ISSN No:-2456-2165

A Goal-driven Personalized Educational Content Recommendation System for Self-Learning

L M R Jayasinghe
Department of Computer Science & Software Engineering
Faculty of Graduate Studies and Research
Sri Lanka Institute of Information Technology
Colombo, Sri Lanka

Dharshana Kasthurirathna
Department of Computer Science & Software EngineeringFaculty of Computing Sri Lanka Institute of Information TechnologyColombo,
Sri Lanka

Abstract:- With the evolution of the Internet, modern learning patterns have improved significantly, helping to bring knowl-edge wherever the learner requires it. When considering the knowledge available in public domains, the speed of information filtering, and information exchangeability, traditional educational frameworks do have not sufficient powers to move with today's world. As a result, emerging trends indicate that internet-based E-Learning technologies will be used to acquire vast amounts of knowledge. Also with this evolution, self-learning is highly encouraged in the present era. Since there are a lot of internet resources, when someone looks for educational content there could be many similar responses, and this is where we may need personalized recommendations to make life easy. Compared to the existing research of personalized recommendations in theeducation domain, there is a gap in providing plenty of relevant personalized educational resources to a student for self-learning. Thereby, the research, Goal-driven Personalized Educational Content Recommendation System for Self-learning is based on the personal competency level and the areas that need to beimproved of a student. The outcome of the research will provide insight into the perspectives and challenges of personalized recommendation-based self-learning based, primarily web-based learning, and provides a path to identifying appropriate solutions using Topic Modeling for key challenges in the education domain, with personalized recommendations for future research.

Keywords:- Machine Learning, Topic Modeling, Text Ex-traction.

I. INTRODUCTION

With the evolution of information technology, students are encouraged to follow self-learning processes to achieve differ-ent desires in their education and career. If someone browses some educational materials, various online resources can be found within a short period of time in today's world. But if we checked the properties like validity, accuracy, efficiencyof those materials, there could be many valuable as non- valuable resources. These materials can be categorized based on different parameters such as free or payable resources, accurate or wrong resources, efficient or non-efficient re- sources. Also, some students may not be able to find the most appropriate tutorials to their skill level. Individual differences,interests, and learning styles of students all have a significant impact on the influence of learning. The facts like the styleof learning, cognitive level differences, student interests are varied in student-centered personalized learning. For example, there could be learning resources that are not appropriate fora certain competency level of a student. If some student is planning to self-study programming and purchases a course and if it is not matching with his/her knowledge, the effort would be in vain. If there is guidance to students that to be directed towards the correct path on self-learning, it will bea great opportunity for them to be an expert in their selected field of study within a short period of time.

When a student focuses on self-learning and finds resources, he/she may find plenty of resources to study. Although there are plenty of resources, sometimes those will not help the student to be an expert in his/her selected field of study. There is a possibility that some of the resources that a student may select will not match with the competency level of that student. Also, there can be scenarios where the student will not get the exact outcome from some tutorials or courses due to the mistakes on those resources. If a student faces thiskind of problem or else if he/she does not find the correct resources suitable for him/her, the self-learning curve will not be successful. Therefore, the difficulty of finding the most appropriate personalized learning material on self-learning is the problem statement that is going to be addressed through this research.

There are two categories of recommendation algorithms namely, Collaborative Filtering and Content-Based Filtering and each category contains different algorithms to perform the relevant recommendation activities. Collaborative filter- ing method relies on the historical interactions which have been captured between the users and the items, in order to produce new recommendations. Collaborative filtering has a tendency to discover what similar users would like and the recommendations to be offered, in order to categorize users into clusters of similar kinds and recommend each user in accordance with the preferences of its cluster [1]. There two types of Collaborative filtering approaches namely memory based approach and model based approach.

The content-based strategy makes use of further user and item related data. This filtering technique suggests additional products based on the user's past behavior or explicit feedback while also using the item's features. Content-Based Filtering is the primary focus for this research and Topic Modeling is the selected technique for the research.

ISSN No:-2456-2165

The rest of the research paper is organized as follows. Section II will deliver the literature review of recent research papers based on this topic and Section III will explore the methodology of the research conducted. Section IV will deliver the testing and evaluation done for the research while Section V will conclude the paper and highlight the future work to be done.

II. BACKGROUND STUDY

There are various research done on implementing personalized content recommendation systems using multiple algorithms and techniques. Topic Modeling, Reinforcement Learning(RL), Q-Learning are some of the ML techniquesand algorithms which have been used to existing research. A detailed background study of content recommendation systemsrelated research work is described in the below section.

