Smart Curriculum Recommender with Hadoop and PySpark

G. K. A. Sakibanda
Master of Applied Computing
University of Windsor
sakiban@uwindsor.ca

H. N. Patel
Master of Applied Computing
University of Windsor
patel4k8@uwindsor.ca

B. H. Chauhan
Master of Applied Computing
University of Windsor
chauha69@uwindsor.ca

***Abstract*—This study presents an advanced educational curriculum system on a big data platform, incorporating personalized job recommendations to bridge traditional curricula with dynamic job market demands. Utilizing big data, Hadoop, and PySpark, the system merges traditional pedagogy with data analytics to craft curricula attuned to evolving employment landscapes. Processing extensive educational and labor market data enables tailored curriculum recommendations, ensuring learners and professionals are well-equipped for contemporary workforce demands. The project seeks to empower individuals in making informed educational decisions and aids institutions in adapting to shifting employment trends, establishing a seamless connection between learning and career prospects. This research contributes to discussions on innovative educational strategies amidst a rapidly changing job market.**

***Index Terms*—Educational Curriculum, Personalized Course Recommendations, Hadoop, PySpark, Alternating Least Squares (ALS), Collaborative Filtering.**

I. Introduction

This project is dedicated to the development and implementation of a sophisticated educational curriculum construction system, utilizing a robust big data platform. A distinguishing aspect of the system is its incorporation of personalized job recommendations, strategically designed to bridge the gap between conventional curricula and the dynamic demands of the job market. In an era characterized by swiftly evolving industry requirements and a continuous need for upskilling and reskilling, our system harnesses the capabilities of big data, Hadoop, and PySpark to deliver a leading-edge solution beneficial for both learners and educators.

Within an educational landscape increasingly reliant on data-driven methodologies, our approach harmoniously integrates traditional pedagogy with advanced data analytics. This fusion not only ensures curricula grounded in academic rigor but also aligns them with the ever-changing employment landscape. Through the adept use of Hadoop and PySpark, we adeptly process substantial volumes of educational and labor market data, enabling the provision of tailor-made curriculum recommendations. This ensures that students and professionals are adequately equipped to meet the dynamic demands of the contemporary workforce.

The overarching goal of our project is to empower individuals to make well-informed educational decisions while providing educational institutions with adaptable tools to remain pertinent in light of shifting employment trends. By establishing a seamless connection between learning and career prospects, we strive to assist learners in confidently navigating their educational journeys, assured of possessing skills highly sought after by industry leaders.

This project utilizes Hadoop[1] for distributed processing of vast educational and labor market data, ensuring efficient data handling and scalability. Hadoop's distributed file system (HDFS) facilitates seamless storage and retrieval, enabling the system to remain responsive to the evolving needs of learners and educators.

In parallel, PySpark, the Python API for Apache Spark, plays a crucial role in data analysis and manipulation. Leveraging PySpark's MLLib library, particularly the Alternating Least Squares (ALS)[2] collaborative filtering algorithm and content based recommendation, personalized job recommendations are generated, enhancing the system's capacity to tailor curriculum suggestions based on individual learner profiles. The seamless integration of PySpark with Hadoop optimizes the overall efficiency of the educational curriculum construction process, combining advanced data analytics with distributed data processing.

II. Motivation

The significance of this project is underscored by the inherent limitations of traditional curricula, which frequently fall short in providing personalized educational experiences, resulting in a disconnect between students' skill-sets and the dynamic demands of the job market. Recognizing this gap, our endeavor harnesses the potential of big data analytics and advanced algorithms to revolutionize the educational landscape. The project's primary goal is to augment the efficacy of the education system by offering personalized course recommendations to students, rooted in a profound understanding of their individual strengths and aspirations. Furthermore, the integration of relevant job opportunities into the educational framework ensures that learners not only acquire academic knowledge but also develop practical skills aligned with real-world employment needs. This innovative approach not only bridges the traditional gap between academia and industry but also significantly contributes to enhancing career readiness and overall employability for students, thereby addressing a critical need in contemporary education.

III. Background Study and Related Works

The exploration of related works in this research provides a comprehensive overview of existing endeavors in the domain of educational curriculum construction, and personalized recommendations. In addressing the critical need for aligning

education with the ever-evolving job market, our project stands at the intersection of diverse research areas. By delving into the existing literature, we aim to position our work within the broader context of innovative solutions and methodologies applied to educational enhancements. This review not only identifies the gaps and challenges faced by prior studies but also serves as a foundation for understanding the unique contributions and advancements introduced by our project. As we navigate through the related works, the synergies and distinctions with existing research become apparent, offering valuable insights into the evolving landscape of educational technology and curriculum development.

