Predicting Video Engagement of Users

林昌毅　曹書恒　蕭文逸　劉淑雯

Abstract——The goal of this project was to analyze the reasons for users to not complete viewing their chosen video. We noted that the score trends of common video subsets between students are similar. Hence, we applied various data mining tools to achieve the following: user-user analysis model and user-video analysis model. From these two aspects, we want to achieve the result of predicting whether students will finish watching all sections of a video and complete the listening tests.

1. INTRODUCTION

Hopenglish serves as a platform for users to learn English in an engaging and interesting manner through video watching and listening tests to assess their understanding. User retention is vital to the operation of the website as it is an indication of the effectiveness and usefulness of the recommended videos for the users. Having a suitable match between the difficulty level of a video and user's level will engage the users better, improve their learning outcomes and sustain their interest in using the site.

Therefore, the ability to predict video engagement of users could help the company have a better understanding of the suitability of videos to their English learners and generate strategies to enhance the quality of the website's service to the users. Using the results from our project can also help the company with recommending appropriate videos that suits the different English-standards of the students, thus enabling them to gain satisfactory learnings.

2. DATASET AND FEATURES

The company provided two datasets about the user information with corresponding learning records, and video information related to the difficulty levels:

A. *Student Behavior Information*

This dataset contains information on 3098 students such as their chosen videos, the corresponding scores and list of saved words from the video. A score of -1 indicates that the user did not complete the test for a particular section.

B. *Video Information*

This dataset is in the form of (postID, wordLevel,VideoSpeed,subtitleLengthRatio,sectionLength,wordList). For example:

- (15055, 9.09, 10.13, 0.04, 2, [hi, kate, it, going, good, ...])
- (14807, 32.4, 13.77, 0.35, 3, [hello, dr, jeff, machat, ...])

The following are a few insights drawn from the datasets:

- Out of 51489 total unique views, approximately 5% of them are incomplete.
- Majority of the incomplete views occurred at the last section of the video.
- Several videos (each with at least 100 views) have significantly higher incompletion rate than others.

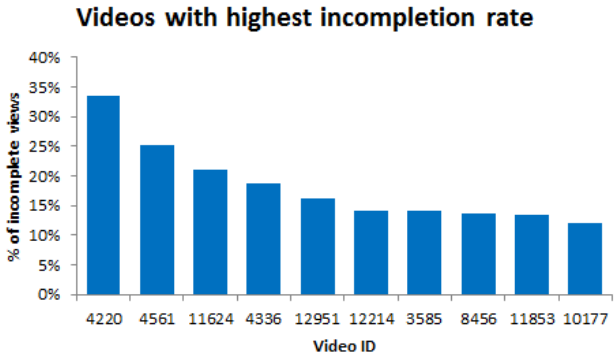


Fig.1 Top 10 videos with highest incompletion rate

3. PREPROCESSING

Prior to generating the prediction models, we conducted several preprocessing steps to clean the dataset and create new variables used during modelling later on:

A. *Filter out data where users have a score of -1 or 0 for all sections of the video*

This was performed as we noted that majority of the users have a score of -1 or 0 for all sections of the last video viewed at the time when data was obtained from the system. Including them would lead to inaccurate and unreliable results.

B. *Video Difficulty Level Classification*

The features from *Video Information* dataset are positively correlated with the difficulty level of a video. We scaled and translated each variable individually such that it falls in the range between 0 and 1. The values of all features of each video were then summed up and used as the difficulty score.

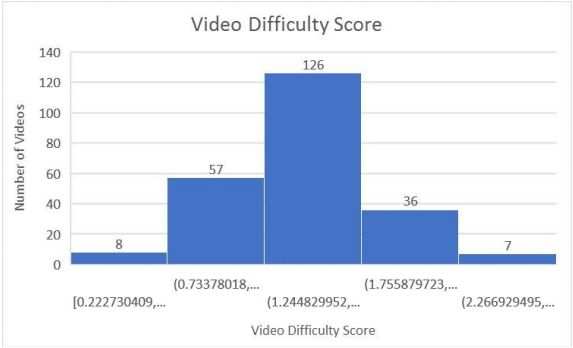


Fig.2 Video difficulty scores

Following which, we divided the videos into 5 levels according to their difficulty score, with Level 5 indicating that it is the most difficult and Level 1 being least difficult.

4. METHODOLOGY

In order to understand why students did not finish viewing a video, we approached this problem by investigating the relation between users who share common video subsets. An incomplete viewing of the video is defined to be when the user has a score of -1 for at least one section of the video.