Y. Tang and W. Wang have conducted a literature reviewon various personalized learning algorithms, and their anal- ysis has been aimed at formulating research based on data mining and recommendation systems in personalized learning algorithms, as well as anticipating future research trends [2]. They have analyzed personalized learning algorithms based on Recommendation Systems as well as data mining techniques. Under the topic of personalized learning algorithms basedon Recommendation Systems, they have discussed about the Content-Based Recommendation Algorithm, Collaborative Fil- tering Recommendation Algorithm, and Hybrid Recommenda-tion Algorithm. However, based on their analysis, they have highlighted that the efficiency of a recommendation systemis determined by the data mining algorithm used. Moreover, they have pointed that RL can be an effective method insolving Sequential Decision Making, which has accomplished outstanding results and has become the current revolutionary conceptual intelligence representative ML method.

Y. Chen et al. have researched the topic of ”Smart Education Recommendations” in 2018 and reached up with a solution that provides relevant educational guidance to students in a smart classroom [3]. They have used RL techniques to provide learning guidance based on sensory inputs. They have proposed building a cyber-physical-social system in a smart-classroom that combines numerous sensors, including cameras and a question creator, to monitor students' learning progress and uses RL techniques to provide learning assis- tance associated with multi-sensor data. Primarily, their smart- learning recommendation system has measured each student'sscore of the questions, heartbeat, and facial expressions to articulate learning proficiency, and then based on their currentproficiency level, the system suggests efficient educational activities to students. They have utilized Markov Decision Process to develop their recommendation model, and theyhave done a simulation demonstrating the significance of their smart-learning recommendation system.

K.J. Noh, et al. have conducted their research on implementing a ”Topic Model Based Media Re-creation Service Recommendation System” to carry out a contextually aware suggestion of appropriate media units for a user's intent and context [4]. Their system forecasts the user's rating of a mediaunit that has been looked for a media re-creation service andit compares the similarity of the subjects of re-creating and recreated media for this purpose using a topic vector based on the media's metadata. While they investigate the process of verifying the accuracy of the proposed recommender's theory,the efficiency of the system generated recommendations were upgraded while demonstrating that the Mean Absolute Erroris reduced without consuming a significant amount of time.

K. Edirisinghe has researched on focusing the topic of ”Reinforcement Learning Algorithms For Personalized Recommendations” and presented the strategies of RL in personal-ized recommendation systems on online systems and, websites[5]. In his study, he has stated that RL can be utilized to analyze the users' consent to the offered recommendations on a regular basis and adopt the future recommendations basedon the users' interests. Also, he has described the ability to use RL for personalized recommendation systems to achieve better performance within his findings. This research shows how effective RL-based personalized recommendation systemsare compared to classic recommendation systems. Several traditional systems create fresh recommendations based solelyon the user's past data, however, these methods have some drawbacks. He has discussed how to eliminate such obstacles utilizing this new technique as a result of his research.

M. C. Urdaneta-Ponte et al. have done a systematic review on educational practices and educational recommendation sys-tems [6]. They have focused on the type of education areas, the different implementation approaches utilized, the recom- mended educational content, and the gaps in the education domain for future work. They have analyzed nearly a hundred articles from main databases such as IEEE, Scopus, ACM,and WoS, to find out the systems that the researchers have developed educational recommending systems for users of formal education. As a conclusion for their systematic review,they have presented the possible future research work and development.

With the intention of identifying semantic relationships between TV program description word groups and TV user groups S. Pyo et al. have conducted their research on implementing an Unified Topic Modeling algorithm for TV Program Recommendation using o Latent Dirichlet Allocation (LDA) [7]. They have developed two LDA models for TV users and description words which can be described as topic models and those two models were integrated via a topic proportion parameter. This specifies the simultaneous grouping of similarTV viewers and related program descriptions in a single subject modeling framework and also they have solved the item ramp-up issue by their unified subject model, allowingto make it possible to recommend new Tv programs to TV viewers with confidence. They have made use of electronic program guide

ISSN No:-2456-2165

data of six months time period and real TV watching record data gathered by a TV polling company. The experimental findings reveal that their suggested unified topic model produces 81.4 percent average precision for 50 themes in TV program suggestion, and that it performs on average 6.5 percent better than the topic model of TV users only.

