paper presents a video viewing behavior analysis algorithm based on multiple linear regression. The residual independence test proved that the algorithm has a good approximation to the data. It can predict the popularity of similar course videos to help producers optimize video design.

1. Introduction

MOOCs [1] are the massive open online courses, and these courses are usually provided by university and shared in the network [2, 3]. Coursera, Udacity, and edX [4] are currently the three largest providers in the world. The Chinese University MOOC [5] is the largest operator in China, which is launched by the Higher Education Agency of China.

MOOCs [2–4] tend to have a large number of learners. These learners are brought a new experience and given a chance to get involved in high education. At the same time, the feedback from the learners can also help the university improve courses.

MOOCs are developing rapidly, but they still have to face many problems, such as high dropout rates, low resource utilization, and lack of effective profit model [6]. Sometimes, the number of candidates can reach tens of thousands, but only 1% of the students completed the course. The Advanced Mathematics in the Chinese University MOOC is taken as an example. Its number of students is 4317 in 2017, while the number of learners who had watched more than 50% of videos is only 18. The low utilization of course videos is almost beyond imagination. How to make MOOC video more attractive is a very important research topic.

Philip J. Guo [7] studied the influence of teachers’ images, video length, and speech speed on learners. Their research shows the video within 6 minutes is the best, which comes from data analysis and questionnaires. Many researchers explore the rules of online learning by analyzing learning behavior data. Among them, viewing video behavior data is a research hotspot. And some scholars have done a lot of valuable work. Tanmay Sinha [8] operationalizes video lecture clickstreams of students into cognitively plausible higher-level behaviors. Their results illustrate how such a metric inspired by cognitive psychology can help answer critical questions regarding students engagement, their future click interactions, and participation trajectories that lead to in-video and course dropouts. Nan Li et al. [9] present a video interaction analysis to provide empirical evidence about this issue. They find out that speed decreases, frequent and long pauses, and infrequent search with high amount of skipping and rewatching indicate higher level of video difficulty. Geza Kovacs [10] analyzes how users interact with in-video quizzes and how in-video quizzes influence users lecture viewing behavior.
behavior. Through data analysis, they found the peak period for students to think about issues. Juho Kim et al. [11] identify five student activity patterns that can explain peaks: starting from the beginning of a new material, returning to missed content, following a tutorial step, replaying a brief segment, and repeating a nonvisual explanation. Christopher G. Brinton [12] studied student video-watching behavior and quiz performance. It is found that some of these behaviors are significantly correlated with changes in the likelihood that a student will be Correct on First Attempt or not in answering quiz questions, and in ways that are not necessarily intuitive. Qing Chen et al. [13] introduce a comprehensive visualization system called Peak Vizor. This system enables course instructors and education experts to analyze the peaks or the video segments that generate numerous clickstreams.

These studies provided many analysis methods for video viewing behavior and proposed some options to optimize video design. However, they are preliminary statistics and processing of the data, and there is no in-depth analysis of the functional relationship. The purpose of our research is to build a mathematical model, which can describe the impact of video length and release time on attraction.

The parameter $R$ is defined in the article, which is defined as ratio of average viewing duration to video length. Its value reflects the popularity of the video. Obviously, the larger the value of $R$, the higher the popularity of the video. Our research shows that there are significant negative linear correlations between $R$ and video length, and between $R$ and release time. Oversized videos tend to have small $R$ values, because learners cannot keep an interest in video for a long time. At the same time, the learner's interest later is lower than the initial stage. However, there is a threshold for this. When the number of videos is less than a certain threshold, the learner's interest in all course videos does not fluctuate significantly.

The multiple regression is used to analyze viewing video data in this article. By comparing the standardized regression coefficients, the release time has a much greater effect on $R$ than the video length. If the length of a single video is reduced, the total number of videos will be increased. So, it is unwise to emphasize video segmentation to shorten the length of video. We need to choose a balance between the number of videos and the length of the video. The algorithm proposed in this paper can predict the attractiveness of course video and help providers optimize video segmentation.