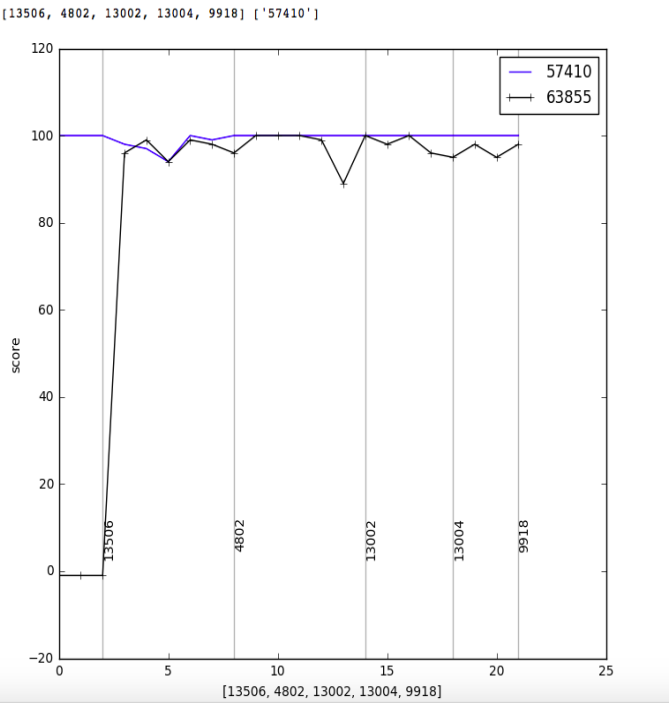
A. *Common Video Subset*

First, we created lists to append the student's listening score to the corresponding video section. Next, we used the list of videos watched by the students to find the common video subsets between every two students. Figure 3 shows the common video subsets for two users No.63855 and No.57410 and figure 4 displays their scores in closer detail.

B. *Measuring distance between users*

After finding the common video subsets between every two students, we created a vector for each student in the pair, consisting of the difference between the video score and average score of all the videos. For example, student A's scores are [80, 90, 80, 70], with an average score of 80. The resulting vector is [0, 10 , 0, -10]. The two vectors were then used to calculate the Manhattan distance between the pair of students.

Fig.3 Member No.63855 and No.57410 has same video sequence:[13506, 4802, 13002, 13004, 9918]



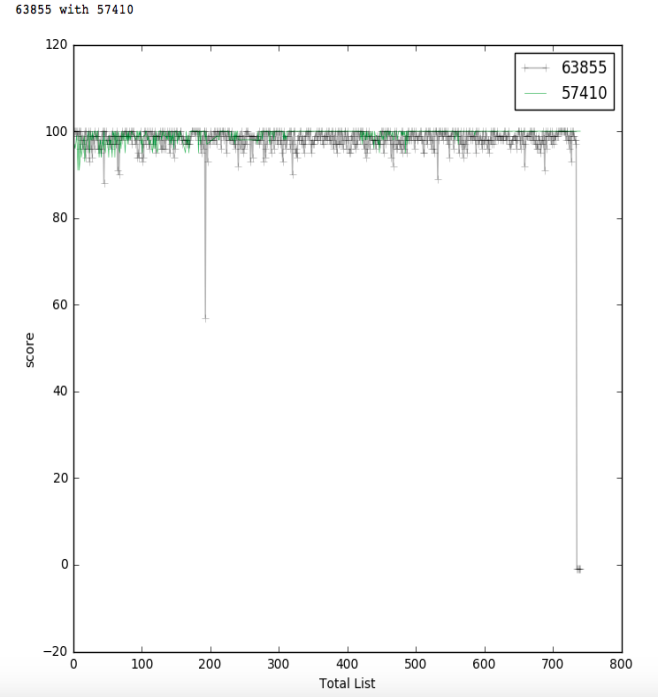


Fig.4 Member No.63855 and No.57410: common section numbers and scores

C. *Clustering*

To obtain clusters of students according to similar listening score trends, we use the distances compute earlier to construct the proximity matrix between every pair of student. Then, the elbow function finds the thresholds for clustering is at 40. Finally, complete linkage hierarchical clustering is applied to find clusters of students with similar listening scores trends.