C. Chiang et al. have researched the topic of "Convergence Improvement of Q-learning Based on a Personalized Rec-ommendation System" with the purpose of improving a per- sonalized recommendation system-based strategy to increase the practicality and effectiveness of RL [8]. Also, they have discussed some of the difficulties that could be faced when using RL agents on different streams. They have discussedthe challenge in choosing a correct action for the RL agent considering the balance between exploitation and exploration. An incorrect action might result in an increase in the cost of learning or a failure of learning. Therefore this is one difficulty that they have discussed in their research. Another issue that they have discussed is that in order to achieve rewardand penalty in real-time, the RL learning agent must interact with the environment; yet, the time spent for learning in the interaction process may be excessive. To address the identifiedchallenges, they have proposed a method that uses a person- alized recommendation system to provide a feedback control candidate activity for the RL model to implement self-adaptive learning through teaching. To evaluate the performance of the proposed approach, they have carried out a real-world visual tracking experiment using a pan-tilt camera system.

R. Raghuveer et al. have conducted their research on finding a content recommendation system using Reinforce- ment Learning techniques for MOOC environments [9]. They have focused on the different learning outcomes that eachindividual learner expects in their learning background and their interests. They have highlighted that a common approachof recommending learning outcomes may not suit all the learners as their expectations may differ from one person to another. To address this identified issue, they have proposedan RL-based algorithm to explore the trainee information and provide efficient support for the learner's capabilities and requirements within a specific learning context. According to the findings, the knowledge mastered from the learning anal- ysis was effective in producing personalized recommendation policies that can accommodate the learners' context-specific needs.

III. METHODOLOGY

The system intends to be a blend of three components, with each having its distinct functionalities and responsibilities. The first component is responsible for extract content of PDF documents and slide decks can get the details into CSV format. Analyzing and recording the timestamps along with thevideo content is the accountability of the second component. Third component is the main research component which is responsible to provide recommendations based on user inputs.The following figure 3.1 illustrates the high-level architecture diagram of the complete solution.

A. Personalized Educational Content Recommendation Model

The overall research consists of the feature where the students will be able to participate in a questionnaire and find his/her skill level and the system will suggest the most appropriate content. Students will be carefully guided andmotivated to be an expert in the field through this system. The initial questionnaire will be a predefined set of structured questions in the field of software engineering. The competencylevel of the student will be identified through the analysis of the given answers and also the areas that need to be improved will be identified based on the analysis. According to the identified weak areas, personalized educational content willbe suggested to the student where he/she can practice moreand improve. When a student completes the suggested course material, he will get another questionnaire to evaluate theprogress and to find the area that further needs to be improved. Based on the answers for the second questionnaire, the student will get another course material, likewise, the student will be continuously monitored until he/she will reach the goal ofbeing an expert in the selected study area.

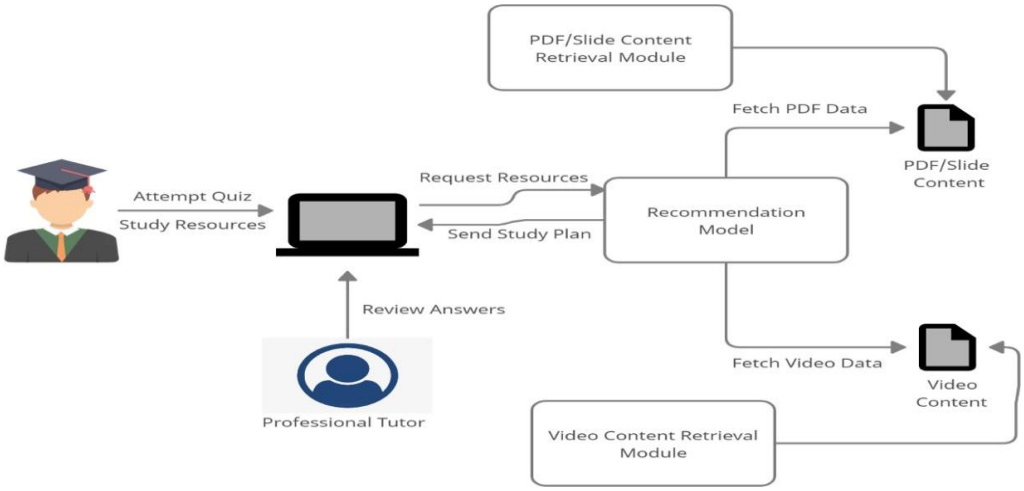

Fig. 1: High-level Architecture Diagram of the System

ISSN No:-2456-2165

A personalized educational material recommendation will be done using a machine learning model. The output gathered from the analysis of the answers to the questionnaire will be the initial input for the resource recommendation. ML model can be developed using Topic Modeling which is one of the main recommendation techniques in ML so that the content can be suggested accordingly.