### 2. Data Description and Preprocessing

MOOC learning behavior data [14, 15] mining has become an important research field. Our data are provided by the Chinese University MOOC. The total time each learner watches each video is recorded. Only the impact of video length and number on attraction is studied here, and course content and teacher performance are ignored. Advanced Mathematics course is chosen as a research object. The time interval for data collection is 2016.09.26 to 2016.11.26. Although the content of each video is also different, and the teacher's performance will fluctuate, these differences are not considered in this paper.

#### Table 1: Course information.

<table>
<thead>
<tr>
<th>Course</th>
<th>Learners</th>
<th>Videos</th>
<th>Recording</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Math</td>
<td>27664</td>
<td>129</td>
<td>41033</td>
</tr>
</tbody>
</table>

![Figure 1: The data contains invalid learning behavior.](image)

Table 1 lists the basic information of the course. The visiting site for the course is https://www.icourse163.org/course/NUDT-9004. It contains 129 videos and the number of students enrolled in this class reached 27664. Many of these classmates did not watch videos, so the data contains 41033 viewing video records.

The sampling period for the data is 10 seconds. Thus, the viewing duration is an integer multiple of 10, and the unit is seconds. The data collection method is to record the total duration of watching a video. In this way, each learner has a record for a video. Then everyone's record is an $n$ dimensional vector, where $n = 129$. Let the number of learners be $m$, where $m = 4805$. Therefore, the viewing duration data is a matrix $D = (d_{ij})_{m \times n}$.

Since too short viewing time is not enough for learning, the data must be preprocessed to remove invalid data. A histogram is drawn to analyze those invalid behaviors, which contained 41033 records.

As shown in Figure 1, there are many learners watching video with a total time of less than 60 seconds (red bar). These people may have just browsed the video or accidentally opened it. For noneffective learning behaviors, these data should be filtered out.

#### 3. Ratio of Viewing Duration to Video Length

To analyze the relationship between the viewing duration and the length of the video, we calculate the average viewing duration of each video. This average viewing duration is calculated after removing the invalid data. Let $w_j$ be the average viewing time of $j$-th video. Its definition is

$$w_j = \text{Average} \left\{ d_{ij} \mid d_{ij} \geq 60, \ i = 1,2,\ldots,m \right\},$$

(1)

$$j = 1, 2, 3, \ldots, n,$$

where $m = 4805$ and $n = 129$.

The average viewing duration of 129 videos and the length of the video are plotted in Figure 2. The blue curve in Figure 2 represents the learner's average viewing duration, and the red curve represents the video length. Observing Figure 1, almost
all the average viewing time is less than the video length. But the average viewing time is sometimes greater than the length of the video, which is caused by the learner repeatedly viewing the same video. Because the length of the video will affect the average viewing duration, it is unreasonable to directly compare the average viewing times to determine the popularity of the video. Obviously, the ratio of the average viewing duration to the length of the video can reflect the students’ preference for this video. Let the length of the \( j \)-th video be \( V_j \) and the average viewing duration of the \( j \)-th video be \( w_j \). The ratio of the \( j \)-th average viewing duration to the video length is

\[
R_j = \frac{w_j}{V_j}, \quad j = 1, 2, \ldots, n. \tag{2}
\]

Two rules can be found based on Figure 2. One is that the longer the length of the video, the smaller the ratio \( R_j \). Another one is that the ratio \( R_j \) of the later period of the course is smaller than that of the early stage. Because the number of the video represents the order of release, the stage of the lesson can be represented by it. So, we can conclude that \( R_j \) is negatively related to the video length and video number. However, these judgments are very simple and require more accurate quantitative analysis.

### 4. Correlation Analysis

The Pearson correlation coefficient has been widely used in many research fields [16], which is a statistical indicator used to reflect the strength of linear relationship between variables. In our study, the correlation coefficient will be used to analyze the relationship between video length, video number, and \( R \). This work also verified the rationality of the regression analysis.

