**MOOCS RECOMMENDER BASED ON LEARNING STYLES**

Porawagama A.S

(IT14142024)

Bachelor of Science Special (Honors) in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2018

**MOOCS RECOMMENDER BASED ON LEARNING STYLES**

Porawagama A.S

(IT14142024)

Dissertation submitted in partial fulfillment of the requirements for the degree Bachelor of Science Special (Honors) Degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2018

**DECLARATION**

I hereby declare that the project work entitled “MOOCs Recommender Based on Learning Styles” (Topic Modelling and Transcript Complexity components) submitted to the Sri Lanka Institute of Information Technology, is a record of original work done by our group under the guidance of Mr. Nuwan Kodagoda (Supervisor) and Ms. Kushnara Suriyawansa (Co- Supervisor), and this project work is submitted in the fulfillment for the award of the Bachelor of Science (Special Honors) in Information technology Specialization in Software Engineering. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma. The diagrams, research results and all other documented components were developed by myself and I have cited clearly any references I have made.

**Authors:**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Registration Number | Signature | Date |
| Porawagama A. S | IT14142024 |  |  |

The above candidate has carried out research for the B.Sc. dissertation under my supervision.

**Supervisor:**

|  |  |  |
| --- | --- | --- |
| Name | Signature | Date |
| Mr. Nuwan Kodagoda |  |  |

**ABSTRACT**

Over the years, Massive Open Online Courses (MOOCs) has evolved as a learning phenomenon providing large number of courses in different domains to a wide range of users. However, learners mostly explore through different platforms and resources before they find the material that best fits their needs. It is a known and proven fact that different learners have different learning styles. As of yet, no system has been implemented which takes the learner’s style into consideration and recommends the best available MOOCs to the learner. The proposal aims at describing processes of building a learner-style centric MOOCs recommender that takes his/her learning style as a valuable input. Different video lecture styles (like talking head, presentation slide with voice-over, animation etc.), available quizzes, resources etc. are considered as significant factors when mapping learning style to MOOCs. We will be focusing only on Computer Science Programming courses on edX, Coursera and Futurelearn platforms. Among various available models, Felder-Silverman Learning Style Model (FSLSM) is selected based on literature review, to identify the learning style of users. Apart from looking for preferred courses, the user can also use the system as an advanced search engine to find any course of choice. The main objective of the system is to present most suitable resources to a learner so that learning and mastering a subject becomes effortless.

# **ACKNOWLEDGEMENT**

The success of the work described in this document was done as our fourth-year research for the subject Comprehensive Design Analysis Project. This project is the result of all the dedicated work of the group members and the encouragement, support and guidance given by many others. Therefore, we would like to express our appreciation to all who gave us the support to complete this significant task.

Authors would like to declare their special thanks for their research supervisor, Mr. Nuwan Kodagoda for his continuous guidance, support and insightful comments which motivated us to move further. Authors would also like to thank their co-supervisor Ms. Kushnara Suriyawasana and lecturer-in-charge, Mr. Jayantha Amararachchi for constantly providing guidance and assistance to implement the research successfully.

Also, authors wish to thank all colleagues and friends for all their help, support, interest and valuable advices. Finally, we would like to thank all others whose names are not listed particularly but have given their support in many ways and encouraged us to make this a success. We will forever be thankful to my former lecturers for supporting us during these past four years.

**TABLE OF CONTENTS**

[**DECLARATION** 3](#_Toc523773408)

[**ABSTRACT** 4](#_Toc523773409)

[**ACKNOWLEDGEMENT** 5](#_Toc523773410)

[**1.** **INTRODUCTION** 7](#_Toc523773411)

[1.1. Background 7](#_Toc523773412)

[1.2. Literature Review 7](#_Toc523773413)

[1.3. Research Problem 10](#_Toc523773414)

[**2.** **OBJECTIVES** 12](#_Toc523773415)

[2.1. Main objectives 12](#_Toc523773416)

[**3.** **METHODOLOGY** 13](#_Toc523773417)

[3.1. System Description 13](#_Toc523773418)

[3.2. Feasibility Study 17](#_Toc523773419)

[3.3. Requirement Analysis and Specification 17](#_Toc523773420)

[3.4. Design 17](#_Toc523773421)

[3.5. Commercialization Aspects of the Product 17](#_Toc523773422)

[3.6. Testing & Implementation 17](#_Toc523773423)

[3.7. Research Findings 18](#_Toc523773424)

[**4.** **RESULTS AND DISCUSSION** 19](#_Toc523773425)

[4.1. Results 19](#_Toc523773426)

[4.2. Discussion 19](#_Toc523773427)

[**5.** **CONCLUSION** 20](#_Toc523773428)

