ISCR - Integrative Intelligent Semantically Driven Scheme for Online Course Recommendation

Harshal Sharma[1,†], Gerard Deepak[2,†]

[1]Birla Institute of Technology and Science, Pilani , Vidya Vihar, Pilani, Rajasthan, India
[2]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

Abstract

The easy access of Internet has caused an exponential increase in the availability of courses on a variety of platforms. The increasing popularity and demand of these online courses subsequently create a need for better and more efficient course recommendation systems. The development of WEB 3.0 has also added to this need as the current course recommendation systems are not semantically compliant. The system is user knowledge-centric. It uses Decision Trees to classify the dataset. The semantic similarities have been computed with the help of normalized compression distance, normalized google distance and Gini-index. The performance has been evaluated and compared with other approaches like OPCR, CRFL, CRQCA and hierarchical clustering plus Jaccard similarity. A clear observation has been made that the proposed, knowledge centric ISCR course recommendation system performs superiorly and attains an average Recall, and F-measure of 98.48

Keywords

Course Recommendation, Decision Trees, RDF Synthesis, Semantically Driven, Semantic Similarity,

1. Introduction

The advancement of the Internet has had a significant impact on all aspects of human lives, especially in the field of education. With the increasing accessibility of the Internet, education is shifting towards an online mode. Many platforms like Coursera, Udemy, Alison, edX, etc., provide a multitude of online courses and have a vast user base. This has prompted an exceptional expansion in the quantity of these courses, making it progressively troublesome to express the user's individual needs accurately. With this arises a need for a course recommendation system.

A course recommendation system analyses the user's history and preferences to present a list of courses that they might find interesting. E-learning has significantly increased information consumption, and such systems are fairly accurate in suggesting relevant courses to users. Such systems are beneficial as people often don't exactly know what they want and might find interesting. Besides courses recommendation these recommendation systems are used by many organizations for a variety of purposes. For example, the main page or the feed

International Workshop on Semantic IoT (IWSIoT-2022) and Multilingual Semantic Web (IWMSW-2022) and Deep Learning for Question Answering (IWDLQ-2022)

*Corresponding author.

gerard.deepak.christuni@gmail.com (G. Deepak)
0000-0002-0877-7063 (G. Deepak)

on apps like Facebook, Instagram, Google, Amazon etc. are different for different users as the data presented to the user on these pages is according to the user's preferences. This idea has been in use for many years, and there have been many advancements in this field. The current models work on content-based filtering, collaborative filtering, and knowledge graphs. But there is always scope for improvement, especially with the developments in Web 3.0.

Motivation: Web 3.0 or frequently referred to as the semantic web is the third generation in the evolution of the World Wide Web and of internet services like websites and other applications. Web 3.0 focuses on making the data available on the Internet machine readable so as to make a semantically compliant and data-driven Web. The advancement of technology and the tremendous increase in the amount of data generated go hand in hand. Therefore, we need a robust recommendation system for the increasing number of courses and users. The semantic web is a knowledge-centric and data-driven web. Furthermore, the current recommendation models are not semantic web compliant as it is highly cohesive. This makes the development of better and more accurate models crucial.

Contribution: A semantically driven; knowledge centric system is proposed for online course recommendation systems. The user generates an input query which is then subject to pre-processing. Query pre-processing involves lemmatization, tokenization, stop word removal, and NER (Named Entity Recognition). Furthermore, on obtaining the individual query words, the RDF (Resource Description Framework) synthesis takes place using the RDF distiller. Entity enrichment has been done using the Wikidata API. Classification of the dataset is done using decision Trees. Semantic similarity has been computed using normalized compression distance, normalized google distance and Gini index.

Organization: The publication's structure is as follows: Section 2 comprises is comprised of the related works around the subject of research. Section 3 contains of the architecture for the proposed framework. Section 4 consists of the implementation of the proposed model along with the performance evaluation and results. Section 5 consists of the conclusion.