Let

\[
\rho_{XY} \triangleq \frac{\text{cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} \tag{3}
\]

be the correlation coefficient of \( X, Y \), where \( \text{cov}(X, Y) \) is the covariance of \( X, Y \) and is defined as \( \text{cov}(X, Y) \triangleq E[(X - E(X))(Y - E(Y))] \) and \( D(X) \) and \( D(Y) \) are the variance of \( X \) and \( Y \). The coefficient of correlation describes the strength of linear relationship between \( X, Y \). For the Pearson correlation coefficient, the following conclusions are made:

\[
\min_{a,b} E \left[ (Y - (a+bX))^2 \right] = D(Y) \left( 1 - \rho_{XY}^2 \right). \tag{4}
\]

When \( |\rho_{XY}| \) is large, the mean squared error is small. So \( X \) and \( Y \) show a strong linear relationship.

In order to analyze the relationship between \( R \) and the video length and the video number, the scatter plots between them are drawn.

Figure 3 is a plot of \( R \) and video length. The abscissa is \( R \), and the ordinate is the video length. There is a clear negative linear correlation between them. The Pearson correlation coefficient is equal to \(-0.758\). There was a significant negative correlation at the level of significance of 0.01 (\( p \)-value < \( 10^{-6} \)).

Because the video number is the corresponding release order, it can be used to reflect the progress of the course. The relationship between the video number and \( R \) can be observed in Figure 4. The abscissa of Figure 4 is \( R \), and the ordinate is the video number. The values of \( R \) at the end of the course are smaller than the beginning. The Pearson correlation coefficient is equal to \(-0.646\). It shows that there is a significant negative correlation between \( R \) and video number (\( p \)-value < 0.05).
The dependent variable $y$ and the independent variable can be expressed as a three-dimensional vector. There are $n$ sets of data, and the $i$th set of data is recorded as $(y_i, x_{i1}, x_{i2})$. When there is a linear relationship between $y$ and $x_1$, $y$ and $x_2$, respectively, a multiple linear regression model can be used to describe the relationship between them.

$$y_i = \tilde{y}_i + e_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$$  \hspace{1cm} (5)

where $b_0$ is a constant coefficient and $b_i$ ($i = 1, 2$) is a partial regression coefficient. $e_i$ represents the residual. The residual is the difference between the actual measured data and the estimated value of the dependent variable, which is not determined by the independent variable. The following expression calculates the sum of squares between the predicted and measured values.

$$Q = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2})]^2$$  \hspace{1cm} (6)

By solving $b_0$, $b_1$, and $b_2$, let $Q$ takes the minimum value.

5.2. Abnormal Data Filtering. By observing Figures 3 and 4, it can be found that there are several abnormal data. These abnormal data may have an influence on the calculation of the regression coefficient. Therefore, we should design an algorithm to filter them out. According to the Durbin-Watson test theory [18], those points will be filtered out that fall outside the interval $(\tilde{u} - 3\tilde{\sigma}, \tilde{u} + 3\tilde{\sigma})$. $\tilde{u}$ is the mean of the residuals and $\tilde{\sigma}$ is the variance of the residuals. The process can be divided into the following 5 steps.

(a) Calculate the regression coefficients $\tilde{b}_0, \tilde{b}_1, \tilde{b}_2$.
(b) Calculate the residuals $\{\tilde{e}_i\}$, $i = 1, 2, \ldots, n$.
(c) Calculate the mean $\tilde{u}$ and standard deviation $\tilde{\sigma}$ of the residuals.
(d) Find abnormal data; it meets $\tilde{e}_{i_q} \not\in (\tilde{u} - 3\tilde{\sigma}, \tilde{u} + 3\tilde{\sigma})$, $q = 1, 2, \ldots, Q$.
(e) Filter out abnormal data, including $R$ set $\{R_{i_k}\}$, video length set $\{V_{i_k}\}$, and video number set $\{N_{i_k}\}$, where $q = 1, 2, \ldots, Q$.

Let the data collection after filtering be $\{R_{i_k}\}, \{V_{i_k}\}, \{N_{i_k}\}$, where $k = 1, 2, \ldots, K$. Obviously, $n = Q + K$ here.

