MOOCs RECOMMENDER BASED ON USER PREFERENCE, LEARNING STYLES AND FORUM ACTIVITY

Anushka Yohan Liyanage
Department of Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
aliyanage44@gmail.com

Tharushi De Silva
Department of Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
desilva.watp@gmail.com

Sachin Pathirana
Department of Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
sachin.lahiru@gmail.com

Sameer Hilmy
Department of Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
smrbasil4@gmail.com

Nuwan Kodagoda
Department of Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

Kushnara Suriyawansa
Department of Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

***Abstract*— With the development of MOOCs (Massive Open Online Courses) as a major source of e-learning materials, the number of MOOCs available today has become dauntingly high. Furthermore, MOOCs are produced in many different video production styles and these styles play an important role in helping the consumer stay engaged and interested in the course throughout. However, due to the sheer number of MOOCs available today, it is becoming increasing difficult to find the MOOCs that suits your personal preferences and the learning style. This paper describes how thousands of MOOCs that belong to different styles are identified efficiently while each consumer's preferences are identified to provide personalized MOOC recommendations. Furthermore, the paper describes how forums can be analyzed to identify how consumers feel about MOOCs that they followed, which is a crucial metric in recommending MOOCs to consumers.**

Keywords*— *E-Learning, Machin Learning, Natural Language Processing, Index of Learning Styles, Containerization, Parallel Processing, Engagement Level Mapping.

I. INTRODUCTION (*HEADING 1*)

MOOCs have been steadily rising in popularity with more than 101 million consumers spread across 5 major MOOC providers that provides more than 11,400 MOOCs [1] as described in a survey conducted by Class Central in 2018. While these numbers do show that there is a market for MOOCs, what it does not show is the low completion rate of MOOCs. For multiple reasons such as lack of interest and peer reviews, a substantial number of consumers drop out of MOOCs [2].

While there are MOOC search engines such as Class Central, MyMOOC that search across multiple MOOC providers as well as search functionalities provided by MOOC providers themselves, it is important to note that these search engines consider the area of study and keywords to produce recommendations. But, to address the issue where consumers lose their interest, one clear solution is to provide MOOCs that resonate with their preference of video style and user experience. Furthermore, peer reviews play a major role in influencing consumers to pick one MOOC over the other[2].

MOOCRec V2 is designed to address these ongoing issues, in order to uplift the rate of completion of MOOCs. MOOCRec V2 service is designed to identify how engaged a consumer with each of the 6 major video production styles. By using this as a baseline, the search results are refined using the consumers' sentiments expressed in online discussion forums for each MOOC. This unique combination of identifying consumer's engagement and the sentiments of thousands of other consumers who have followed a given MOOC, the service is capable of generating highly personalized recommendations for the consumers.

II. LITERATURE REVIEW

The completion rate of MOOCs has been found out to be low by many researches done on the engagement of consumers with MOOCs over the duration of of courses [3].Felder-Silverman Learning Style Model [4] has highlighted the importance of acknowledging that an individual can inherit a learning style or few that he or she will be more inclined

towards and dimensions that an individual can fall into, such as Active or Reflective, Visual or Verbal, Sensory or Intuitive, Sequential or Global.

Moreover, it was realized that Index of Learning Styles (ILS) Questionnaire [5] [4] exists to determine the dimension that an individual fit into, from the aforementioned set of learner dimensions[6], devised by Richard M. Felder and Barbara A. Soloman [6]. The Questionnaire is lengthy and consists 44 multiple choice questions [7][6]. It is proven that longer questionnaires have lower response rates by Micheal J.Roszkowski and Andrew G.Bean [8].

Upon further investigation, Petteri Nurmi and Tei Laine states that through HCI User Modeling, we can find the goals, knowledge background, traits, context of work and also the interests of a user who interacts with a system [9]. Using an Analytical Model, a system can simulate the cognitive process that carries out while a user interacts with a system [10]. Hence it explains that given user's engagement to a specific web content can be analyzed by using on screen HCI Analytical Techniques such as mouse hover, scroll, rate to content, flip, and skip watching the content and such by User Experience of On-Screen Interaction Techniques research done in 2013[10].