2. Related Works

Zhang et al., [1] described a model for a course recommendation system for MOOCs [Massive Open Online Courses]. MCRS is based on distributed computational framework. Its basic algorithm is distributed association rule mining, which is rooted on improvement of Apriori algorithm. Ibrahim et al., [2] described system for recommending courses which are personalized and ontology-based and hence the name OPCR. It combines content-based filtering with collaborative-based filtering. Furthermore, it uses an ontology mapping technique. Lin et al., [3] proposed an adaptive course recommendation in MOOCs. They presented Dynamic Attention and hierarchical Reinforcement Learning [DARL] mechanism which improves the course recommendation in terms of adaptivity. They presented a dynamic attention mechanism that can be used to monitor alterations in user preferences.

Pardos et al., [4] designed for serendipity in course recommendation in a university. One of their models was based on descriptions mentioned in course catalogs and another set by course enrollment histories. Zhang et al., [5] proposed a new approach of learning for recommendation of courses in MOOCs using Hierarchical Reinforcement Learning. The attention mechanism performs poorly because the additional courses the user is interested in dilute the effects of contributory courses. To address this, they presented hierarchical reinforcement learning algorithm to revisit and make changes to the user's profile and based on this, enhance the model. Agorista et al., [6] proposed an approach towards Course Recommendation using Markov Chain Framework for. It uses a random walk-based methodology to record the connections between several courses in chronological order.

Jing et al., [7] described course recommendation in MOOCs. The course recommendation algorithm uses collaborative filtering based on user's interests, demographic profiles and course requirements relations. Bhumichitr et al., [8] presented a recommender system for university elective courses. It utilizes collaborative based recommendation using Alternating Least Square [ALS] and Pearson Correlation Coefficient. Then, based on a dataset of students' academic records, it compares their performance. Chang et al., [9] described a hybrid system for course recommendations. It integrates collaborative filtering and artificial immune systems. The testing parameters used are confusion matrix analysis and average error. Ibrahim et al., [10] presented a novel approach and built an ontology-based personalised course recommendation framework, an ontology-based hybrid-filtering system [OPCR]. Sulaiman et al., [11] proposed employing fuzzy logic to create a course recommendation system. using the Mamdani fuzzy inference system as a foundation, it applied the fuzzy rules technique to determine each related student's skill and interest level [CRFL]. Gulzar et al., [12] presented a model for recommending courses to scholars depending on their requirements and domains of interest using N-gram classification technique [CRQCA]. In [14-18] several ontological frameworks which also include knowledge graphs and a conceptual knowledge model in support of the literature of the proposed framework have been discussed.

3. Proposed System Architecture

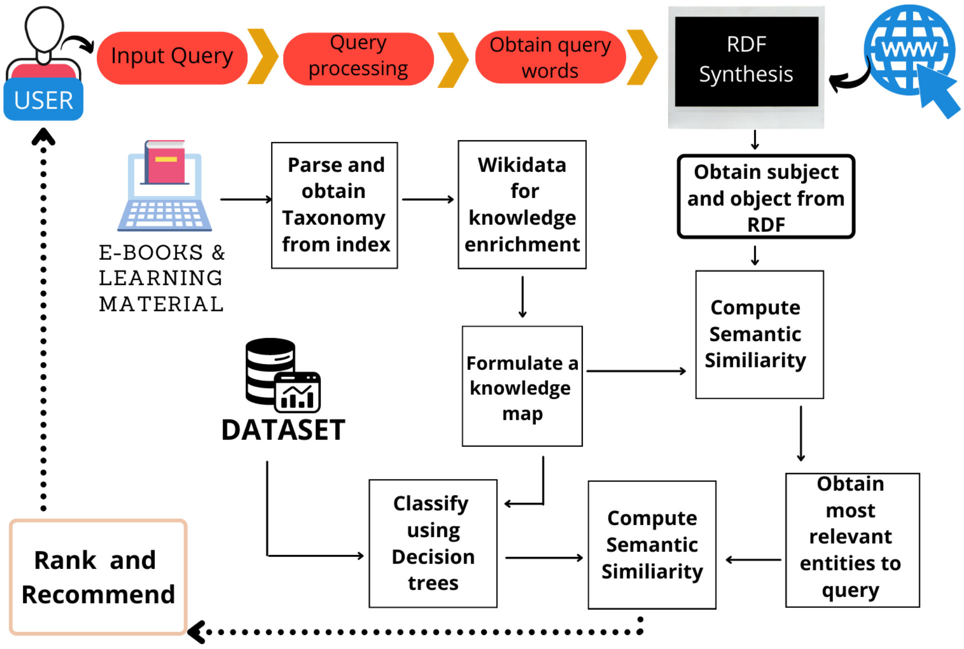
USER
Input Query
Query processing
Obtain query words
RDF Synthesis
WWW
E-BOOKS & LEARNING MATERIAL
Parse and obtain Taxonomy from index
Wikidata for knowledge enrichment
Obtain subject and object from RDF
Compute Semantic Similiarity
Formulate a knowledge map
DATASET
Classify using Decision trees
Compute Semantic Similiarity
Obtain most relevant entities to query
Rank and Recommend