5.3. Regression Coefficient and Residual Test. After the previous 5 steps, the abnormal data are filtered out. Multiple linear regression is performed using data sets $\{R_{i_k}\}, \{V_{i_k}\}$, and $\{N_{i_k}\}$. Let the regression function be

$$R_{i_k} = \tilde{R}_{i_k} + e_{i_k} = b_0 + b_1 V_{i_k} + b_2 N_{i_k} + e_{i_k}$$  \hspace{1cm} (7)

It is not enough to just calculate the regression equation. It is necessary to verify the correctness of the model. The verification method is to test the residual $|e_{i_k}|, k = 1, 2, \ldots, K$ for independence. When the residuals obey the normal distribution $N(0, \sigma^2)$, it shows that the model has a higher degree of approximation. Kolmogorov-Smirnov [19] test can be used here.
6. Experiment

6.1. Abnormal Data Filtering. Firstly, a multiple linear regression is performed on the preprocessed data; then the residuals are calculated. However, the video number and the video length are two variables with different dimensions. Their influence on $R$ cannot be directly compared. The standardized regression coefficients can be used to analyze the contribution of the independent variable to $y$. The standardized formulae for the variables are $ar{x}_{ij} = (x_{ij} - \bar{x}_j)/\sigma_{x_j}$, $\bar{y}_i = (y_i - \bar{y})/\sigma_y$ $(i = 1, 2, \ldots, n)$. The standardized regression equation is $ar{y}_i = \bar{b}_0 \bar{x}_{i1} + \bar{b}_1 \bar{x}_{i2} + \bar{e}_i$. Note that the coefficient $b_0$ here is 0. The results are shown in Table 2.

By comparing the two standardized regression parameters in Table 2, we can get a clear conclusion. Compared to the length of the video, the video number has a greater influence on $R$. The coefficients in Table 2 cannot be used for data fitting. Its purpose is to calculate residuals and find abnormal points.

Table 3 lists the statistic residuals. The mean of the residuals is less than $10^{-4}$ and the variance is 0.1053. According to the proposed algorithm, abnormal data are removed. In this course, there is only one data set that needs to be removed. After filtering, there are 128 data sets, which are marked $\{R_k\}$, $\{N_k\}$, and $\{V_k\}$. They will be used to do multiple regression analysis to get new coefficients.

6.2. Calculate the Regression Coefficient. The nonstandardized regression coefficients can be used to predict $R$, while the standard regression coefficients can be used to compare the influence of video length and number on $R$. Here, they are all calculated.

Observing the data in Table 4, it was found that the standardized coefficients of the video number and video length variables are -0.622 and -0.155, respectively. For this course, the video number had almost 4 times more influence on $R$ than the video length. The regression equation can be used to calculate $\hat{R}$.

The blue line in Figure 7 is $R$ calculated from the data, and the red line is the curve predicted by the regression equation. As can be seen from Figure 8, the predicted value reflects the change trend of $R$. The regression equation can be used for video design of similar courses. After providing the video length and number, the values of $R$ can be predicted. The predicted values $\hat{R}$ can serve as important reference. The video recorder can optimize the design by video segmentation or integration to increase the attractiveness of the video. Because the appeal of the video is influenced by the course content, it is important to emphasize that the parameters used for prediction should come from similar courses.

6.3. Residual Test. The residual independence test is a very important step in the algorithm. When the residual satisfies the normal distribution [20] of $N(0, \sigma)$, the accuracy of the mathematical model can be guaranteed.

Formula $e_i = R_i - (b_0 + b_1 V_{i1} + b_2 N_{i2})$ is used to calculate residual $\{e_i\}$. The residual set $\{e_i\}$, $k = 1, 2, \ldots, K$ contains 128 data sets ($K = 128$). The method of residual independence test is to test normal distribution of $\{e_i\}$, $k = 1, 2, \ldots, K$. 

Table 2: Multiple linear regression coefficients for unfiltered anomaly data.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Non-standardized</th>
<th>Standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>1.107</td>
<td>0</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-0.000207</td>
<td>-0.654</td>
</tr>
<tr>
<td>$b_2$</td>
<td>-0.001430</td>
<td>-0.299</td>
</tr>
</tbody>
</table>

Table 3: Residual statistics.