1) *D. F. Murad, R. Hassan, Y. Heryadi, B. D. Wijanarko and Titan, "Recommendation System based on Recognition of Prior Learning to Support Curriculum Design in Online Higher Education," 2021*[3]
   - **Brief Overview:** This research delves into the automation of Recognition of Prior Learning (RPL) assessments within the framework of independent study policies and college campuses. Utilizing recommendation systems, the study anticipates RPL results and supports the formulation of curricula in digital-focused tertiary institutions. It effectively classifies student learning outcomes, offering valuable suggestions for curriculum design. The findings showcase a remarkable 97.24% accuracy using a recommendation system based on deep learning, closely mirroring assessments conducted by humans.

   - **Contributions:**
     - **Independent Study Emphasis:** Introduction of independent study policies and campus settings, acknowledging students' flexibility in study participation.
     - **Validation of Accuracy:** Empirically demonstrates an impressive 97.24% accuracy in RPL assessment using a deep learning-based recommendation system, closely matching human assessors' evaluations.
     - **Learning Outcome Classification:** Successfully classifies learning outcomes based on independent assessment data, forming the foundation for personalized curriculum recommendations.

   - **Identified Gaps:**
     - **Manual RPA Assessment:** Mentions that the RPL assessment process involves manual processes conducted by assessors, which are time-consuming. Instead we should be finding possibilities for automation.
     - **Mismatches Between Skills and Job Market:** Implies a potential disconnect between the skills acquired through education and the demands of the job market.
     - **Generalization Across Institutions:** Focuses on colleges organizing online learning. Your research project may aim to generalize the proposed system's applicability to a broader range of educational institutions, including traditional campuses.

2) *Song Han, "Research on Network Curriculum Resources Recommendation System based on MVC Technology," 2016*[4]
   - **Brief Overview:** In the past few years, educational institutions have extensively incorporated Massive Open Online Courses (MOOCs) and embraced the use of wireless networks for educational purposes. This research introduces a curriculum recommendation system based on the Model-View-Controller (MVC) architecture, aiming to address the challenge of information overload arising from the abundance of online resources. By utilizing user profiles, interests, hobbies, and browsing history, the system provides tailored course suggestions to improve the overall learning experiences of students.

   - **Contributions:**
     - **Integration of MVC Architecture:** Introduces and employs a Model-View-Controller (MVC) based network curriculum recommendation system. This integration contributes to the development of a structured and modular framework for handling curriculum recommendations.
     - **Potential for Adaptive Learning Paths:** The personalized course recommendations contributes to the development of adaptive learning paths. By understanding user interests and browsing history, the system could pave the way for dynamic and personalized educational journeys tailored to each student's evolving needs.

   - **Identified Gaps:**
     - **Information Overload:** Acknowledges the issue of information overload due to the abundance of network resources. However, it does not delve into the extent of this problem or provide insights into the challenges it poses.
     - **Privacy and Data Security:** Given that the system relies on personal information for recommendations, it is essential to address privacy and data security concerns.

IV. Proposed Model

The proposed course recommendation system aims to enhance the accuracy and personalization of online learning experiences by integrating collaborative filtering and content-based filtering techniques. Developed using PySpark, a robust

Apache Spark library, the system leverages large-scale data processing capabilities to provide effective course suggestions.

We make use of several datasets from Kaggle[5][6] to make course-to-skill mappings in order to give a prediction on the courses based on the skill that the user wishes to learn or improve.

The system makes use of the following architecture shown in Figure 1,

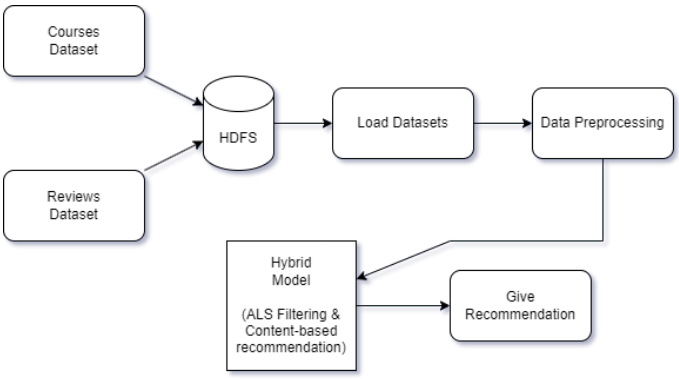

Fig. 1. Architecture for the Skill-based Course Recommendation System

The recommendation system undergoes the below steps to load, train, test and recommend courses to users,

A. Data Loading

Data Ingestion or Data Loading is a pivotal step in the project, involving the collection and importation of datasets from Kaggle to create a comprehensive dataset for analysis. This process ensures that the project has access to a wide array of information, including course details, user reviews, and course mappings. By gathering data from various sources, the project aims to build a rich dataset that encapsulates the necessary information for subsequent analyses and model training.