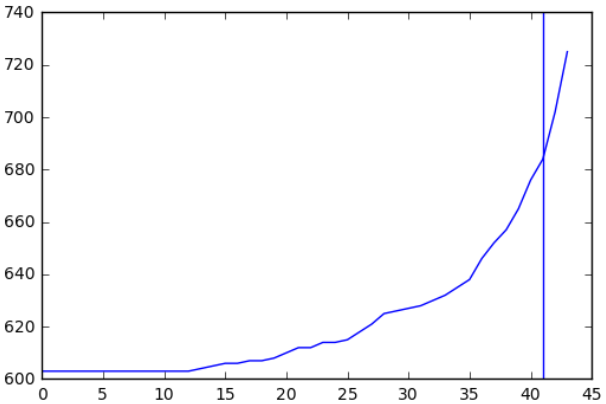


Fig.5 Compute the 'elbow' , where y-axis is clusterings and x-axis is threshold

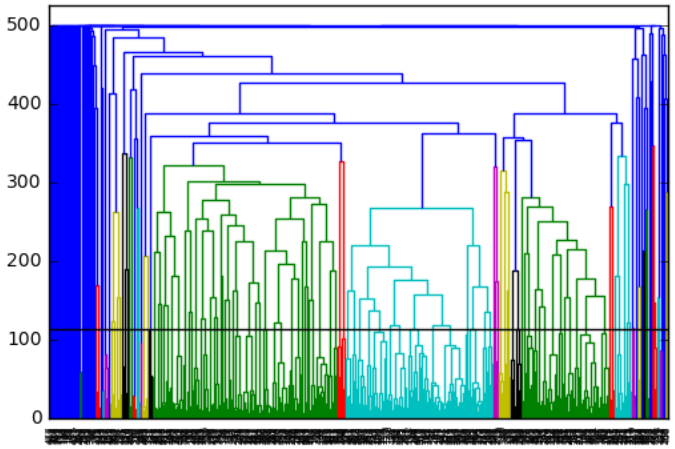
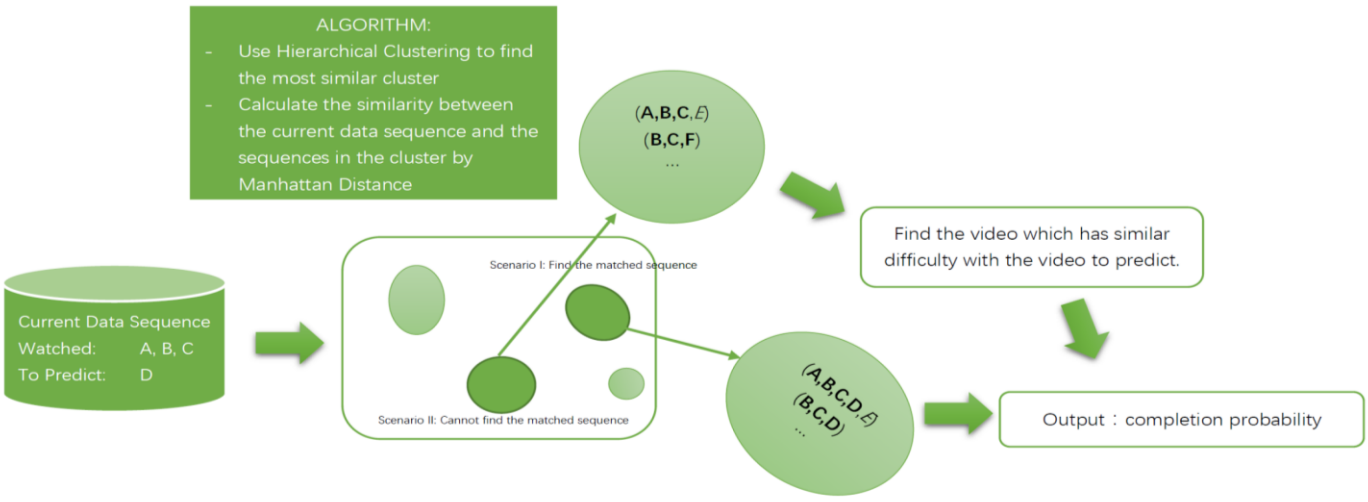


Fig.6 Clustering results, where y-axis is clusterings, x-axis is student's ID, dividing line is threshold from Fig.5

D. *Prediction*

To predict whether a user will complete a video, we first determine if there has been other students who share the same video sequence. In the case where no other students in the records share the same subset of the user, we will find a video that share similar difficulty level as the one to be predicted based on the video level classification earlier. The resulting output of the prediction model is the probability of whether the user completes the entire video.



5. RESULTS AND DISCUSSION

Different performance measurements were used to evaluate the model: precision , recall and F-measure. The dataset was split into different ratios of training and testing datasets and we plotted the results in the following graphs. In general, the model does not fare well when training and testing dataset are divided equally; the results tend to be better when we increase the proportion of the training dataset to 70%.