Figure 1: Proposed ISCR model

Fig. 1 depicts the proposed system architecture for knowledge driven course recommendation model. The user's input query, which goes through pre-processing, serves as the model's driving force. Lemmatization, tokenization, stop word removal, and NER (Named Entity Recognition) are all components of query pre-processing . Tokenization involves separating individual words from sentences and obtaining the individual tokens for the separated words. Lemmatization has been done to derive the base form of the word from its inflectional form. Stop word removal focuses on eliminating stop words like of, the, and, other articles. Named entity recognition involves marking the entities by recognizing them. Standard Python Natural Language Toolkit (NLTK) based libraries are used for lemmatization, stop word removal, tokenization, and NER.

Furthermore, on obtaining the individual query words, the RDF (Resource Description Framework) synthesis takes places. RDF synthesis takes places using the RDF distiller, which is a triadic format of knowledge obtained from the World-Wide Web (WWW). RDF consists of subject, object and predicate. However, the RDF is not used as it is. Only the subject and the object are obtained. The reason pertaining to this is that the subject and the object both are either in the form of a term or sentence while the predicate can be a URL or a sentence or a word. Owing to the heterogeneity of the predicate, only the subject and the object are used. Nonetheless, subject and object when occur together can imbibe stronger level of semantics. As a result, RDF subject and object alone are used.

E-books and other learning materials are stored in a repository. The learning material involves e-books pertaining to subjects such as, history, psychology, civics, journalism, environmental science, ecology etc. All the e-books constituting these domains are crawled and stored in a localized repository which are then passed to extract the taxonomy from the indexes of the e-books. These indexes are formalized domain-wise and are subjected to the Wikidata API for further knowledge enrichment. In other words, entity enrichment takes place by subjecting the indexes to the Wikidata API and a knowledge map for individual domains or subjects is formalized. Wikidata is an open and free, multilingual knowledge graph built and managed by the Wikimedia Foundation, which is readable and editable by both humans and machines. For many initiatives, Wikidata serves as a central repository for their structured data. Upon obtaining the individual knowledge map for the domains, the dataset is classified using the features obtained from the domain-based knowledge maps. The classification is done using Decision Trees.

Classification Decision Trees
These types of trees classify the data using yes or no questions. For example: - if we want to classify heart disease patients, we can do so by judging the sample space, here the people we want to classify based on whether or not they suffer from chest pain. If a person answers yes, then they are classified into having heart disease, if they answer No, then they are classified into not having heart disease. The same can be done by using various other factors that influence the subject, here heart disease.

Regression Trees
In contrast to classification trees these types of trees use numeric values to classify a dataset. For example: - testing the dosage of a new drug developed to cure coronavirus disease. Given the data for the percentage effectiveness of the drug upon taking a certain dose, we can make a decision (regression) tree by dividing the given data into categories of dosage. Then, we use these numeric values to find the percentage effectiveness of the drug at any given dosage, like is the dosage greater than 10mg? If yes, then is it lesser than 15mg, and finally we answer by taking the mean of the percentage effectiveness of the drug in the given range of doses taken.

