Development of Course Recommendation System with Udemy Dataset

J.A.Y.V. Jayakodi, M.B.D.L. Bandara, A.L.F. Shanaz, M.N.A. Hinas, M.N.M. Aashiq, K. Sendurpriyan, S.F.M. Azam

Abstract— Appropriate course selection in Massive Open Online Course (MOOC) platforms is daunting and challenging since there is a lot of courses in the platforms based on requirements and user backgrounds. The following project aims to propose one hybrid course recommendation system which will definitely meet this challenge. This system leverages sophisticated technologies such as Natural Language Processing (NLP) and Machine Learning (ML) to provide tailored recommendations based on an extensive Udemy course dataset. Analyzing the user feedback after recommendations showed an average result relevance score of 87%, hence proving the system to be efficient in recommending precise, personalized courses. The novelty in the design is the combination of content-filtering and collaborative-filtering models, therefore increasing the capabilities of the system by considering specific course content and user characteristics. The system will provide real-time learning, facilitated through powerful data analysis with a user-friendly graphical interface using Python as the development language.

***Index Terms* – Course recommendation, Recommender system, Natural Language Processing, Machine Learning, User-user similarity**

I. INTRODUCTION

EDUCATION is the most powerful weapon in the world. With the improvements in technology, most students have been attracted to online learning. Nowadays, the online education system has become the most popular learning method in the world due to its versatility. Most of the users are enrolled in online courses for their own learning or career development. Therefore, a boom is seen in MOOC providers and the thousands of courses available online. These courses are not only limited to students' educational purposes but also various departments of entertainment education categories like sports, music, dancing, yoga, etc.

JAYV. Jayakodi is with Department of Electrical and Telecommunication Engineering, South Eastern University, Oluvil, Sri Lanka (Email: yasinduviran@seu.ac.lk)

MBDL. Bandara is with Department of Electrical and Telecommunication Engineering, South Eastern University, Oluvil, Sri Lanka (Email: dilinalb@gmail.com)

Shanaz A.L.F. is a Senior Lecturer attached to the Department of Computer Science and Engineering, South Eastern University of Sri Lanka. (Email: shanazlathif@seu.ac.lk)

Dr. MNA. Hinas is a Senior Lecturer attached to the Department of Computer Science and Engineering, South Eastern University of Sri Lanka. (Email: ajmalhinas@seu.ac.lk)

M.N.M. Aashiq is a Lecturer attached to the Department of Computer Science and Engineering, South Eastern University of Sri Lanka. (Email: aashiqmnm@seu.ac.lk)

K. Sendurpriyan is with Department of Computer Science and Engineering, South Eastern University of Sri Lanka. (Email: sendurpriyan@seu.ac.lk)

S.F.M. Azam is with Department of Computer Science and Engineering, South Eastern University of Sri Lanka. (Email: azamsathik34@seu.ac.lk)

Selecting the most useful courses makes the learning process more convenient. Because of the huge number of courses in the online environment, finding the best and most comprehensive courses for users has become a major challenge. The recommendation process performs a specific role when achieving this challenge. Most of the MOOC platforms use this technique to help users find relevant courses by searching using simple keywords. However, the quality of most recommendation systems is not sufficiently high due to the complexity of assessing user knowledge and the continuity in the process.

When considering the recommendation process, two main techniques are typically used, namely content-based and collaborative filtering methods. These methods recommend courses based on the content of the course and the past activities of the user respectively. However, some courses with good content can be missed when using each technology separately. To fulfill this challenge at present, hybrid recommendation techniques are implemented by combining these two techniques. Hybrid technology can be employed to use the course content and the user's historical data for the recommendation process.

This paper presents a recommendation methodology that recommends courses based on user-user similarity where personal interests and career objectives of the targeted user are taken into consideration. This methodology aims to provide effective course recommendations by reducing the high personalization to introduce diversity and novelty to the recommendation process with real-time learning capabilities.

II. RELATED WORKS

Since much research has been carried out related to the recommendation field, categorizing the recommendation systems for the study with justifiable parameters seems an uneasy task. However, to understand the systems, a broad

ISSN: 2961-5992

classification of three categories has been introduced by the researchers namely content-based, collaborative, and hybrid recommender systems.