B. Topic Modeling

Topic modeling is an unsupervised machine learning approach that can scan a collection of documents, detect word and sentence patterns within them, and automatically cluster word groups and similar phrases that best characterize the collection of documents [10]. Since topic modeling is an unsupervised ML approach, it does not require pre-training or predefined training data or list of tag that has been earlier classified by people. Topic modeling is a quick and simple technique to begin data analysis because as it does not require training. In order to identify subjects within unstructured data, topic modeling encompasses counting and grouping words with similar word patterns. A topic model groups comparable feedback and phrases and expressions that are used frequently by identifying patterns like word frequency and word distance.Topic modeling is capable of identifying the hidden themesin collections, categorizing the documents into the revealed themes and using the classification to summarize and search the documents [11]. Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are two topic modelingmethodologies mostly analysts use of.

- *Latent Semantic Analysis:* Latent Semantic Analysis is an NLP methodology that observes the relationships between a collection of documents and the terms contained within them. It scans unstructured data using a mathematical technique named singular value decomposition to discover hidden relationships between the terms and concepts [12]. Concept searching and automatic document classification are the two main uses of LSA. However, it has also been used in areas such as search engine optimization, text summary publishing, analyzing source code in software engineering, and other fields.
- *Latent Dirichlet Allocation:* LDA is a generative probabilistic model of a corpus. The fundamental concept of LDA is that documents are modeled as random mixtures over latent topics, and with each topic being defined by a distribution over words [13]. LDA is also laid under the same underlying assumptions: the statistical mixture hypothesis andthe distributional hypothesis which capable of determininga statistical distribution. The goal of LDA is to map every document in the corpus to a collection of topics that coverthe majority of the words contained in the document. LDA Recommendation model was selected for the implementation of this research.

C. Implementation of LDA Recommendation Model

For topic modeling there are four major phases that we should follow up and those stages can be briefly described as follows. Text Processing, generate dictionary of vocabulary, map corpus using dictionary and training the topic model are the main phases. The automation of interpreting electronic textis referred to as text processing. This enables machine learningmodels to obtain structured data about the text for analysis, manipulation, or generation of new text [14]. Text processing consists of removing special characters, punctuation marksand clean the documents, tokenizing the documents, stem the tokens and remove numerical tokens. Processes laid under testprocessing is described follows.

As the initial step, the required base Python packages and tools which need to eliminate expressions from the scriptsand plotting packages will be imported. As the next step,the required data set should be loaded to the script. Data exploration and preparation should be done as the next step. As the data preparation technique, the minimum and maximumdocument lengths should be determined, so that the basic bounds can be set to decrease the document corpus basedon unfitting lengths. Also, we can dig deeper into the corpus and see which document titles appear the most frequently in the article texts of other documents. This will help to derive the most prevalent documents from the documents collection.

In the pre-processing stage there are several things that are going to be performed such as removing punctuations andthe other characters, removing the stop words, removing the numerical tokens, and then the formatted text will be subjectedinto lemmatization and cleaning. The cleaned data can bevisualize using bigram and trigram models. After removingthe characters and punctuations and clean the datasets, the next steps are to generate the dictionary of vocabulary and mapcorpus using dictionary. In these stages, word to ID mapping, create the words bag of each topic and create the cluster ofall word bags of all the documents will be executed. After completing these two stages, the process of creating the LDA model can be initiated. Topic identification can be concluded with the LDA model implementation. Compute the coherence score and model perplexity will be covered as the phases of LDA implementation. Coherence score and model perplexity can be calculated and draw a plot as below and the topic model graph can be visualize using pyLDAvis library. The results of the designed plot and the topic modeling graph will be discussed in the Results and Discussions chapter.

IV. TESTING & EVALUATION

A. Evaluation Strategy

There are 30 students in Software Engineering stream who has been selected via an initial survey. A questioner with structured questions will be given for all the 30 students and their answers will be evaluated by a IT professional and assigned with a score for different sections. Based on the scorea level will be identified to each section, so that the identified levels will be set as the input for the model and a study plan will be generated in-person. Then they will be split into 2 groups with 15 individuals for each. First group will be given access to the developed system to study with a study plan.The other group won't have access to the system and will be instructed to study on their own for the same amount of time. Then all the 30 students will be faced to another round of exam, and they will be evaluated again by the IT professional, and the measure of improvement will be recorded. If the students who studied using the developed system exhibits a better improvement of their studies than

ISSN No:-2456-2165

the other group, itwill state that the developed system succeeds as expected.