An ordinary decision tree is depicted inverted.. The top of the decision tree is called the Root Node. Every factor or question that splits the data is called an Internal Node or Node. Nodes which are not divided further are called Leaf nodes or leaves. Decision Trees are a simple and effective algorithm. Often combined version, using both forms of decision trees are used, so as to increase the accuracy.

Based on the domain to which the query belongs to, the query enriched RDF subject object entities have been used to compute the semantic similarity with the domain specific knowledge map. The semantic similarity has been computed using the normalized compression distance, normalized google distance and Gini index is also computed. Normalized compression distance and normalized google distance are used with a threshold of 0.5.

$$NCD(X,\ Y) = \frac{|C(XY)|\ - \min(|C(X)|,\ \ |C(Y)|)}{\max(|C(X)|,\ \ |C(Y)|)} \tag{1}$$

Equation (1) defines the Normalized Compression Distance (NCD). E (X, Y) is the information distance between two string X and Y, which is equal to the duration of the shortest program. which converts X into Y and vice versa in some fixed programming language. E (X, Y) is roughly equivalent to C(XY), which is the outcome of compressing a file made up of X concatenated with Y using a particular compression method C.

$$NGD(x,\ y) = \frac{max[\log f(x), \log f(y)] - \log f(x,\ y)}{\log N - \min[\log f(x),\ \log f(y)]} \tag{2}$$

Equation (2) defines the Normalized Google Distance (NGD), where N is the number of singleton search items on the pages multiplied by the total number of web pages searched by Google. The number of results for the search terms x and y are represented by f(x) and f(y), respectively. The number of web pages where both x and y are present at the same time is represented by f(x, y).It is desired to maximize the number of instances at this phase, but it is also crucial to maintain relevancy. Thus, to maintain relevancy and maximize the number of instances, we enhance the space where the semantic similarity is computed using these two measures by setting threshold of 0.5. Along with this Gini index is also computed with a step deviation of 0.2. the reason behind using two semantic similarity measures and Gini index is to enhance the number of relevant entities in the term space, and moreover, to keep the rate of relevancy high.

Furthermore, we are obtaining the most relevant entities to the query and these entities are again used to calculate semantic similarity with instances classified with the decision trees. At this stage only the normalized google distance with a threshold of 0.75 is used. The reason behind this is that the relevancy has been computed already and has reached a refined stage. Hence, a single semantic similarity measure with a high threshold is sufficient. Moreover, this will also reduce the computational complexity and make the process of recommending courses computationally inexpensive.

Finally, the courses are ranked and recommended in the increasing order of the normalized google distance to the user. Also, further clicks from the user are recorded, and this process continues recursively until the user is satisfied and if there are no further user clicks then the entire recommendation stops.

4. Implementation and Performance Evaluation and Results

The experimentations on the proposed Integrative Semantically driven Course Recommendation (ISCR) architecture were executed on Kaggle's Coursera Course dataset. This dataset contains data of 890 courses. It also contains detailed description of the courses in 6 columns like rating, difficulty, certificate, number of students enrolled etc. This is primarily the main dataset on which the experimentations have been done. However, to enhance the relevance with respect to domain, several courses pertaining to history, psychology, civics, journalism, environmental science, ecology etc. were crawled. It is seen that around 142 courses regarding history, 14 courses regarding civics, 78 courses regarding psychology, 24 courses regarding environmental science, 8 courses regarding ecology were crawled and incorporated in the framework. Out of the added the mentioned numbers were true positives and nearly an equal number of false positives were also found. Around 50 percent of the courses added were labelled as neutral as they were neither true positives nor false positives but were relevant to subjects. As a result, these courses were annotated and categorized according to the Coursera Course Dataset.

The performance of the proposed ISCR framework for course recommendation which is semantically inclined and integrative in nature is compared with the help of precision recall accuracy, f measure percentages and false discovery rate (FDR) as potential metrics. The reason for using precision, recall accuracy and f-measure percentages is because it quantifies the percentage of the relevance of results and false discovery rate is used as the preferred metric. It also computes the number of false positives captured by the framework. Furthermore, standard formulation for precision recall, accuracy, f measure and false discovery rate have been used. The experimentations were conducted for 4471 queries on the customized dataset described later below. To compare the performance of the proposed ISCR, it is base-lined with OPCR, CRFL, CRQCA models. It is also benchmarked with the combination of hierarchical clustering and Jaccard's similarity.