Evidently, with more investigation, it is found that a two-dimensional mapping which is prepared considering learning style characteristics and MOOCs characteristics exist[5] in order to help decide, which learning material type suits which learning style., i.e. if a MOOC video contains a given percentage of a talking head video style, then the MOOC video is more likely suitable for an intuitive, verbal and global learner[5]. Therefore, based on above facts, it was concluded that based on human computer interaction techniques, we can identify how engaged that individual is, with the given task.

Users who learn through peer-to-peer interactions need to find MOOCs which have an active community, so that the user too will be able to participate in it. Previous research has shown that discussions have a positive impact on learning effectiveness [11].

The research paper [12] describes an intelligent web crawler called 'iRobot' which can identify web forums with minimal initial data and figure out how to traverse the forum. Though the system in that paper does not find content relevant to a specific MOOC or topic it has a major advantage in the fact that it can identify crawling paths through forums by itself, therefore it will be able to cover a larger number of links.
The web crawler described in [13] had aimed at gathering data using regular expressions which is a template-based processing method. As a result, the information acquired is structured which makes it easier to process them later, but the disadvantage to this method is that the regular expressions must be explicitly written for the data that the system is attempting to collect. So, the system could miss out on potential information if it is not designed to do so.

A general sentiment analyzing model which was trained using 300,000 twitter posts was presented in [14]. The results of the research paper show that the model was able to identify user sentiments with decent accuracy. A multinomial Naïve Bayes classifier had been chosen to use the model data to make the classifications.

Sentiment analysis has been done in a much different approach in [15] by B. Pang et al. Before the data is used to trained or classified, the system proposed compresses and removes any unnecessary subject domain text while retaining polarity information. Basically, it compresses the data before processing, as a result the performance was increased. In addition to that, the process increases the accuracy because the removal of unnecessary data makes the data cleaner and has a lower chance of being misidentified by the Naïve Bayes classifier.

Though previous research has shown positive results in the field of web crawling and sentiment analysis, they were not regarding the context of discussion forums relating to MOOCs. Therefore, research is needed to determine whether information (ratings derived from sentiments, other forum thread metadata) contained within forums can be used to improve MOOC recommendations.

MOOC resources include multiple modalities such as lecture videos, audio transcriptions, slides, textbooks, forum discussions and clickstream log data. Among them, lecture video is arguably the central and omnipresent component for knowledge transfer, to which other data modalities support [16]. Thus, we focus on designing a method that can organize video resources to dynamically fit different learners. Generally, the types of content in MOOC videos include talking head, slide, coding, animations, writing (khan academy), conversations etc. In production of some MOOC videos, there are lot of transitions between visual views, e.g.: switching from a talking head to a slide or switching from a slide to an animation.

III. METHODOLOGY

A. *Parallel Video Analysis*

The approach to classifying multiple, distinct parts of a single video file simultaneously is based on SIMD principle (*Single instruction, Multiple Data-streams*) of parallel computing. Due to the nature of SIMD, a centralized analyzer is needed in this use case to devise a classification for the full video file.

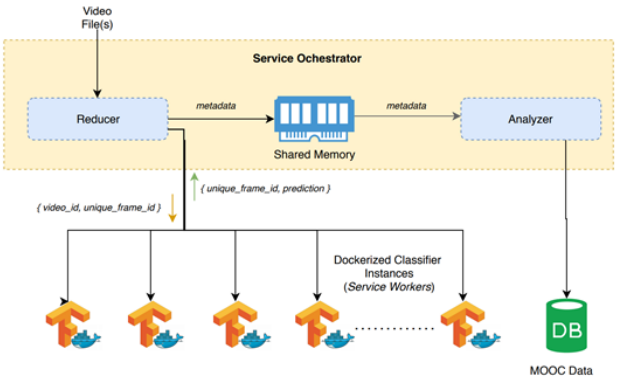

Figure 1

As shown in the architecture, service workers are simply tasked with classifying a given part of the video file. The classification give for all such parts of a video is then put together to provide a single classification for the entire video as follows.