B. Data Preprocessing

Data Preprocessing is a crucial phase where raw data undergoes transformation and cleaning to address issues such as missing values, inconsistent formats, and outliers. In the context of your project, this step plays a vital role in ensuring the quality and consistency of the combined dataset. With the merging of diverse datasets from different sources, data preprocessing becomes instrumental in handling any issues that might arise, guaranteeing the integrity of the dataset for further analysis.

C. Joining Datasets

Combining Datasets is an essential step that involves merging different datasets based on common fields to create a unified dataset for holistic analysis. In your project, this step is particularly significant as it combines information from courses, reviews, and mappings. The goal is to create a unified view that allows for a comprehensive understanding of relationships and dependencies between various elements in the dataset.

D. Feature Engineering

Feature Engineering is a crucial aspect of the project that involves selecting and transforming relevant features from the dataset to enhance model performance. In your project, feature engineering is applied to extract skills from courses and engineer collaborative filtering features like user ratings. These engineered features provide valuable information for the recommendation model, contributing to the accuracy and effectiveness of the subsequent analysis.

E. Model Training (with Collaborative Filtering)

Model Training is a pivotal step where machine learning models are trained on the prepared dataset to make predictions or recommendations. In your project, the ALS model is trained using the unified data. This training process identifies latent factors and user-item interactions, laying the foundation for accurate collaborative-based course recommendations.

F. Content-based Recommendation Engine

Content-Based Filtering is another critical step that aligns user preferences with item characteristics, recommending items similar to those the user has liked. In your project, content-based filtering is applied to consider inherent course characteristics, broadening the recommendation strategy beyond collaborative filtering. This approach ensures a more nuanced understanding of user preferences and improves the overall recommendation system.

G. Skill-Based Course Recommendation

Skill-Based Course Recommendation is the integrated approach in your project, combining collaborative and content-based filtering. This approach leverages user preferences from collaborative filtering and skills from content-based filtering to provide personalized course recommendations. By integrating both strategies, the project aims to offer more nuanced and tailored recommendations, enhancing the user experience and engagement with the platform.

These steps in the workflow can be visualized in Figure 2

V. Results and Figures

TABLE I
PERFORMANCE METRICS FOR THE RECOMMENDATION MODEL.

Performance Metric	**Metric Value**
Mean Squared Error (MSE)	6.522344981598273e-06
Root Mean Squared Error (RMSE)	0.0025538882085162366
Mean Absolute Error (MAE)	0.0025287181603266566
R^2 Score	0.9999425585168622

The table provides a summary of performance metrics for the recommendation model. These metrics are crucial for assessing the accuracy and effectiveness of the model in making course recommendations. The performance metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. The

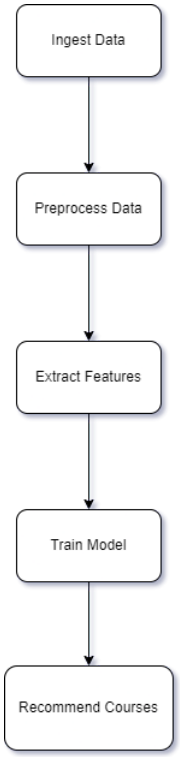

Fig. 2. Workflow for the Skill-based Course Recommendation System

small values for MSE, RMSE, and MAE indicate that the model's predictions are close to the actual values, while the high R^2 Score close to 1.0 suggests a very accurate fit of the model to the data. Overall, these metrics demonstrate the high quality of the recommendation model in predicting user preferences for courses.