Table 1
Comparison of Performance of the proposed ISCR with other approaches

Model	Avg Precision%	Avg Recall %	Avg Accuracy %	Avg F-Measure %	FDR
OPCR [2]	90.23	93.18	91.71	91.68	0.10
CRFL [11]	90.37	94.69	92.53	92.47	0.10
CRQCA [12]	87.63	90.22	88.93	88.90	0.12
HC+JS	85.31	88.12	86.72	86.69	0.15
Proposed ISCR	96.07	98.48	97.275	97.26	0.04

From Table 1 it is indicated where the proposed ISCR yields the highest precision of 96.07%, highest average recall percentage of 98.48%, highest average accuracy of 97.275%, highest average f measure percentage of 97.26%, with the lowest FDR of 0.04%. It is clearly observable that OPCR yields an average precision of 90.23%, an average recall of 93.18%, an average accuracy of 91.705%, an average f measure of 91.68% with an FDR of 1-0.90 = 0.1. The CRFL model yields an average precision of 90.37%, an average recall of 94.69%, an average accuracy of 92.53%, average f measure of 92.47% with an FDR of 0.1. The CRQCA model yields an average precision of 87.63%, an average recall of 90.22%, an average accuracy of 88.925%, an average f measure of 88.9% with an FDR of 0.13. The HC+JS model, the combination of hierarchical clustering and Jaccard's similarity yields an average precision of 85.31%, an average recall of 88.12%, an average accuracy of 86.715%, an average f measure of 86.69% with an FDR of 0.15. The reason why the proposed ISCR model yields the highest average precision percentage, highest average recall percentage, highest average accuracy, highest average F-measure percentage and lowest FDR is because it is integrative and hybridized in nature, it is semantically driven, it is driven by RDF. The model ensures RDF synthesis for the query terms and perfect taxonomy, which is domain specific is used and knowledge is enriched using the Wikidata API. Since it is RDF driven, the lateral semantics is very high because the subject and object co-occurrent is used together. Most importantly the dataset is classified using decision trees from which the features are derived from an initially formalized knowledge graph which is conceived based on the taxonomy as well as the RDF synthesizer. Most importantly the relevance computation mechanism is very strong in terms of semantic similarity computation using normalized compression distance, normalized google distance and the Gini Index. Owing to all these factors and since it is driven by the query and RDF synthesis is carried out along with the incorporation of a perfect taxonomy from the domains based on the standard E-books, the proposed model is much better than the baseline models. Moreover, the semantic similarity computation scheme which is to compute the relevance is very stringent and strong as it uses NCP, NGP and the Gini-index.

The reason why the OPCR model lacks, although it uses a perfect ontological model is that ontologies are shallow in nature and apart from being shallow, they are static. So hierarchical ontology semantic similarity model alone is used along with collaborative and content-based filtering. The combination of contest-based filtering, collaborative filtering along with the hierarchical ontology similarity also become insufficient because the ontology itself is static. So even though the schemes for relevance computation are slightly stronger, owing to shallow nature of the ontology this model does not perform as expected. The CRFL model which uses the fuzzy logic approach for course recommendation uses fuzzy rules and techniques for fuzzification. Although it considers rules, it does not work well and moreover there is an absence of ontologies which is static knowledge or auxiliary knowledge into the model. So as a result, this model also lacks and fuzzy based computation is always approximate computation and is not concrete and the relevance computation mechanism is not strong. This is the reason why the CRFL model lacks compared to the proposed model. The CRQCA model also lacks as it uses a query classification approach but however, the N-gram is alone used for classification and there is some amount of auxiliary knowledge in the model but the entire learning takes place based on the dataset depending on the domain of the courses. Due to these reasons N-gram becomes quite shallow and learning happens from the dataset which creates a lag in the performance of the CRQCA model. Finally, the hybridization of hierarchical clustering with Jaccard's similarity also does not perform well. Although hierarchical clustering ensures good strategy clustering and Jaccard's similarity provides high relevance computation mechanism, the model lacks in its accumulation of auxiliary knowledge. As a result, the hierarchical clustering plus Jaccard's similarity model lacks when compared to the proposed model and even compared to other models it lacks abruptly.