Content-based filtering recommendation systems allow the usage of content information of items to generate recommendations via matching items and users accordingly. For instance, the content information can be keywords or item titles which enables similarity calculation methods such as cosine/dice similarity [1]. Replication of search engines can also be achieved using content-based methods where users are allowed to search for items within recommender systems [2]. In general, collaborative filtering methods use users' ratings [3] on items to generate recommendations by finding similarities between users or items i.e., Netflix, and Amazon. These systems use various technologies and algorithms i.e., K- Nearest Neighbor (KNN), Random Forest, Support Vector Machine (SVD), Apriori, and SPADE to recommend items [4,5,6,7,17,19]. Instead of using the algorithmic approaches, other techniques such as association rules mining [8] and cross–user domain filtering systems based on score distribution of courses [9] are also used to build collaborative filtering recommender systems.

Hybrid recommendation systems build on combining both content-based and collaborative methods to address individual issues associated with the aforementioned methods. The following text describes briefly some hybrid recommendation systems that employ different techniques to obtain course recommendations.

Stating the issue of collaborative filtering-based systems tend to narrow down the recommendation context and can lead the user towards a filter bubble, a serendipitous recommender system approach was introduced using Machine Learning and Natural Language Processing [10].

To initiate hybrid recommender systems, user profile data is necessary since the recommendation process highly depends on relevant data availability. For instance, a Personalized Self-Directed Learning Recommendation system was developed to guide online learners where course recommendations are generated based on personal interests and preferences of the targeted user, which are gathered via user profiles and personalized queries [11]. On the other hand, a Hybrid Content-Aware course recommendation system was publicized on MOOCs platform named XuetangX, where the text context of the courses, user interests from historical access behavior and course prerequisites have been considered for the recommendation process [12].

Some heuristic approaches are also used to develop recommendation methods in the context of course recommendation systems such as a user-based recommendation system that combines both Collaborative Filtering and Artificial Immune Systems to predict courses by employing cosine similarity and Karl Pearson (KP) correlation [13]. Stating that the highly personalized curriculum available in universities can be often overwhelming for students, an interactive course recommendation system namely CourseQ was introduced by combining visualization techniques with recommendation techniques [14]. In a traditional collaborative recommendation system, some potential courses with good content can be removed due to selections of the threshold value. To minimize this error, a hybrid course recommender system [15] with an innovative hybrid technique is developed by employing extended association rules that combine students' interests and historical records to generate course recommendations.

III. Methodology

The rapid increase in usage of online learning platforms which are generally known as MOOCs leads towards high personalization of course recommendations and confuses users with the question of which course is best to select. To address this issue associated with modern recommendation systems, this system is developed with a user-user similarity measurement method where users with similar background information are preferred when undertaking the recommendation process.

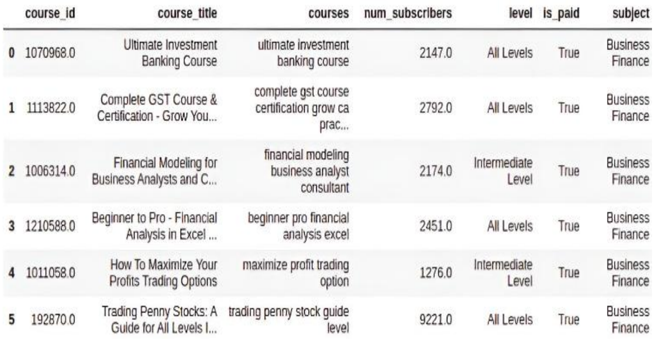
In the phase of developing the course recommendation system, various steps were followed to obtain the desired outputs. In the first step, two models namely the User-user similarity model and the NLP model are built based on two datasets associated with course data and user profile data. Next, an interactive web-based Graphical User Interface (GUI) was developed and then, the two models were integrated into GUI. Finally, the system was evaluated through a user survey to determine the degree of success in achieving desired parameters.