[**6.** **REFERENCES** 21](#_Toc523773429)

[**7.** **APPENDICES** 22](#_Toc523773430)

[7.1. ER Diagram 22](#_Toc523773431)

[7.2. Use Case Diagram 22](#_Toc523773432)

[7.3. Use Case Scenarios 22](#_Toc523773433)

[7.4. Frontend Interfaces 22](#_Toc523773434)

# **INTRODUCTION**

### Background

Massive Open Online Courses (MOOCs) have taken a giant leap in the field of e-Learning, specifically open education. Over the years, various platforms such as edX , Coursera , Udacity , Futurelearn etc. have emerged with the intent of providing massive educational resources to its users. One of the major and fundamental component in MOOCs is the video lectures and its production style. There are standard styles, like talking head, presentation slides with voice-over, animation, screencast etc. that are most commonly used. Along with the video lectures, quizzes and learning resources are also made available to help in the learning process.

In the recent years, there has been increasing attention towards the characteristics of learners such as learning styles. Different learners have the ability to learn in different ways and hence poses their own style of learning. Because of this behavior, the learner mostly explores through different MOOCs platform to find the most optimal learning resources that best fit their needs, preferences and learning style. Until now, no system has been developed that takes learner’s learning style into consideration when recommending the best available MOOCs across different platforms.

Hence, we will be focusing on two different aspects: learning styles and MOOCs. Our sole purpose throughout the research project is to integrate both the factors together and develop an educational platform where learners from any domain can benefit from personalized learning services.

### Literature Review

* Topic Modelling and Transcript Complexity

Discovering abstract topics in MOOCs video transcripts can further help to filter the required videos. Wide range of research have also been conducted in this area. [5] has suggested three different mechanisms for automatically detecting the topic of the text by reviewing the results and evaluation’s scores. It has been observed that the proposed three approaches are applicable to find out the appropriate topic for the single document short texts and they extract the Topic (title) that can concisely convey the message of the long text. The study concluded that the most relevant, frequent and suitable words are Nouns for suggesting the topic for a newspaper article. Similarly, the research in [5] presented an automated phrase mining framework with two novel techniques: the robust positive-only distant training and the POS-guided phrasal segmentation incorporating part-of-speech (POS) tags, for the development of an automated phrase mining frame-work AutoPhrase. Text readability is another research domain incorporated with the proposed system. [6] developed a readability measurement method that estimates readability scores for texts intended for EFL learners by examining the various linguistic features of the texts: lexical, syntactic and discourse features.

### Research Problem

With the rapid advancement of technology, learning resources available over the web are more distributed than ever. Specifically, MOOCs (Massive Open Online Courses) are a relatively recent learning phenomenon with different platforms available like, edX, Coursera, Udacity, Futurelearn, Khan Academy etc. which provides large number of courses to a massive audience in different domains. MOOCs has transformed the era of learning since its introduction and one of the main reason behind it is openness. Learners coming from different background has access to massive resources at any point of time. It truly has brought a revolution in field of e-Learning.

As the popularity of MOOCs increases, more platforms are stepping up into the play where each of them has unique format of displaying the course information. This will bring much inconvenience to the users and there is a high demand of a unified platform where learners can explore the courses based on their preference. An advanced MOOC search engine that allows the users to filter through multiple fields and parameters unlike other current general search engine is a key necessity and of high demand.

To make is possible with the advanced search feature the topics modelling mechanism is being used to identify abstract topics from the courses to filter out the search results faster. Linguistic complexity calculation will also help the users to identify the courses based on their language level and filter out the best fit courses

Considering the above facts, we can conclude that there is a high demand of a unified platform for recommending MOOCs based on the learning style of user and at the same time, serve as a search engine allowing the users to filter through more advanced fields.

# **OBJECTIVES**

### Main objectives

The proposed system is a research project focused on several objectives. Upon the completion of the project, we are supposed to fulfil these research objectives. The general objectives of the research project are mentioned as follows:

* Develop a unified platform for searching MOOCs distributed on various platforms: edX, Coursera and Futurelearn using advanced filters.
* Recommend the most suitable courses from different platforms based on learners’ learning style and domain knowledge.
* Develop user-friendly interfaces so that users can easily get what they need.
* Help users identify their learning style based on Felder and Silver Learning Style Model.
* Explore and understand the application of deep learning techniques in various scenarios.