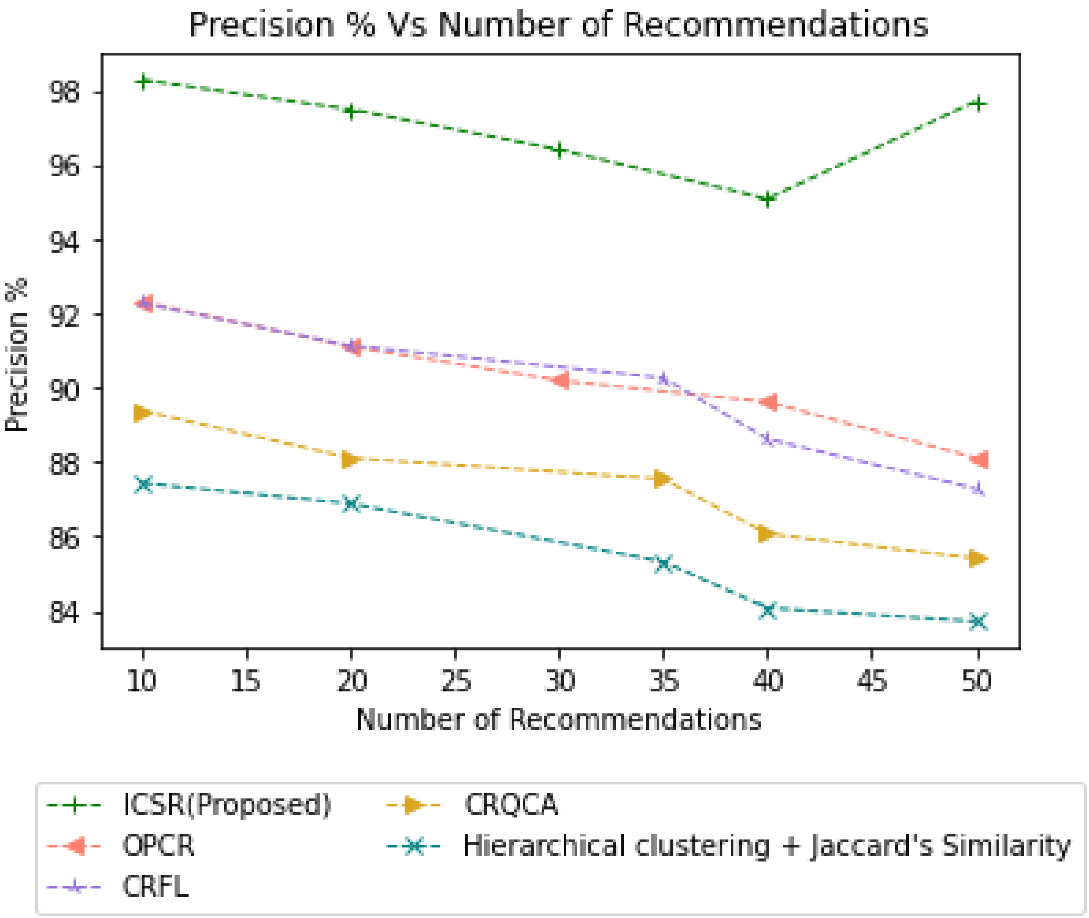

Figure 1: Precision % vs Number of Recommendations

Fig. 2 represents precision % vs number of recommendations curve. From the figure we can infer that ISCR yields the highest precision in the hierarchy of precision with the number of recommendations distribution curve. The next immediate position is taken by the OPCR model. The next position is occupied by CRFL, followed by CRQCA and the lower most position is occupied by the hierarchical clustering plus Jaccard's similarity.