$$P(Head) = \frac{\sum[\,P(h_1) + P(h_2) + \cdots + P(h_n)\,]}{n} \quad (1)$$

$$P(Animation) = \frac{\sum[\,P(a_1) + P(a_2) + \cdots + P(a_n)\,]}{n} \quad (2)$$

$$P(Slide) = \frac{\sum[\,P(s_1) + P(s_2) + \cdots + P(s_n)\,]}{n} \quad (3)$$

$$P(Code) = \frac{\sum[\,P(c_1) + P(c_2) + \cdots + P(c_n)\,]}{n} \quad (4)$$

$$P(Khan) = \frac{\sum[\,P(k_1) + P(k_2) + \cdots + P(k_n)\,]}{n} \quad (5)$$

Once all the parts of a video file are classified, the probabilities of each label are added separately, and their averages are taken. The label with highest average is considered as the classification label for the entire video.

The parallelization is achieved by containerizing the classifier into Docker containers. These containers subscribe to a message queue which contains information about parts of video files referred to as "video chunks" henceforth. Whenever a container intercepts a message in the queue it will classify the video chunk and return the classification label back to the queue.

B. *Identifying user's preffered learning material type*

To identify user's learning style in a more practical manner, MOOCRec V2 is replacing the ILS questionnaire which helped user identify the learning style with an introductory video which will be identifying user's preference by tracking Human Computer Interaction.

In 'Identifying user's preferred learning material type of a MOOCs user through HCI techniques phase, the system displays an interactive introductory video session which consists of all MOOCs learning material types that are recognized by the proposed system namely animations, talking head, presentation slides, khan academy writing, code/tutorial that are suitable for all types of learners. These videos are of same duration and of same subject weight.

Our HTML 5 video player is designed in a way where users can play, pause, skip to next video, replay the same video, skim through video as well as rate the video content. Furthermore, the HTML player is designed adopting HCI techniques such as affordance, constraints, attention and workload models. During this session, we collect user feedback by using analytical HCI techniques such as dropdown point analysis, skim through rate, getting user ratings, mouse scroll motion captures (for transcript viewing), button clicks, etc. As a result, our algorithm determines, during which phase of the interactive video session, the user was mostly engaged with the system.

By doing so, we are left with a sizable amount of information about the engagement of the user across multiple video production styles that belong to different learning dimensions. Using this information, the application predicts the video production style(s) that is in tune with the user's learning dimension. To determine the corresponding learning style for a given video style, we carry out a mapping of MOOCs with FSLSM (a video style to learning style mapping) which was derived through literature review. Finally, this prediction result is derived through a decision-making algorithm.

Consider a single segment(i) in the interactive session.

$$A = \left[\frac{\Delta t_i}{T} * (-1)\right] \quad (6)$$

where Δt_i is the skipped duration of segment i and T is the total lenght of that segment

$$B = [\,\Sigma(P_{n)} * 1\,] \quad (7)$$

where P_n represent a positive action in segment i

$$C = [\,\Sigma(N_n) * (-1)\,] \quad (8)$$

where P_n represent a positive action in segment i

$$D = Q_1 + Q_2 \quad (9)$$

where Q represent a question asked in the segment i

$$\boldsymbol{Score(S_i) = A + B + C + D} \quad (10)$$

By adding the positive points gained by positive interactions and the ratings provided by the consumer at the end of each segment, as well as the negative points gained by skipping, seeking through the segments, we can give a final score to each segment that represent how positively engaged, the consumer was, with the session.

Table I. Learning-Dimension to Video Style Mapping based on FSLM

	Talki-ng Head	Anim-ation	Code/ Tute	Present-ation slides	Khan Academy Writing
Sensory			✓	✓	✓
Intuitive	✓	✓			
Visual		✓		✓	✓
Verbal	✓		✓	✓	✓
Active		✓	✓		
Reflective				✓	
Sequential			✓	✓	
Global	✓	✓			

C. Video Classification

MoocRec1 has classified only coding, slides and talking head styles. In moocRec2, we are classifying all 6 styles given above. Transfer learning has been used to train the VGG16 model used by MoocRec1. Transfer learning generally refers to a process where a model trained on one problem is used in some way on a second related problem. In deep learning, transfer learning is a technique whereby a neural network model is first trained on a problem similar to the problem that is being solved. One or more layers from the trained model are used in a new model trained on the problem of interest [13].