# **METHODOLOGY**

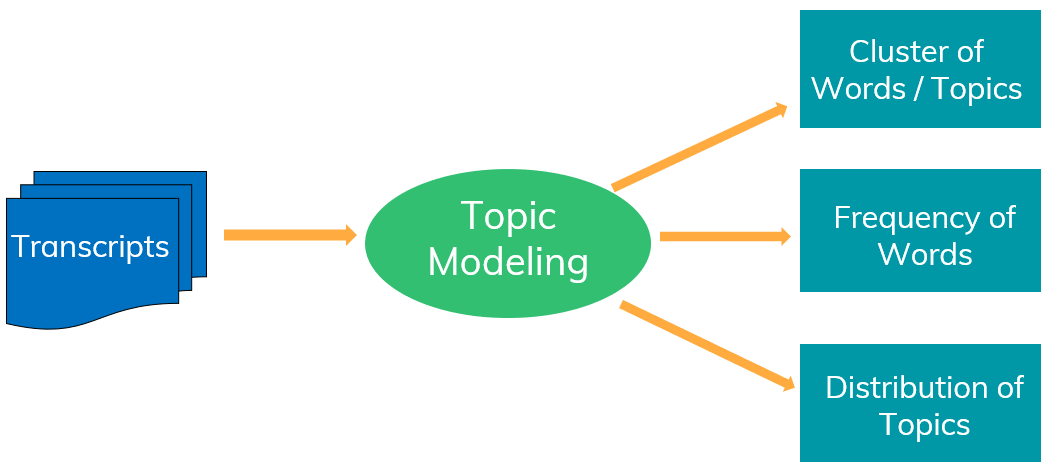
### System Description

The system is one of the first kind in the MOOCs domain, which first identifies the learner’s learning style. Out of several models defined, Felder and Silverman Learning Style Model (FSLSM) is adopted. Index of Learning Style (ILS) questionnaire is used as a mechanism to identify the learning styles. The FSLSM classifies learning styles into four dimensions and identifies two types of learners for each dimension[1]. The dimensions are perception, input, processing and understanding[1]. Video lectures of MOOCs from three platforms, edX, Coursera and Futurelearn are classified into various production styles using deep learning techniques. These classified video styles along with quizzes and resources available, MOOCs are labeled with a specific learning style, based on literature review. After the learning style has been identified, relevant MOOCs are suggested to the learner. Besides recommendation, the system can also be used as an advanced MOOC search engine by filtering using various parameters. Transcripts of video lectures, accent of speaker (English as First Language, FFL or English as Second Language, ESL), course description, provider details, ratings, comments etc. are also extracted, analyzed and stored which can be used for filtering MOOCs.

At the end of the project, we will be focusing on two measurable outcomes:

* **Accuracy:** The accuracy of the overall system can be measured by calculating what percentage of the MOOCs recommended by the system (considering the learning style) matches with the MOOCs that the learners would actually select in an open environment. In order to carry out the validation and measurement, experiments can be conducted taking students from same educational background and domain knowledge but with different learning styles. First, they will be freely allowed to select a specific course from any of three platforms. And finally, they will be given to use the system which also recommends a prioritized list of courses based on his/her learning styles. Hence, the course selected by the student can be validated with the course recommended by the system. With the process, an overall accuracy of the system can be measured.
* **Speed:** The overall speed of the system can also be measured by calculating how long does it take for MOOCs retrieval and recommendation. Our major goal will be to achieve a fast working system, hence accuracy might be tradeoff for speed.

#### Topic Modelling and Transcripts Complexity

For all the three selected MOOCs platform, transcripts are available for each video lecture. Specific topics can be extracting from the document which is useful in the scenario where the learner wants to search for MOOC with specific topics. Similarly, determining complexity of the language in transcript can be another important factor for filtering a MOOC by the learner. Phrase mining technique can be used to get short texts from the massive texts. Then for topic modeling, Nouns only approach can be used since it is the most relevant for topic suggesting as it has given a better result for the approach. Various linguistic features can be used to measure the complexity where features can be: lexical, syntactic or discourse.

The measurable outcomes for this component are:

* Accuracy: The accuracy of topic modeling can be measured by comparing topics outputted by the topic modeling algorithm with the topics actually contained in the transcript document, which is extracted manually.
* Speed: The speed of the algorithm can be measured with the amount of time taken to show the topics in the transcripts.

#### Topic modelling with Latent Dirichlet allocation

Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body for discovering the abstract “topics” that occur in a collection of documents. In MOOCs platform, transcripts are available for each video lecture. Specific topics can be extracting from the document which is useful in the scenario where the learner wants to search for MOOC with specific topics.

LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. Latent Dirichlet allocation (LDA), first introduced by Blei, Ng and Jordan in 2003[99], is one of the most popular methods in topic modeling. LDA represents topics by word probabilities. The words with highest probabilities in each topic usually give a good idea of what the topic is can word probabilities from LDA.