A. Datasets Description

The course dataset contains details of 3678 Udemy courses categorized into 4 subject groups. Fig. 1 shows the topic distribution of the course dataset.

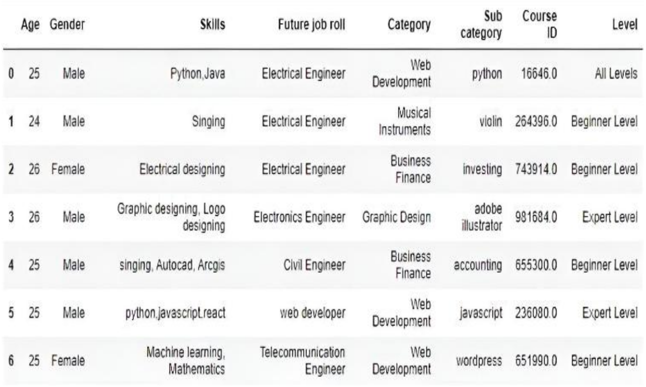
To develop a collaborative filtering model that suggests courses based on user-user similarity, the requirement for a user profile dataset has risen, and the challenge of collecting user profile data was addressed by conducting a user survey.

This surveying approach has been implemented as a questionnaire by carefully studying the well-recognized MOOC platforms such as Coursera, Edex and Udemy and their methods of collecting user-related data i.e., Email, Age, Gender, Education and Skills. Students from Faculty of Engineering, South Eastern University of Sri Lanka participated in this user profile data-acquiring process, and 100 records were documented to the User profile database. Fig. 2 represents the structure of the User profile database.



	course_id	course_title	courses	num_subscribers	level	is_paid	subject
0	1070968.0	Ultimate Investment Banking Course	ultimate investment banking course	2147.0	All Levels	True	Business Finance
1	1113822.0	Complete GST Course & Certification - Grow You...	complete gst course certification grow ca prac...	2792.0	All Levels	True	Business Finance
2	1006314.0	Financial Modeling for Business Analysts and C...	financial modeling business analyst consultant	2174.0	Intermediate Level	True	Business Finance
3	1210588.0	Beginner to Pro - Financial Analysis in Excel ...	beginner pro financial analysis excel	2451.0	All Levels	True	Business Finance
4	1011058.0	How To Maximize Your Profits Trading Options	maximize profit trading option	1276.0	Intermediate Level	True	Business Finance
5	192870.0	Trading Penny Stocks: A Guide for All Levels I...	trading penny stock guide level	9221.0	All Levels	True	Business Finance

Fig. 1. Topic distribution of course dataset



	Age	Gender	Skills	Future job roll	Category	Sub category	Course ID	Level
0	25	Male	Python,Java	Electrical Engineer	Web Development	python	16646.0	All Levels
1	24	Male	Singing	Electrical Engineer	Musical Instruments	violin	264396.0	Beginner Level
2	26	Female	Electrical designing	Electrical Engineer	Business Finance	investing	743914.0	Beginner Level
3	26	Male	Graphic designing, Logo designing	Electronics Engineer	Graphic Design	adobe illustrator	981684.0	Expert Level
4	25	Male	singing, Autocad, Arcgis	Civil Engineer	Business Finance	accounting	655300.0	Beginner Level
5	25	Male	python,javascript,react	web developer	Web Development	javascript	236080.0	Expert Level
6	25	Female	Machine learning, Mathematics	Telecommunication Engineer	Web Development	wordpress	651990.0	Beginner Level

Fig. 2 User profile database structure

B. *Content-Based Filtering Model*

The purpose of developing a content-based filtering model is to achieve semantic processing of course data and user interests in the form of search terms to provide course recommendations using the available course content from the course dataset. For the stage of course dataset preprocessing, some NLP techniques such as Stop words removal and Lemmatization to obtain base words were applied to the dataset to obtain fast execution time in the expected system.