OpenCV has been used to split video frames and image into frames. Then image frames are classified into video styles and composition of each style is calculated.

If n is the total number of frames split by the classifier and a is the number of frames classified as animations, then the composition of animation in a single video is given by:

$$animation = a / n * 100 \quad (12)$$

Finally, total composition of the video of the video styles of a course is calculated by calculating the average of each video production style.

D. Forum Crawling, Scraping and Analysis

The forum analysis process requires a significant amount of data to be able to calculate the ratings. The larger the amount of data the more accurate the ratings tend to be as outliers are balanced out.

To satisfy this, three web crawlers were implemented to gather data from the MOOC platforms Coursera, Edx and FutureLearn. The Coursera and FutureLearn crawlers gather data by simulating a web browser and extracting data from the rendered web pages. The Edx crawler uses a hybrid approach as it simulates a web browser to authenticate and navigation purposes but collects the HTTP responses while they are on their way to the browser. The Edx crawler is capable of gathering more in-depth information than its counterparts due to the fact that it extracts JSON data which are received directly from the server.

Once the data gathering processes have concluded the analysis process will begin. The result of the analysis process is two rating scores, the overall course rating score and the forum activity score. The overall course rating is a measure of how useful the learner the course is while the forum activity score is a measure of how active a forum is. The high-level diagram of the forum analysis component is shown in Figure XX.

Online discussion forums regarding MOOCs generally fall into two categories; questions or discussions. Questions are threads in which discourse about the course content takes place. On the other hand, discussion-type threads are regarding the MOOC itself i.e. how well the teacher explains the topic, the language fluency, whether the topic of the subject is covered, etc. Complaints usually fall into this category and will be detected by the sentiment scores of the thread text data as it will show up as having relatively lower scores. Sentiment of text refer to how positive or negative the piece of text is.

The course rating is calculated based on the course rating available on the MOOC platform the course was extracted from and the sentiment scores of discussion-type threads as shown in equation 13.

$$C = (K_1 * P) + (K_2 * S) \quad (13)$$

Where,

C = Course Rating
P = The rating the course has on its platform
S = The sentiment score calculated based on the discussions
K_1 and K_2 are constants of values between 0 and 1

The forum rating is calculated based on the statistics derived from the metadata of the forum threads. Metadata here refers to the attributes of the threads and posts such as ‘thread creation date’, ‘no. of unique users on thread’, ‘no. of replies(posts) of a thread’, etc.

IV. Research Findings and Results

Identifying user’s preferred learning material type module was evaluated by letting a sample set of 10 Information Technology undergraduates of age group 21-25 years, freely follow our interactive introductory session allowing them to get their preferred learning material type and their learning style as a result from our system.

To measure our system’s accuracy, first we let the same set of students fill the ILS questionnaire to identify their learning style. Then we compared the learner types result we

got for each person from ILS questionnaire, with the learner type result that was given by the system to validate our MOOC learning material-to-FSLSM mapping. The results of the computations and evaluation techniques used are shown in table bleh.

Table II: Computations and Evaluations

HCI evaluation method	Technique used	Test carried out	Test result	
			Success (%)	Error (%)
Evaluating implementations	Experimental evaluation	System is tested with real user participation.	100%	-
Query techniques	Questionnaire	Real users fill ILS questionnaire with maximum concentration.	90%	10%
Observational methods	Think aloud	User is asked to describe what he thinks about his preference of video segments to check if user's selected preference matches with what system suggests.	80%	20%
Evaluating implementations	Experimental evaluation	Main test; If system result = ILS questionnaire result If users actual preferred video style = video style identified by the system	66.67%	33.3%

Furthermore, we compared the time taken for an average MOOCs learner to complete the ILS questionnaire with respect to the time taken to complete our interactive session, in order to measure the time effectiveness of this component. The results of the computation are shown in Table III.

Table III: Computations and Evaluations 2

Time taken to complete ILS questionnaire for an average MOOCs user	Time taken to complete interactive session for an average MOOCs user
7 mins	12 mins

The performance of the three web crawlers were measured in terms of speed and efficiency. The efficiency of the crawlers in retrieving data from courses is shown in Table IV. The speed of the web crawlers is shown in Table V. Three courses from the three platforms were selected to be used to the take performance readings.

With regards to Table IV, the 'actual no. of threads' data points were available on the respective platforms. The data was collected using the platform specific crawler on a computer with an Intel Core i7-6700HQ CPU at 2.60 Gigahertz. The 'extracted no. of threads' data were tracked by the web crawlers itself. The data shown can be considered accurate as it deals with no. of threads which is a discrete value.

Table IV: Web Crawler Efficiency Data

Platform	Course	Actual No. of Threads	Extracted No. of Threads
FutureLearn	Introducing Robotics	242	242
Edx	Python for Data Science	1971	1971
Coursera	Machine Learning	134,946	134,946

With regards to Table V, the elapsed time was tracked by the crawlers themselves. The data shown is an average of five separate executions of the task. The time elapsed data are accurate up to five seconds (± 5 seconds). The error was found by calculating the standard deviation of the data points (standard error). This time information includes the amount of time taken to save the extracted information to the database which was hosted locally to minimize the database handling performance impact.

Table V: Web Crawler Speed Data

Platform	Course	Time Taken to Extract Data (seconds)
FutureLearn	Introducing Robotics	77
Edx	Python for Data Science	498
Coursera	Machine Learning	1472

The most computationally intensive part of the analysis phase is sentiment analysis which involves natural language processing. Forum data of a single thread was sent concatenated as a paragraph before being analyzed by Google NLP because the network transfer speeds make it so that it is unfeasible to analyze posts individually. As a result, an accuracy drop of approximately 5% was observed. The results shown of three separate runs are shown in Table VI which is accurate up to five seconds (± 5 seconds).

Table VI: Sentiment Analysis Performance Comparison

Platform	Course	Average Time Taken to Analyze all Threads of Course (seconds)	
		Google NLP	*VADER Sentiment*
FutureLearn	Introducing Robotics	26	20
Edx	Python for Data Science	240	108
Coursera	Machine Learning	588	342

The actual composition of MOOC videos has been calculated manually and compared with the output of the classifier.

Table i

Platform – Coursera		
Course – Front End JavaScript Frameworks: Angular		
Video Name - Welcome to Angular		
Video Style	***Actual Composition (%)***	***Using Algorithm (%)***
Talking Head	10.7	11.3
Code	0	0
Slide	89.3	88.7
Animation	0	0
Writing	0	0
Conversation	0	0

Table ii

Platform – Khan Academy		
Course – Introduction to logarithms		
Video Name - Introduction to logarithms		
Video Style	***Actual Composition (%)***	***Using Algorithm (%)***
Talking Head	0	0
Code	0	0
Slide	0	5
Animation	0	0
Writing	100	95
Conversation	0	0

While parallelizing a classifier to work on different parts of a single video file gives the idea of drastically reduced classification times, the Table VI shows how classification time reduces for 1000 frames based on the number of service workers working on it.

Table VI: Change of time-consumption based on # of service workers

Number of Workers	**Average Time (seconds)**	**% Decrease in time compared to most time-consuming run**
1	474	0%
2	347	27%
3	289	39%
4	306	35%

V. Conclusion & Future Works

While MOOC platforms have come a long way, the struggle to find the perfect MOOC has been the same way ever since. While MOOC search engines provide different criteria such as field of study, language, they do not consider the personal preference of the consumer. This research paper describes a practical and a novel approach to solving the struggle of searching for MOOCs that matches a consumer's individual preference by proposing an HCI based approach to directly identifying consumers' preference in video styles. Furthermore, this paper describes how sentiments expressed in user reviews within online discussion spaces associated with MOOCs can be leveraged to identifying MOOCs of high quality.