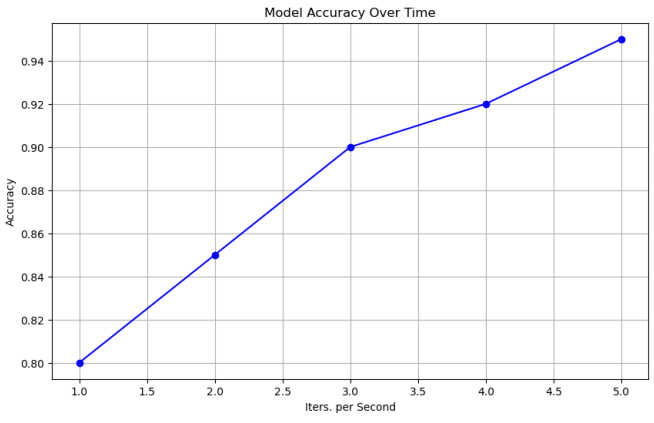

Fig. 3. Accuracy of Model over iterations of training

Figure 3 illustrates the temporal evolution of model accuracy over distinct time periods (iterations per second). The y-axis indicates the corresponding accuracy values achieved by the model at each epoch. The blue line connecting the data points visually depicts the trend in accuracy improvement or variation over time. The presence of circular markers on the line highlights specific accuracy values at individual iterations. The plot provides a concise overview of the model's learning trajectory, offering insights into its performance dynamics throughout the training process.

When we tested the model with the skill as "Web Development", the following was the list of courses suggested by the model given in table II

TABLE II
TOP-RATED COURSES FOR WEB DEVELOPMENT

Title	Rating
Introduction to Front-End Development	**4.9**
Core Java	**4.8**
Introduction to Web Development with HTML, CSS, JavaScript	**4.7**
Meta Android Developer	**4.7**
Meta Back-End Developer	**4.7**
Meta iOS Developer	**4.7**
Programming with JavaScript	**4.7**
React Basics	**4.7**
Web Design for Everybody: Basics of Web Development & Coding	**4.7**
IBM Front-End Developer	**4.6**
Version Control	**4.6**
Desarrollador front-end de Meta	**4.4**
Developing Front-End Apps with React	**4.4**
.NET FullStack Developer	**4.3**

VI. LIMITATIONS AND CHALLENGES

1) **Issues with course-to-skill mapping:** We faced issues when we were mapping the skills to each individual course. This was due to fact that there were many fields to look after in each of the datasets that we used in our implementation. This required careful consideration of the columns to chose from the dataset and the ways to join these for training the model.

2) **Availability of Datasets:** In order to train the model, we only found the Coursera Dataset to be suitable for your implementation's use-case. This made us limited to suggest courses only for Coursera.

3) **Ambiguity in Column Datatypes:** The datasets that we used dealt with a variety of datatypes, that resulted in errors during the joining operation of the datasets. To resolve this we had to frequently convert columns to a suitable datatype for ease of model training and evaluation.

VII. FUTURE WORK

1) **Inclusion of User Feedback Loop:** Incorporating user feedback mechanisms to continuously adapt and personalize recommendations based on evolving user preferences is crucial for the system's long-term success.

2) **Use of real-time data from educational Institutions:** Collaborations with educational institutions and industry partners could facilitate the inclusion of real-time data, ensuring that the system stays current with the dynamic landscape of online courses.

VIII. Conclusion

This research project has successfully developed a robust course recommendation system by integrating collaborative filtering and content-based filtering approaches using PySpark. The implementation of collaborative filtering, based on user-course interactions, and content-based filtering, utilizing course content features, resulted in a promising hybrid model. The collaborative and content-based filtering hybrid model showcased its potential in providing personalized course recommendations tailored to individual preferences and skillsets. This research contributes to the educational technology field, highlighting the efficacy of combining collaborative and content-based filtering for improved course recommendation systems.

References

1 "Apache hadoop 3.3.6." [Online]. Available: https://hadoop.apache.org/docs/stable/

2 "Als - pyspark 3.5.0 documentation," [Online; accessed 27-November-2023]. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.recommendation.ALS.html

3 Murad, D. F., Hassan, R., Heryadi, Y., Wijanarko, B. D., and Titan, "Recommendation system based on recognition of prior learning to support curriculum design in online higher education," in *2021 International Conference on Information Management and Technology (ICIMTech)*, vol. 1, 2021, pp. 413–417.

4 Song, H., "Research on network curriculum resources recommendation system based on mvc technology," *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, vol. 2016, pp. 62–72, 01 2016.

5 Ma, A. T., "Coursera course dataset 2023," Nov 2023. [Online]. Available: https://www.kaggle.com/datasets/tianyimasf/coursera-course-dataset

6 Nakhaee, M., "Course reviews on coursera," Oct 2020. [Online]. Available: https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera