SmartLearn: A Personalized Recommender System for Online Learning

Name:Xutao Zhang

Code: z5641904

Date:2025/7/5

Introduction

With the role digital transformation is playing in changing education, lifelong learning has turned out to be a necessity amongst those who want to empower themselves to stay with the changes that are happening in job markets very fast. Against this understanding, the given project suggests the development of a recommender system, which can be called SmartLearn, the course recommendation application aimed at adults, professional spheres, and people who want to educate themselves in a self-learning mode. The objective is to create a system that will be able to learn and suggest online courses to the user on the basis of the interests, previous learning activities and their own objectives.

This report will demonstrate the full design of the SmartLearn recommender system; this will include the application scope, data sets that may be considered, the recommendation algorithms, evaluation measures as well as a simulated user study. In the given proposal, the authors focus on the integration of content-based and hybrid recommendation approaches that would be used in providing course recommendations in an effective manner that is context-aware, diverse, and individualized to each user. It also provides the factors on scalability, cold-start problems and incorporating the feedback of the users and creation of business values.

1. Scope

SmartLearn is tailored to help the customers locate a related online course in a myriad of topics, such as data science, business, health, psychology, design, or language learning. The main target users will be the adult learners whose ages lie between 25 and 50 years and are either on a personal development program or career advancement program. Such users might lack time and knowledge to physically sieve thousands of online learning opportunities offered on various platforms.

The goal of the system is to offer a limited (6-8) and individually customized list of courses suggestions per logged-in user and per session. The user interface will be in the form of web application and mobile application to make it easily accessible and convenient to the user. When a user logs to the system, he or she will receive a customized dashboard that shows suggested courses, and includes some indicators like the duration of the course, its difficulty level, course provider, average rating and a concise description about the course. The users will be able to react to the recommendations either by clicking on like, save this later, not interested, or enroll.

Feedback of the user will be done in both explicit and implicit ways. Direct feedback: That may present users as liking or disliking a course, saving it to a wish list or enrolling. User behavior as time spent in reading course description or scrolling through a course category is made into implicit feedback. Such encounters will be recorded and applied in improving further recommendations.

Cold-start users with no interaction history will be profiled by the short questionnaire during onboarding, in which they can choose what subjects they would like to learn, what their learning goal is and which options on course they prefer (e.g., short-term vs. long-term, beginner-level vs. advanced). This starting point will be used by the system to bootstrap content-based recommendations and provide them with additional popular and high-rated courses on the same subject areas.

The system is made to manage the dynamic content since online courses are frequently updated. New courses data will be integrated into the system with weekly updates, and new user profile retraining from new interaction logs obtained. Reduced decision fatigue and by making it easier to find courses SmartLearn will provide value to the user. On the business side, the system can combine with profit by becoming an affiliate marketer and share revenue partnership with Coursera, edX, Udemy, and earn commission on course enrollments. Moreover, a pro version may include such features as custom learning track, certification overview, and career development dashboards.

2. Dataset(s)

A list of online training courses will be used as a dataset to SmartLearn, available on such public data repositories as Kaggle, Hugging Face, or scraped metadata of sites that provide MOOCs (Massive Open Online Courses). The Online Courses Dataset provided by Kaggle is one of the appropriate solutions since it contains the metadata of more than 6 thousand courses offered by Coursera, Udemy, and edX. Some of its data attributes are course title, course description, course provider, course subject category, course price and course set of difficulty, average rating, and course duration.

The dataset can be considered quite rich and wide to allow generating plausible recommendation on diversified topics. The content-based filtering will be the most important and as such the description field will be most important since it comprises textual information, and the textual information will be the representation of learning objectives and skills that are covered in each course. The other valuable aspects are subject category (to aggregate likeminded courses), difficulty (to prepare the appropriate level of user), price (to cater to the various budgets), and the rating (to be used as a proxy substitute to the course quality).

Still, there are limitations to this dataset. To begin with, it lacks actual logs on user interaction like the completion rates, click through, or user responses in details. Second, since the dataset is filtered to be used publicly, it can be cleaned or restricted to a particular time frame, i.e., it might not indicate real-time conditions of courses offers and user activity. Third, the most popular datasets available in Kaggle are not time-dependent and cannot be checked against the live APIs; manual updating or simulation is needed to produce dynamic recommendation conditions.