At the same time, this paper shows how containerization based distributed computing can be used to classify an enormous number of video frames of MOOCs that amounts to more than 9400 MOOCs, without relying on expensive GPU based parallelization.

In future, the proposed system can be updated to identify new video production styles should they be introduced to the consumers. At the same time, a machine-learning based approach can be taken to analyze the completion rate of MOOCs recommended by the proposed system, in order to increase the rate of completion of MOOCs by recommending more and more MOOCs that the consumer's preferences resonate with.

VI. Acknowledgment

It is with our great appreciation that we convey our gratitude to the administration of Sri Lanka Institute of Information Technology (SLIIT) for enabling us to complete this research by providing an appropriate and helpful environment. We would like to extend our gratitude towards each and everyone who helped us reach the completion of this research.

VII. References

[1] D. Shah, "By The Numbers: MOOCs in 2018," *Cl. Cent.*

[2] D. F. O. Onah, J. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses: behavioural patterns," *EDULEARN14 Proc.*, vol. 1, pp. 5825–5834, 2014.

[3] M. Khalil, "Learning Analytics in Massive Open Online Courses," no. April, 2018.

[4] R. M. Felder, "A Longitudinal Study of Engineering Student Performance and Retention. IV. Instructional Methods," *J. Eng. Educ.*, vol. 84, no. 4, pp. 361–367, 1995.

[5] S. A. It, "MOOCREC : LEARNING STYLES-ORIENTED MOOCS RECOMMENDER AND SEARCH ENGINE Department of Software Engineering Sri Lanka Institute of Information Technology Sri Lanka MOOCREC : LEARNING STYLES-ORIENTED MOOCS RECOMMENDER AND SEARCH ENGINE October 2018," no. October, 2018.

[6] S. Graf, Kinshuk, and Tzu-Chien Liu, "Identifying Learning Styles in Learning Management Systems by Using Indications from Students' Behaviour," 2008.

[7] R. M. Felder and B. A. Soloman, “Index of Learning Styles Questionnaire.”

[8] M. J. Roszkowski and A. G. Bean, “BELIEVE IT OR NOT ! LONGER Q U E S T I O N N A I R E S H A V E LOWER R E S P O N S E RATES,” vol. 4, no. 4, pp. 495–496, 1990.

[9] T. L. Petteri Nurmi, “Presentation: Introduction to User Modeling,” p. 25, 2007.

[10] S. S. Sundar, S. Bellur, J. Oh, Q. Xu, and H. Jia, “User Experience of On-Screen Interaction Techniques: An Experimental Investigation of Clicking, Sliding, Zooming, Hovering, Dragging, and Flipping,” *Human–Computer Interact.*, vol. 29, no. 2, pp. 109–152, 2014.

[11] C. Kent, E. Laslo, and S. Rafaeli, “Interactivity in online discussions and learning outcomes,” *Comput. Educ.*, vol. 97, pp. 116–128, 2016.

[12] R. Cai, J. Yang, W. Lai, Y. Wang, and L. Zhang, “iRobot: An Intelligent Crawler for Web Forums,” *Proceeding 17th Int. Conf. World Wide Web (WWW 2008)*, pp. 447–456, 2008.

[13] Q. Gao, B. Xiao, Z. Lin, X. Chen, and B. Zhou, “A high-precision forum crawler based on vertical crawling,” *Proc. 2009 IEEE Int. Conf. Netw. Infrastruct. Digit. Content, IEEE IC-NIDC2009*, pp. 362–367, 2009.

[14] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining Microblogging Microblogging = posting small blog entries Platforms,” *Analysis*, pp. 1320–1326, 2010.

[15] B. Pang and L. Lee, “Sentiment analysis using subjectivity summarization.pdf,” 2002.

[16] X. Zhang, C. Li, S. W. Li, and V. Zue, “Automated segmentation of MOOC lectures towards customized learning,” *Proc. - IEEE 16th Int. Conf. Adv. Learn. Technol. ICALT 2016*, pp. 20–22, 2016.

[17] https://machinelearningmastery.com/how-to-use-transfer-learning-when-developing-convolutional-neural-network-models/