B. Software System

Students can login to the system to attempt questionnaire, view their marks on different attempts, and to get the study materials to improve themselves in their selected study areas. The system allows students to select different topics for the questionnaire where they can attempt different questions in different streams to identify their weak areas. The system will record the scores of every attempt of the student so that the student can self evaluate them by looking at the scores. The professional tutor can login to the system and review the answers of the students, that they have attempted. Fig. 2 depicts the dashboard designed for the students and the tutors to interact with the system.

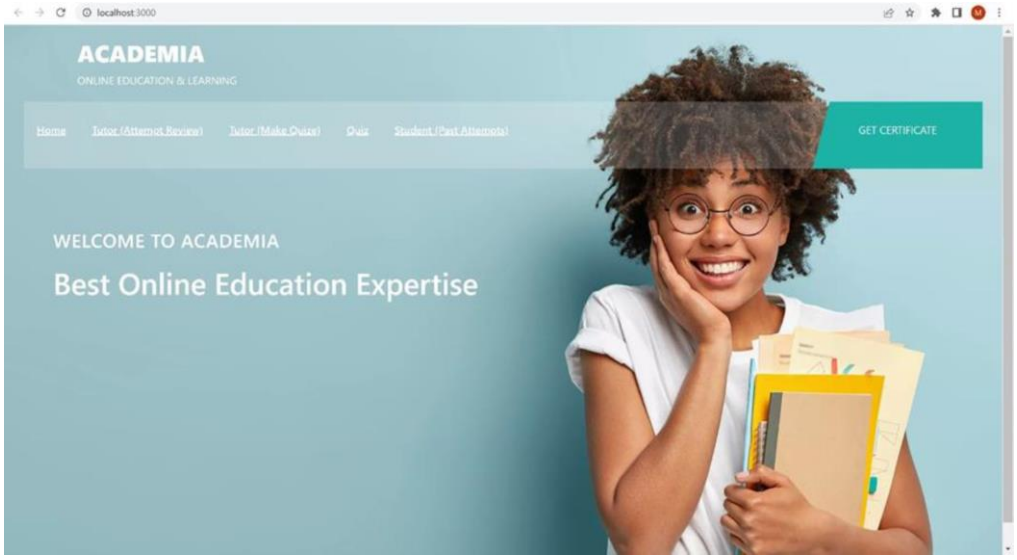

Fig. 2: Dashboard of the Application

Students can attempt to the questionnaire and see their provided answers and the marks that they got in the different attempts from the professional tutor with the use of following UI Fig. 3.

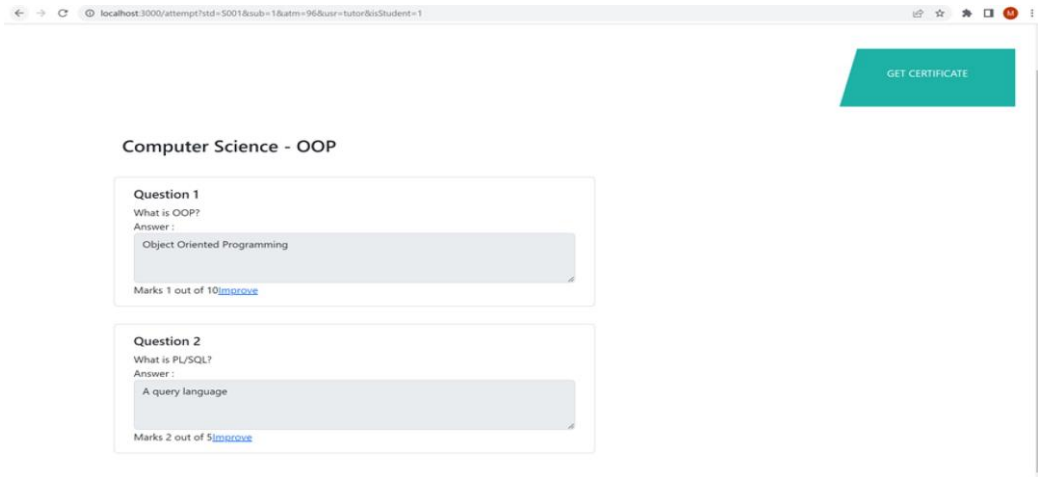

Fig. 3: UI Designed for Student Persona

The study materials will be provided in the following UI where the students can follow them at any time that they needed. The materials will consists of academic pdf documents, slide decks and video materials with timestamps.