Nevertheless, the dataset also has a great representational power in the experimentation endeavors especially to content-based and hybrid recommendation technologies. Keyword matching and topic preferences will be used to create synthetic user profiles and user behavior to serve as the basis of simulation interaction data.

3. Method(s)

In order to provide personalised recommendations, SmartLearn will use a hybrid recommender system architecture whose structure integrates the advantage of both content based filtering and collaborative filtering approaches.

The content-based will involve taking the course descriptions through TF-IDF vectorization process to form the item profiles. Such profiles will be compared to their user profiles based on their liked or saved courses. In the instances where the course descriptions are too short/insufficient, other more advanced generative models such as the sentence-transformer (SBERT) will then be applied to extract semantic embeddings which can be matched upon based on course intent and not just simple word overlaps.

The composing part of collaborative filtering will be executed with the library of surprise. Since it does not involve actual user interaction, synthesised logs of interaction will be generated to test user-course matrices. An algorithm of matrix factorization, like SVD (Singular Value Decomposition), will be used to learn the latent user and item properties, so that the system will be able to recommend courses that involving similar users had positive interactions with.

A weighted hybrid model will be developed in order to combine these two approaches. It will use composite score relationship on each candidate course; a blend of content similarity, collaborative filtering prediction and popularity (e.g. course ratings or the number of people enrolled). Individual weights of each of the components can be optimized empirically according to the performance in validation.

There are various advantages to this type of hybrid model. Content-based filtering will result in offering file-based personalised suggestions even to the first-time users (solving the cold- start issue), and collaborative filtering will offer diversity in the recommendations as well as serve as an identification strategy of courses beyond the explicitly stated desires of the user. Embeddings give the model the ability to detect semantic connections and instead of giving a direct keyword match they can give recommendations.

4. Evaluation

The SmartLearn evaluation strategy can entail an offline evaluation based on historical/simulated data, as well as simulated user study to gather interaction level data.

Evaluation of the models Offline This standard evaluation of top-N recommender systems includes Precision@5, Recall@5, F1-score, and Normalized Discounted Cumulative Gain (NDCG@10). Such measurements will check that the recommended courses with the highest rankings are of interest to the user. Diversity and coverage will also be assessed so that the system does not overfit to the small user profiles and so that it can bring out new and surprising content.

To evaluate collaborative filtering capability, k-fold cross-validation shall be done using synthetic interaction matrix. Prediction accuracy of ratings or interactions measurement will be using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Moreover, an analysis of the individual model performance (content-based vs. collaborative) and the hybrid model will be undertaken to prove that the combination of these two strategies is worthwhile.

To simulate the testing of the system, user study will be undertaken. Five model users will be designed with different interests (e.g. User A: Data Science, User B: Business Strategy, User C: Psychology, User D: Arts and Design, User E: Language Learning). Each user will use the system through three simulated weeks by choosing courses they like among a list of randomly chosen or tailored recommendations.

By the close of Week 3, logs of user interaction will be used to retrain the system. During Week 4, the users will get a new recommendation list produced with the help of the hybrid model. The new recommendations will be rated in the light of user-specified preferences. To show an improvement in the model, Precision@N and NDCG metrics will be recalculated to Week 4. The findings will be displayed as plots or tables in order to visualize them.

Besides quantitative measures, a feedback questionnaire will be playing out any user whereby he/she will be asked to rate recommendations according to their relevance, novelty, and satisfaction levels. Such qualitative feedback will come in handy to determine system effectiveness in user-experience terms.

Last but not least, the system will be examined in terms of the feasibility of computation. Although developing embedding-based models can be computationally expensive, computing vectors is time-consuming; so that after computing vectors, it has low latency and efficient to make recommendation based on cosine similarity in real-time. The collaborative part will be re-trained in a batchwise manner once a week, whereas the updates to the user than can be done in an incremental way to capture the new interactions.

Conclusion

This paper proposes a project that outlines SmartLearn, a virtualized recommender system that would recommends online coursework under the framework regarding the lifelong learning process. It utilizes the public accessible data, the powerful recommendation algorithms, and design principles that consider the needs of the end user in order to provide the most suitable and ever-changing learning opportunities to the user. The proposed system is to solve cold-start problems, have differences and individualization, and have the possibility of practical application with affiliate monetization.

The SmartLearn approach, a combination of content analysis and simulated working in a team, is the interesting and productive method that may help adult learners to learn new courses and to enrich their education. The project is doable in this course, and it has opportunities to develop creativity and technical advancement in the implementation.