User input for the next stage is obtained as the search terms from the front end of the system and processed by means of NLP techniques as declared in the previous step. Afterward, data vectorization is used to organize the raw data from the course dataset for similarity comparison. In this work, Term Frequency – Inverse Document Frequency (TF – IDF) vectorization [16] is utilized for the analysis and vectorization of course descriptions. This is a technique where higher weights are given to terms most relevant to a user's search query and ignoring common words. Compared with other vectorization techniques such as Count Vectorizer, TF-IDF has enhanced precision, which works quite well for semantic-based recommendations. Latent Semantic Analysis (LSA) was also considered as a form of content filtering, but ultimately discarded since LSA aims to capture abstract topics and not the concrete query of a user. Equations (1) and (2) demonstrate the TF-IDF vectorization value assignment with respect to the reference term.

$$Term\ Frequency = \frac{Term\ i\ frequency\ in\ document\ j}{Total\ words\ in\ document\ j} \tag{1}$$

$$Inverse\ Document\ \ Frequency = log_2\left(\frac{Total\ documents}{Documents\ with\ term\ i}\right) \tag{2}$$

TF-IDF vectorization assigns high values for terms that are specified as the reference term thus terms like article words i.e., “the”, “a”, etc. have less value in the vectorization process and give the advantage of highlighting the specific words in the search term, which is the input of the user.

Upon finalizing the TF-IDF vectorization values are ranked and top-10 course recommendations are conveyed to the user by employing the cosine similarity where most similar courses to the search terms are assigned values closer to zero and non-similar courses are assigned values further from zero.

C. *Collaborative Filtering Model*

Since the traditional approach of developing a collaborative filtering model uses user ratings for each available course on the MOOC platform to generate recommendations, it lacks the degree of personalization. Hence a user-user similarity implementation is used to achieve the recommendation personalization using machine learning techniques.

During the development of this model as the first step, feature encoding was applied to the data fields of the user profile database since the machine learning algorithms only understand numerical values. This process was done by creating binary lists for each data field which is considered to predict course recommendations.

The K-Nearest Neighbor (KNN) algorithm was implemented to discover the degree of similarity between users. KNN algorithm is vastly used in various categories of collaborative filtering recommendation systems other than course recommendation systems [20] and is advantageous over other algorithms due to it is innately capable of performing real-time learning and hence adapting dynamically to new profiles introduced [17]. Computationally, it is less expensive than matrix factorization methods, with a high level of accuracy in assessing similarities while using TF-IDF vectorization [18]. The matrix factorization was also successful in collaborative filtering; the high computational cost made it impractical for a real-time application. Since the user profile database will be subjected to a constant updating process when a new user logs in to the system, the real-time learning ability of the KNN algorithm is useful in regard to predicting course recommendations.

To identify similar neighbors of a given user, a similarity measurement had to be done and this task was achieved by introducing a similarity checking function that uses cosine similarity measuring principles. Each of the four data fields was given a weight of one and neighbors were ranked based on their similarity values from close to zero to further away from zero and assigned to a list alongside their similarity values with respect to the reference user. The ranked neighbors are then directed to obtain their course selection from the user profile database and return the course IDs accordingly which then illustrates the course predictions obtained based on user-user similarity.

D. *Front End Development*

A user-friendly graphical interface was developed as a web-based application to generate course recommendations by employing the aforementioned Collaborative and Content filtering models. Streamlit, a powerful dynamic Python-based web framework was used to design and implement the activities related to the web application. Since this web framework interprets the Python code and generates object-oriented activities accordingly, the behavior of the web application was soft-coded from the beginning as per the requirements.

Using the front-end input objects, data required for both models are routed to generate course recommendations as previously mentioned in the backend development section.

Fig 3 illustrates the complete system architecture of the described hybrid course recommendation system where both content-based filtering and collaborative filtering models are integrated with GUI to provide course recommendations.