ISSN No:-2456-2165

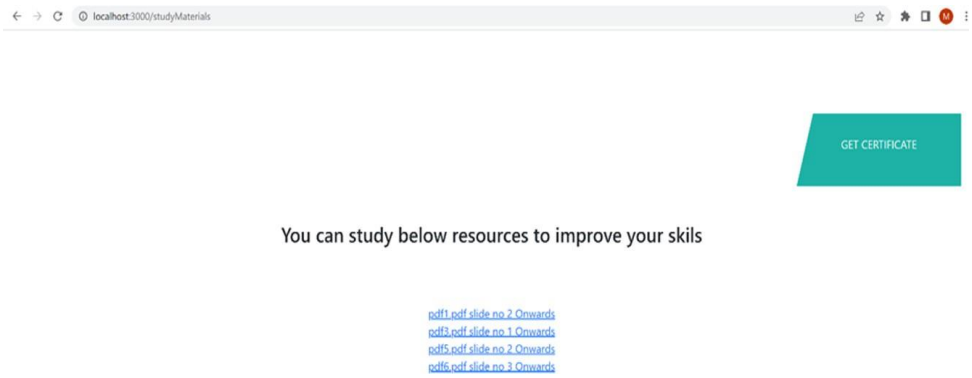

Fig. 4: UI Designed to View Recommendations Suggested by the Model

C. Topic Modeling Model Evaluation

The created Topic model graph is visualized in the following Figure Fig. 4. The graph consists of two panels. The "Intertopic Distance Map" on the left panel shows circles representing various topics, with the distance between them indicating how closely related they are. The size of a topic's circle also reflects the relative frequency of the topic throughout the corpus. By clicking on a topic's circle or entering its number in the "chosen topic" box in the upper left, user can choose that topic for more in-depth examination. The bar chartof the top 30 terms is in the right side. The top-30 most salientterms in the corpus are displayed in a bar chart when thereis no topic picked in the graph on the left. Saliant refers toa term's capacity to discriminate between various themes as well as its frequency in the corpus. One can change the bar chart to display the pertinent terms for each topic by selecting it on the left. The parameter can be used to adjust relevance. Asmaller emphasizes the term's uniqueness more, whilst a largers corresponds to the likelihood of the term occurring across topics. The graph's axes also serve as the primary coordinates for the multidimensional scaling.