5. Conclusion

A semantically driven and knowledge centric model is proposed to recommend online courses. The model is based on RDF synthesis and also uses Decision Tress and various methods for computing semantic similarity. It also uses Wikidata API for knowledge enrichment. The proposed ISCR model achieved an average accuracy of 97.275% with an average precision of 96.07%. It is evident from the results that the proposed ISCR model outperforms the OPCR, CRFL, CRQCA and Hierarchical Clustering plus Jaccard Similarity models in every aspect. This makes it a better, efficient and semantically compliant model for online course recommendation.

References

[1] Zhang, H., Huang, T., Lv, Z., Liu, S., Zhou, Z. (2018). MCRS: A course recommendation system for MOOCs. Multimedia Tools and Applications, 77(6), 7051-7069.
[2] Ibrahim, M. E., Yang, Y., Ndzi, D. L., Yang, G., Al-Maliki, M. (2018). Ontology-based personalized course recommendation framework. IEEE Access, 7, 5180-5199.
[3] Lin, Y., Feng, S., Lin, F., Zeng, W., Liu, Y., Wu, P. (2021). Adaptive course recommendation in MOOCs. Knowledge-Based Systems, 224, 107085.
[4] Pardos, Z. A., Jiang, W. (2020, March). Designing for serendipity in a university course recommendation system. In Proceedings of the tenth international conference on learning analytics knowledge (pp. 350-359).
[5] Zhang, J., Hao, B., Chen, B., Li, C., Chen, H., Sun, J. (2019, July). Hierarchical reinforcement learning for course recommendation in MOOCs. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 435-442).
[6] Polyzou, A., Nikolakopoulos, A. N., Karypis, G. (2019). Scholars Walk: A Markov Chain Framework for Course Recommendation. International Educational Data Mining Society.
[7] Jing, X., Tang, J. (2017, August). Guess you like: course recommendation in MOOCs. In Proceedings of the international conference on web intelligence (pp. 783-789).
[8] Bhumichitr, K., Channarukul, S., Saejiem, N., Jiamthapthaksin, R., Nongpong, K. (2017, July). Recommender Systems for university elective course recommendation. In 2017 14th international joint conference on computer science and software engineering (JCSSE) (pp. 1-5). IEEE.
[9] Chang, P. C., Lin, C. H., Chen, M. H. (2016). A hybrid course recommendation system by integrating collaborative filtering and artificial immune systems. Algorithms, 9(3), 47.
[10] M. E. Ibrahim, Y. Yang, D. L. Ndzi, G. Yang and M. Al-Maliki, "Ontology-Based Personalized Course Recommendation Framework," in IEEE Access, vol. 7, pp. 5180-5199, 2019, doi: 10.1109/ACCESS.2018.2889635.
[11] Sulaiman, M. S., Tamizi, A. A., Shamsudin, M. R., Azmi, A. (2020). Course recommendation system using fuzzy logic approach. Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), 17(1), 365-371.
[12] Gulzar, Z., Leema, A. A. (2018). Course recommendation based on query classification approach. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 13(3), 69-83.
[13] M. Anirudh, Gerard Deepak, A. Santhanavijayan. "Chapter 31 SIITR: A Semantic Infused Intelligent Approach for Tag Recommendation", Springer Science and Business Media LLC, 2022
[14] Gupta, S., Tiwari, S., Ortiz-Rodriguez, F., Panchal, R. (2021). KG4ASTRA: question answering over Indian missiles knowledge graph. Soft Computing, 25(22), 13841-13855.
[15] Tiwari, S., Abraham, A. (2020). Semantic assessment of smart healthcare ontology. International Journal of Web Information Systems, 16(4), 475-491.
[16] Gaurav, D., Tiwari, S. M., Goyal, A., Gandhi, N., Abraham, A. (2020). Machine intelligence-based algorithms for spam filtering on document labeling. Soft Computing, 24(13), 9625-9638.
[17] Rai, C., Sivastava, A., Tiwari, S., Abhishek, K. (2021). Towards a Conceptual Modelling of Ontologies. Emerging Technologies in Data Mining and Information Security: Proceedings of

IEMIS 2020, Volume 1, 1286, 39.

[18] Tiwari, S., Al-Aswadi, F. N., Gaurav, D. (2021). Recent trends in knowledge graphs: theory and practice. Soft Computing, 25(13), 8337-8355.