The transcripts of the MOOC courses will be used for Data Pre-processing and following steps are performed before applying LDA models[1]:

* Tokenization: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
* Words that have fewer than 3 characters are removed.
* All stopwords are removed. (Stop Words are words which do not contain important significance to be used in Search Queries)
* Words are lemmatized — words in third person are changed to first person and verbs in past and future tenses are changed into present.

LDA, an unsupervised generative probabilistic method for modeling a corpus, is the most commonly used topic modeling method. Given a corpus *D* consisting of *M* documents, with document *d* having N d words (*d* {1,..., M}), LDA models *D* according to the following generative process [2].

*(a)* Choose a multinomial distribution *φ**t*for topic *t* (*t* ∈{1,..., *T*}) from a Dirichlet distribution with parameter *β*.

*(b)* Choose a multinomial distribution *θ**d*for document *d* (*d* ∈{1,..., *M*}) from a Dirichlet distribution with parameter *α*.

*(c)* For a word *w**n*(*n* ∈{1,..., *N**d*}) in document *d*,

*(i)* Select a topic *z**n*from *θ**d*.

*(ii)* Select a word *w**n*from *φ**zn*.

LDA is a distinguished tool for latent topic distribution for a large corpus. Therefore, it has the ability to identify sub-topics for a technology area composed of many patents and represent each of the patents in an array of topic distributions. With LDA, the terms in the collection of documents produce a vocabulary that is then used to generate the latent topics. Documents are treated as a mixture of topics, where a topic is a probability distribution over this set of terms. Each document is then seen as a probability distribution over the set of topics. We can think of the data as coming from a generative process that is defined by the joint probability distribution over what is observed and what is hidden.

#### Linguistic Complexity with Neuro-linguistic programming

The Linguistic complexity of the transcripts is measured using three main variables[3], that are calculated separately.

1. Syntactic Complexity [4]: There are 14 different measures of syntactic complexity. Some of them are, Mean Length of Clause (MLC), Mean Length of Sentence (MLS), Mean Length of T-Unit (MLT) etc. To calculate the syntactic complexity of the transcripts, first they are fed as input which is parsed using Stanford parser. Then we count the occurrences of words, sentences, clause, t-unit etc. Then finally we calculate the complexity using those 14 measures and store the syntactic complexity indices.
2. Semantic Complexity: Semantic complexity correlates with the number of ways meaning can be derived and interpreted from an utterance. It is also associated with the types of syntactical structures necessary for it to be an intelligible utterance, and the number of different pathways meaning can be retrieved from.

### Feasibility Study

### Requirement Analysis and Specification

### Design

### Commercialization Aspects of the Product

### Testing & Implementation

#### Implementation

#### Testing

##### Unit Testing

##### Integration Testing

##### System Testing

##### Functionality Testing

##### Performance Testing

##### Security and Portability Testing

### Research Findings

# **RESULTS AND DISCUSSION**

### Results

### Discussion

# **CONCLUSION**

# **REFERENCES**

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research 3 (2003) 993-1022

[2] “Topic Modeling and Latent Dirichlet Allocation (LDA) in Python” [Online]. Available: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 [Accessed: 12-June-2018]

[3] Gabriele Pallotti, “A simple view of linguistic complexity,” University of Modena and Reggio Emilia, Italy, vol. 31, pp. 117-134, July 2014.

[4] Xiaofei Lu, “Automatic analysis of syntactic complexity in second language writing”, International Journal of Corpus Linguistics, vol. 15, pp. 474 –496, 2010.

[5] K. Kotani and T. Yoshimi, “A Machine Learning Approach to Measurement of Text Readability for EFL Learners Using Various Linguistic Features,” US-China Educ. Rev. B, vol. 6, pp. 767–777, 2011..

[6] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han, “Automated Phrase Mining from Massive Text Corpora,” IEEE Trans. Knowl. Data Eng., pp. 1–1, 2018. [7] A. Sajid, S. Jan, and I. A. Shah, “Automatic Topic Modeling for Single Document Short Texts,” in 2017 International Conference on Frontiers of Information Technology (FIT), 2017, pp. 70–75.

[8] “Complete Guide to Topic Modeling.” [Online]. Available: https://nlpforhackers.io/topicmodeling/. [Accessed: 12-May-2018].

[9] “What is a good explanation of Latent Dirichlet Allocation?.” [Online]. Available: https://www.quora.com/What-is-a-good-explanation-of-Latent-Dirichlet-Allocation. [Accessed: 12-May-2018].

[10] “Analysis of Linguistic Complexity in Professional and Citizen Media” [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1461670X.2017.1305285. [Accessed: 12-May-2018].

# **APPENDICES**

### ER Diagram

### Use Case Diagram

### Use Case Scenarios

### Frontend Interfaces