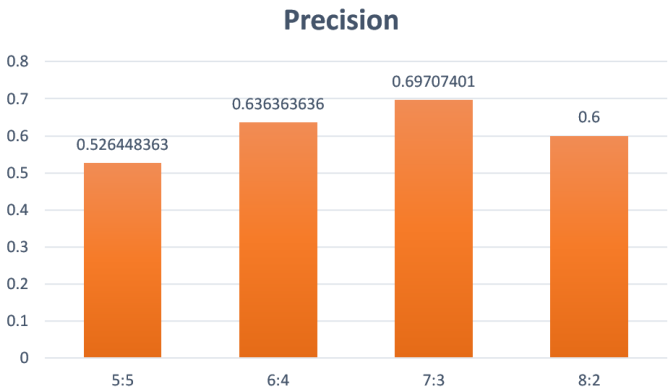


Fig.8 Precision result, where x axis is training dataset and testing dataset

The precision value is above 0.5 for all cases, with the highest being 0.70 when training dataset consists of 70% of the full dataset.

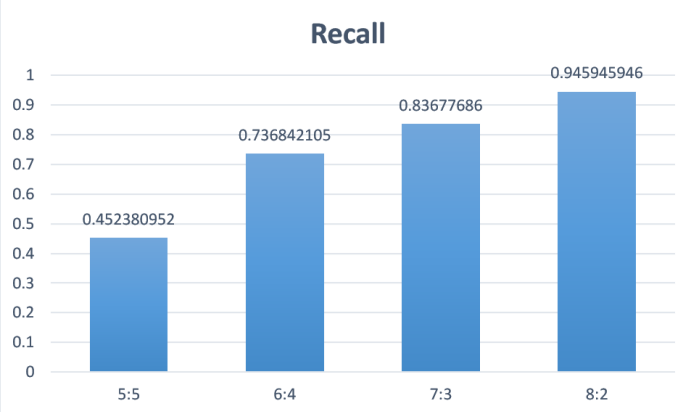


Fig.9 Recall results, where x axis is training dataset and testing dataset

In the case of recall results, the greater the training dataset, the higher the recall rate. When training dataset consists of 80% , the recall rate is highest at a value of 0.9.

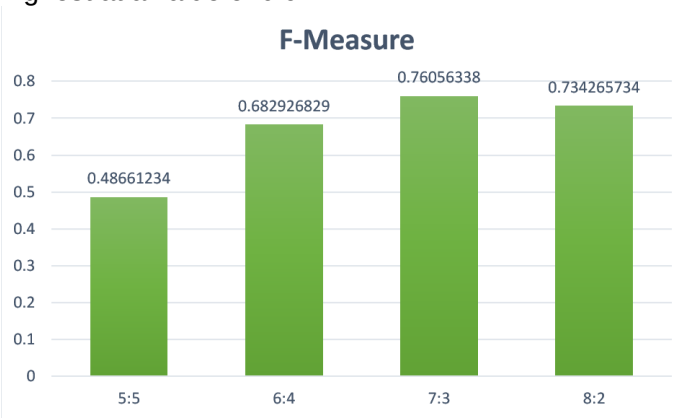


Fig.10 F-measure results, where x axis is training dataset and testing dataset

Similar to precision results, the f-measure value is greatest when the ratio between training to testing data is 7:3.

6. CONCLUSION AND FUTURE WORK

In this project, we attempt to predict whether a user will complete watching a video and finish the listening tests on the English-learning platform. The model was built by first finding the common video subsets between every pair of students and measuring the distance using their listening score trends. Following which, complete linkage hierarchical clustering was performed to construct clusters of students with similar listening score trend, and the clusters will then be used to predict whether a student will finish listening test.

Being able to forecast whether a user will complete watching a video will enable Hopenglish to identify if the video should be recommended to the student. This helps them improve the quality of service for its users, thus sustaining their interest in using the site. In addition, it can help to boost the business when users benefit from the videos and spread it through word-of-mouth, generating more interest to use the website.

This project can further be enhanced on by applying the model to understand inactive users on the platform such as whether their inactivity is due to the lack of video engagement.

REFERENCES

[1] Hierarchical clustering (scipy.cluster.hierarchy). https://docs.scipy.org/doc/scipy-0.18.1/reference/cluster.hierarchy.html.

[2] Manhattan distance. https://xlinux.nist.gov/dads/HTML/manhattanDistance.html.

[3] Time-series similarity measures. http://quant.stackexchange.com/questions/848/time-series-similarity-measures.