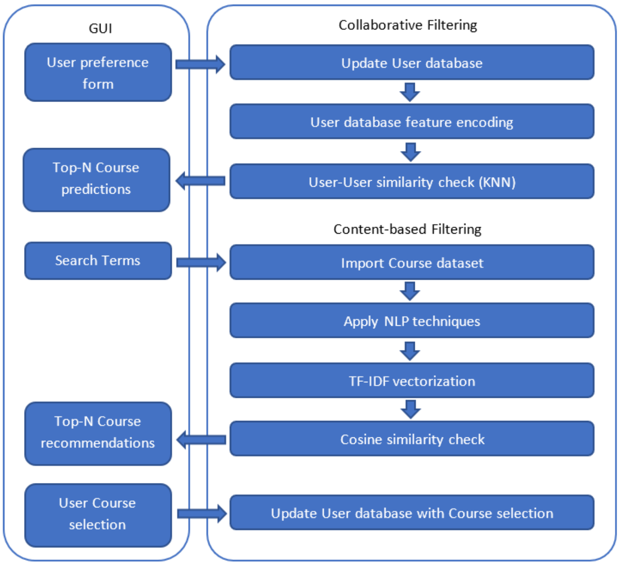


Fig. 3 Recommender system architecture

Using the front-end input objects, data required for both models are routed to generate course recommendations as previously mentioned in the backend development section. Fig 3 illustrates the complete system architecture of the described hybrid course recommendation system where both content-based filtering and collaborative filtering models are integrated with GUI to provide course recommendations.

IV. Evaluation

A user study was done to determine the performance of the newly introduced course recommendation system based on the parameters i.e., diversity, novelty, and relevancy. Existing users from the user database and new users have participated to provide feedback regarding the satisfactory level of their experience through a carefully designed questionnaire. The performance of the hybrid recommendation system was evaluated using feedback from 100 users. The following metrics were recorded:

- Relevancy:

87% of users reported that the recommended courses matched their interests, as measured by a post-survey questionnaire.

- Novelty:

72% of participants discovered courses they had not previously considered, highlighting the system's ability to introduce new options.

- Diversity:

Users observed a diversity score of 78%, indicating that the system successfully provided a variety of course options beyond their typical preferences.

Responses to respective questionnaires are analyzed and the above Fig. 4 visually represents the proportion of courses across four main topic categories based on the Udemy dataset. Each slice corresponds to a specific category, with its size proportional to the percentage of courses in that category.

Comparison with Existing Systems: Whereas XuetangX focuses on course prerequisites and past user interactions, this system integrates real-time learning, individual user profiles, and course content for a better-personalized learning experience.

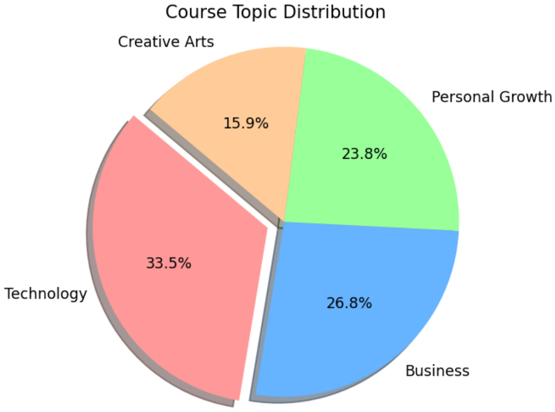


Fig. 4 Course topic distribution

Traditional collaborative filtering systems, like Netflix, usually rely greatly on user-item ratings and sometimes completely ignore item content. This system overcomes the limitations of a ratings-only approach by combining user similarity with content-based recommendations.

This means that the combination of qualitative and quantitative feedback demonstrates the strength of a hybrid system in providing personalized, diverse, and creative recommendations.

V. Conclusion

In conclusion, this system demonstrates how a hybrid course recommendation system assists users in selecting the most suitable course. Most of the users have been influenced by online education from home because of this pandemic situation. Therefore, it is very useful for students to employ this concept which can be used in any place such as institutes, universities, and academies by using their historical dataset.

A hybrid recommendation system comprises two parts namely collaborative and content-based filtering systems. Here, a collaborative filtering system is developed by using the k-nearest neighbor algorithm and cosine similarity which is used to find similar users. Machine learning techniques were implemented using Sci-py.

The content-based filtering model is mainly based on the content of the course. NLP methods are used to implement and find similar courses according to the user's search terms. When considering the content of the courses, stop words That have the least importance for the recommendation process such as a, an, the, and etc. were removed using the NLTK library to improve the efficiency of the recommendation process. TF-IDF vectorization is used to vectorize the words

ISSN: 2961-5992

and cosine similarity is used to find similar courses according to the course content.