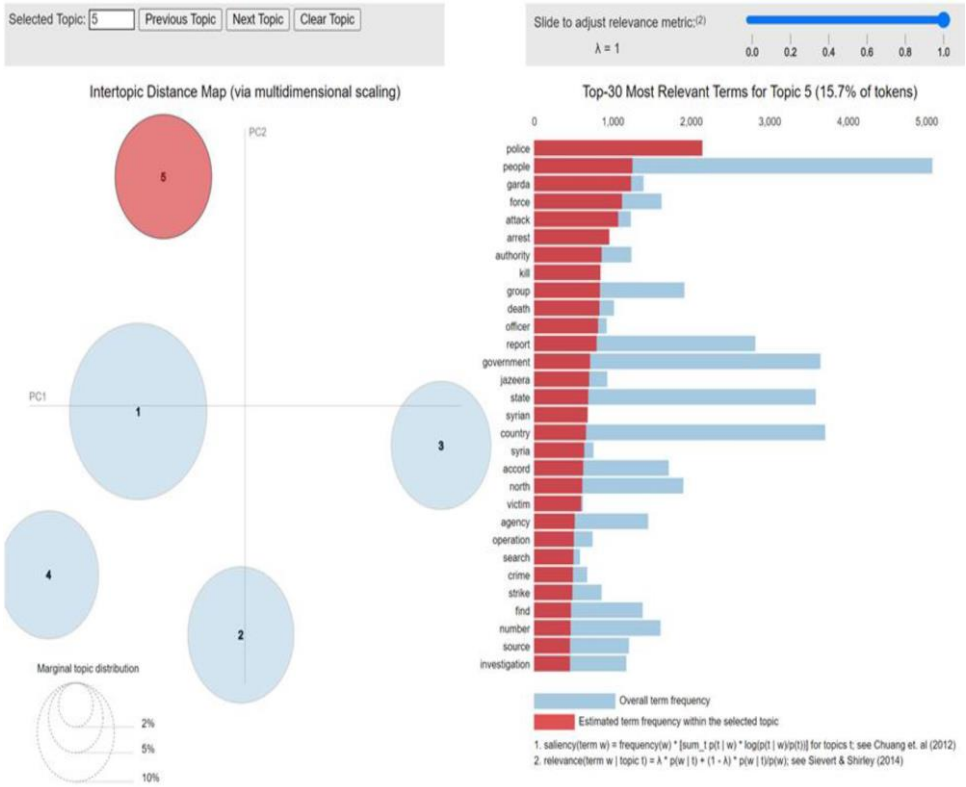

Fig. 5: Topic Model Graph

Coherence score and model perplexity are two measurements that can be used to evaluate the topic model. One frequent illustration of an intrinsic assessment metric is perplexity. It originates from the community of language modeling and tries to reflect the wonder a model feels upon encounteringnovel data. It is quantified as the held-out test set's normalizedlog-likelihood. Topic Coherence measures the score of a single topic by calculating the degree of semantic similarity between high scoring terms in the topic. These metrics assist in separat- ing topics that can be understood semantically from topics thatare the result of statistical inference [15]. There are different coherence

ISSN No:-2456-2165

measures such as C v, C p, C uci, C umass,C npmi, C a. Among from these different measures C v is used to evaluate the implemented model. One-set classificationof the top words, and an explicit validation measure basedon the normalized point-wise mutual information and cosine similarity make up the C v measure. Coherence score and model perplexity plot is visualized below. This plot illustrates the behavior of the coherence score over number of topicsused in evaluating the topic model. Coherence score and modelperplexity plot is visualized below.

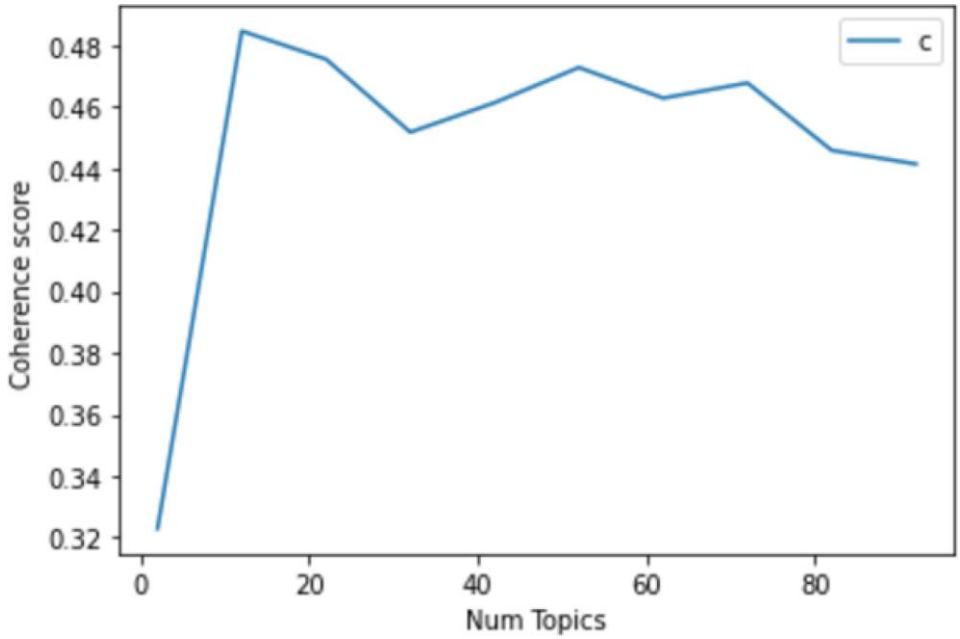

Fig. 6: Plot of Coherence Score

V. LIMITATIONS & FUTURE WORK

While developing this system there was a requirement to suggest video content along with the timestamps of relevant study areas for a student. To fulfill this requirement, therewas a need to generate the timestamps of videos when theyare uploading. There is a limitation on generating timestamps with the existing systems. There is an API provided by googleto convert speech to text and generate timestamps and thatAPI was used to generate the timestamps of the videos forthis system. Although this API provides correct results, thisis a paid service, therefore when the system has large number of videos, the cost that needs to be spend on this will be increased. As of now the system can proceed with a lesser number of videos for users and it has been identified as a limitation.