<table>
<thead>
<tr>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>-0.2420</td>
<td>0.4321</td>
<td>$&lt; 10^{-2}$</td>
<td>0.1053</td>
</tr>
</tbody>
</table>

Table 4: Multiple linear regression coefficients for filtered data.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Non-standardized</th>
<th>Standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{b}_0$</td>
<td>1.085</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{b}_1$</td>
<td>-0.002841</td>
<td>-0.622</td>
</tr>
<tr>
<td>$\hat{b}_2$</td>
<td>$-4.712 \times 10^{-5}$</td>
<td>-0.155</td>
</tr>
</tbody>
</table>
Mathematical Problems in Engineering

6 Mathematical Problems in Engineering

0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

0 5 10 15 20 25 30 35 40

Video number
R

First Aid Knowledge

0 2 4 6 8 10 12 14 16 18 20

Video number
R

Game Theory

Figure 9: The relationship between R and video number for courses with less than 40 videos.

Figure 9 is a histogram of the residuals, which is very close to the normal distribution graphically. The Kolmogorov-Smirnov test [19] is used to examine whether the residuals are normally distributed. The result shows that the residuals follow a normal distribution at a significant level of 0.05. The residuals’ mean value is < 10^-7, and the standard deviation is equal to 0.12374. This means that the mathematical model has a high degree of approximation to the data.

7. Conclusion

This article analyzes the relationship between video length and video number and R. Our research shows that there is a significant negative linear correlation between video length and R, and between video number and ratio R. The video length and video number are independent. Based on these results, the multiple linear regression equation is used to establish the functional relationship between them. The Durbin-Watson [18] test theory and 3σ principles helped us effectively remove abnormal data. Because the effect of abnormal data on small samples is significant, it is necessary to remove abnormal data. The standardized regression coefficients in Tables 3 and 4 all indicate that the video number has a greater influence on R.

In order to reduce the length of the video, the content of the video needs to be decomposed. However, decomposition will bring about an increase in the number of videos. Therefore, reducing the video length and reducing the number of videos cannot be achieved at the same time. We need to find a balance between these two issues.

We also found an interesting rule. When the number of course videos is small, video appeal is hardly affected by the release time. For courses with less than 40 total videos, the R value is hardly affected by the video number. This means that learners have a threshold for the “patience” of the video number.

Two courses of Chinese University MOOC were chosen to illustrate this phenomenon, which are the Game Theory and the First Aid Knowledge. There are 38 videos in the Game Theory. The First Aid Knowledge contains 19 videos. It can be seen from Figure 9 that R is hardly affected by the release time. There is a threshold M for learners’ patience with video number. Therefore, the number of videos for a course should be controlled within M. However, we do not think that the threshold is 40; it requires more in-depth study. When the content of the course is really large, you can consider splitting it into subcurriculums release. These subcurriculums independently assess and issue certificates, which can guarantee learners’ interest.

So the following suggestions are given for video design.

(i) Do not publish videos that are too long. The teaching content must be streamlined to reduce the length of the video. Philip J. Guo [7] thinks the video within 6 minutes is the best.

(ii) Try to control the number of videos within M. When a certain limit is exceeded, the release time of the video will seriously affect R.

(iii) Do not publish long video later in the course. The value of R is often very low.

(iv) After the video is recorded, it must be clipped to balance the number and length of the video. When there are too many videos, you can consider splitting them into several courses and publishing them separately.

(v) The method proposed in this article can predict R. MOOC operators can use this method to establish multiple linear regression models for various courses. It will help video producers optimize the video design and make the videos more attractive. In order to ensure accuracy, please use the regression equation of similar courses to predict.

In future work, we should research the best number of videos for a course. The courses studied in this article are all taught in Chinese. Whether language has an influence on the regression coefficient is another issue that we are concerned about. We expect that researchers with such data sets will be able to give answers.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.
Acknowledgments

This study is supported by the Open Research Fund of National Ministry of Education Higher Education Research Center Grant NO. 2017008 and supported by the Open Research Fund of Hunan Provincial Key Laboratory of Network Investigational Technology Grant NO. 2017WLZC003.

Supplementary Materials

The supplementary materials are provided by the Chinese University MOOC, which are viewing video behavior data of three courses. Advanced Mathematics contains 129 videos. The other two courses are Game Theory which contains 38 videos and First Aid Knowledge which contains 19 videos. (Supplementary Materials)

References