The targeted objective of reducing the degree of personalization is achieved using these techniques and performance parameters are evaluated with a higher success rate where recommendations are introduced as novel, diverse, and relevant to user preferences and profiles.

References

[1] J. Naren, M. Z. Banu, and S. Lohavani, “Recommendation system for students’ course selection,” in Smart Systems and IoT: Innovations in Computing, Singapore: Springer Singapore, 2020, pp. 825–834.

[2] Q. Cheng and Y. Gao, “Courducate-An MOOC Search and Recommendation System,” [Online]. Available: http://www.cs.virginia.edu/~hw5x/Course/IR2015/docs/Projects/Samples/2.pdf.

[3] B. C. Uzo, “Towards an Efficient Machine Learning Algorithm for a Graduate Study Elective Course Recommendation System,” vol. 11, no. 41, pp. 5036–5042, 2021.

[4] R. Obeidat, R. Duwairi, and A. Al-Aiad, “A collaborative recommendation system for online courses recommendations,” in 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), 2019.

[5] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars, and W. Suparta, “A proposed framework in intelligent recommender system for the college student,” J. Phys. Conf. Ser., vol. 1402, p.066100, 2019.

[6] K. Nasaramma, M. Bangaru Lakshmi, G. Prasanna Priya, and G. HimaBindu, “Recommendation system for student e-learning courses.” 2019.

[7] Y. Madani, M. Erritali, J. Bengourram, and F. Sailhan, “Social collaborative filtering approach for recommending courses in an E-learning platform,” Procedia Comput. Sci., vol. 151, pp. 1164–1169, 2019.

[8] H. Zhang, T. Huang, Z. Lv, S. Liu, and Z. Zhou, “MCRS: A course recommendation system for MOOCs,” Multimed. Tools Appl., vol. 77, no. 6, pp. 7051–7069, 2018.

[9] L. Huang, C.-D. Wang, H.-Y. Chao, J.-H. Lai, and P. S. Yu, “A score prediction approach for optional course recommendation via cross-user-domain collaborative filtering,” IEEE Access, vol. 7, pp. 19550–19563, 2019.

[10] Z. A. Pardos and W. Jiang, “Designing for serendipity in a university course recommendation system,” in Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, 2020.

[11] T. B. Lalitha and P. S. Sreeja, “Personalised self-directed learning recommendation system,” Procedia Comput. Sci., vol. 171, pp. 583–592, 2020.

[12] X. Jing and J. Tang, “Guess you like: Course recommendation in MOOCs,” in Proceedings of the International Conference on Web Intelligence - WI ’17, 2017.

[13] Chang, P.C., Lin, C.H. and Chen, M.H., 2016. A hybrid course recommendation system by integrating collaborative filtering and artificial immune systems. Algorithms, 9(3), p.47.

[14] B. Ma, M. Lu, and Y. Taniguchi, “Exploration and Explanation: An Interactive Course Recommendation System for University Environments,” 2021.

[15] S. I. Konomi, “Design a Course Recommendation System Based on Association Rule for Hybrid Learning Environments,” pp. 1–7, 2019.

[16] Avinash, M. and Sivasankar, E., 2019. A study of feature extraction techniques for sentiment analysis. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 3 (pp. 475-486). Springer Singapore.

[17] Hodovychenko, M.A. and Gorbatenko, A.A., 2023. Recommender systems: models, challenges and opportunities. Herald of Advanced Information Technology, 4(6), pp.308-319.

[18] Godinot, A. and Tarissan, F., 2023. Measuring the effect of collaborative filtering on the diversity of users’ attention. Applied Network Science, 8(1), p.9.

[19] Amin, S., Uddin, M.I., Mashwani, W.K., Alarood, A.A., Alzahrani, A. and Alzahrani, A.O., 2023. Developing a personalized E-learning and MOOC recommender system in IoT-enabled smart education. IEEE Access, 11, pp.136437-136455.