Therefore, there is a future continuation plan on this research to implement an own methodology to generate timestamps of videos rather using existing google API. So this will be helpful to expand the research further with more video content on the recommendation model.

VI. CONCLUSION

As the outcome of the literature survey, a research gap was identified where there is a lack of personalized recommen- dation systems in the domain of education. Although there is some existing research on recommendation systems, that are capable of providing initial recommendations, they donot have the ability to drive a student towards a goal. This research implementation "A Goal-driven Personalized Edu- cational Content Recommendation System for Self-learning"is to overcome the above-mentioned identified research gap. The main two research areas are machine learning and topic modeling. The outcome of the research is a web application where students will be able to participate in a questionnaire and find his/her skill level and the system will suggest themost appropriate content. Students will be carefully guidedand motivated to be an expert in the field through this system. The implemented system initially will accommodate content related to the field of Information Technology, although it may be implemented in a way that can be easily extendable fordifferent other fields as well.

REFERENCES

[1.] V. Dey, "Collaborative filtering vs content- based filtering for recommender systems," Aug 2021. [Online]. Available: https://analyticsindiamag.com/ collaborative-filtering-vs-content-based-filtering-for-recommender-systems/

[2.] Y. Tang and W. Wang, "A Literature Review of Personalized Learning Algorithm," *Open Journal of Social Sciences*, vol. 06, no. 01, pp. 119– 127, 2018.

[3.] Y. Chen, "Towards Smart Educational Recommendations wi th Rein- forcement Learning in Classroom," no. December, pp. 1079–1084, 2018.

[4.] K.-J. Noh, Y. Park, and K. Moon, "Topic model-based recommendation system for media re-creation service," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, 2016, pp. 132–134.

[5.] K. Edirisinghe, "Reinforcement learning for

personalized recommenda- tions," 05 2020.

[6.] M. C. Urdaneta-Ponte, A. Mendez-Zorrilla, and I. Oleagordia-Ruiz, "Recommendation systems for education: Systematic review," *Electron- ics (Switzerland)*, vol. 10, no. 14, 2021.

[7.] S. Pyo, E. Kim, and M. kim, "Lda-based unified topic modeling for similar tv user grouping and tv program recommendation," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1476–1490, 2015.

[8.] L. Chiang, M. Y. Cheng, T. Y. Ye, Y. L. Chen, and P. H. Huang, "Convergence Improvement of Q-learning Based on a Personalized Recommendation System," *2019 International Automatic Control Con- ference, CACS 2019*, no. 2, 2019.

[9.] V. R. Raghuveer, B. K. Tripathy, T. Singh, and S. Khanna, "Reinforcement learning approach towards effective content recommendation in MOOC environments," in *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)*. IEEE, dec 2014, pp. 285–289. [Online]. Available: http://ieeexplore.ieee.org/document/7020289/

[10.] F. Pascual, "Topic modeling: An introduction," Sep 2019. [Online]. Available: https://monkeylearn.com/blog/ introduction-to-topic-modeling/

[11.] R. Kulshrestha, "A beginner's guide to latent dirichlet allocation(lda)," Jul 2019. [Online]. Available: https://towardsdatascience.com/ latent-dirichlet-allocation-lda-9d1cd064ffa2

[12.] Market Muse, "What is latent semantic analysis (lsa) - latent semantic analysis (lsa) definition from marketmuse blog," Jun 2021. [Online]. Available: https://blog.marketmuse.com/glossary/ latent-semantic-analysis-definition/

[13.] Goyal and I. Kashyap, "Latent dirichlet allocation - an approach for topic discovery," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1, 2022, pp. 97–102.

[14.] "Text processing: What, why, and how," Apr 2022. [Online]. Available: https://www.datarobot.com/blog/text-processing-what-why-and-how/

[15.] S. Kapadia, "Evaluate topic models: Latent dirichlet allocation (lda)," Aug 2019. [Online]. Available: https://towardsdatascience.com/ evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[16.] T. B. Lalitha and P. S. Sreeja, "Personalised Self-Directed Learning Recommendation System," *Procedia Computer Science*, vol. 171, no. 2019, pp. 583–592, 2020. [Online]. Available: https://doi.org/10.1016/j. procs.2020.04.063

[17.] Shyalika, "A beginners guide to q-learning," Jul 2021. [Online]. Available: https://towardsdatascience.com/ a-beginners-guide-to-q-learning-c3e2a30a653c