BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **23**, No 2
Sofia • 2023 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2023-0019

Novel Approaches for Searching and Recommending Learning Resources

Tran Thanh Dien[1], *Nguyen Thanh-Hai*[1], *Nguyen Thai-Nghe*[2]

[1]*College of Information and Communication Technology, Can Tho University, Vietnam.*
[2]*College of Information and Communication Technology, Can Tho University, Vietnam (corresponding author)*
E-mails: thanhdien@ctu.edu.vn nthai@cit.ctu.edu.vn ntnghe@cit.ctu.edu.vn

Abstract**:** *This study proposes models for searching and recommending learning resources to meet the needs of learners, helping to achieve better student performance results. The study suggests a general architecture for searching and recommending learning resources. It specifically proposes* (1) *the model of learning resource classification based on deep learning techniques such as MLP;* (2) *the approach for searching learning resources based on document similarity;* (3) *the model to predict learning performance using deep learning techniques including learning performance prediction model on all student data using CNN, another model on ability group using MLP, and the other model on per student using LSTM;* (4) *the learning resource recommendation model using deep matrix factorization. Experimental results show that the proposed models are feasible for the classification, search, ranking prediction, and recommendation of learning resources in higher education institutions.*

Keywords**:** *Learning resources; learning performance prediction; learning resources recommendation; deep learning; machine learning.*

1. Introduction

Open learning has become an innovation movement in education and has been constantly developing. The term open learning generally refers to activities that enhance learning opportunities within formal education systems or extend learning opportunities outside formal education systems [1]. Open learning includes but is not limited to, classroom instruction, interactive learning approaches, the culture and ecology of the learning community, and the development and use of learning resources [2]. Open educational resources are integral to open learning. They are learning, teaching, and research materials in any publicly available or copyrighted format and medium that have been released under an open license, allowing access, use, or reuse, change, and completely free and legal re-share content [3]. Learning resources are educational resources that are developed and provided for the teaching and learning process to meet learning goals [4]. Learning resources can be provided through the systems such as e-Learning systems, curriculum, and lecture management systems, education management systems, publishing management systems, etc. Learning resource management systems have their characteristics, but the common purpose is to provide web-based features to support the teaching and learning process of institutions or self-learners needs [5-7].

With the rapid development of information technology, the demand for online learning is raising. In addition, travel is limited due to the pandemic and other issues, thereby increasing the demand for online learning and the use of materials for online teaching and learning. As the demand for online learning increases, the demand for searching learning resources increases. Therefore, it is necessary to have more effective methods of searching for learning resources as well as recommending learning resources that are suitable for learners' needs. Although there are related studies on searching and recommending learning resources, new approaches to searching for learning resources and meeting better the needs of learners, should be proposed. However, learning resources are mainly in formats of .doc, .pdf, etc., so it is necessary to solve the problem of searching unstructured documents. On the other hand, it is necessary to have effective search methods because learning resources are increasingly diverse in many different fields (or topics). For instance, classification is suggested to determine the field of the query, then search on the corresponding field instead of the whole data. Another problem is that semantics needs paying attention to make the search process more effective. In addition, it is necessary to have methods of rating prediction and recommendation of learning resources suitable for each learner.

In this study, state-of-the-art approaches are proposed to search for and recommend learning resources meeting their needs and capacities. The main contributions of the study are to propose a general architecture for searching and recommending learning resources. This approach is then divided into smaller sub-systems for (1) classifying learning resources based on deep learning techniques, (2) semantic-based searching learning resources through ensemble the similarities of cosine and word order, (3) predicting student performance with various models based on deep learning techniques, and (4) recommending learning resources based on deep matrix factorization that extends from standard matrix factorization.

In the remainder of this study, Section 2 presents some state-of-the-art related to document classification, document search problems, rating prediction, and recommendation. The proposed approaches and models for classifying, searching documents, and recommending learning resources used in this study are briefly described in Section 3. The methods and the experimental results of the proposed methods are presented in Section 4. Section 5 consists of the conclusions.

2. Related work

In this section, we summarize the related studies to the issues of document classification and search, ranking prediction, and learning resource recommendations that have been mentioned in previous works.

In learning resource search systems, especially large-scale resources, the first stage of the search progress is to process the query to determine a topic, and then search on such topic. Therefore, query classification plays an important role in narrowing the search space, increasing speed, and improving the accuracy of search results [8-11]. The main purpose of the search system is to provide the learning resources as desired by users from the vast search space. Normally, the search systems calculate the similarity between the search query and the learning resources (or documents), thereby finding a list of documents sorted in descending order of similarity. However, in order to limit the search space and make the search process faster and more accurate, text classification in general, and query classification, in particular, is a very important task to assign labels to the taxonomy set of given topics [12].

There are many studies on query classification focusing on the regular expression approach that relies on hand-written grammar rules to determine the class of the input question [13]. With this approach, previous studies have proposed a way to represent constrained text meanings, along with a flexible strategy to match queries with searched text passages based on semantic similarity and weight relationships between words. This approach has achieved certain successes but still has many limitations [11]. Modeling for this method is time-consuming and labor-intensive, requiring the cooperation of experts in the field of linguistics when building query patterns. In addition, the handwritten grammar rules and the grammar of each type of query are not flexible. When a new query appears, it is necessary to be provided with new rules to handle. The problem of grammatical ambiguity is difficult to deal with, depending on the characteristics of each language. Another problem is that when the answer set is expanded or changed, it requires a complete rewrite of the previous rules, so the system is very difficult to scale. A new approach to data classification that is widely used by researchers currently is based on machine learning techniques. For instance, authors in [14] use the SVM algorithm to solve the text classification problem and have compared its performance with the decision tree algorithm. The results show that classification with SVM is actually better than classification with a decision tree. In addition, the use of the single value analysis technique SVD (Singular Value Decomposition) to analyze and reduce the number of dimensions of the features has improved the classification efficiency with SVM. Another study builds word splitting module according to the N-gram model, then model the text using TF*IDF technique [15]. With the data set modelled into vectors, the author conducts classification based on the Naïve Bayes algorithm. The classification results are quite satisfactory, but this study has not compared the Naïve Bayes method with other classification methods. Besides the commonly used supervised and unsupervised learning techniques, reinforcement learning has also recently been used in text classification.

Reinforcement Learning is one of the most promising approaches to data-driven decision-making for improving student learning in interactive e-Learning systems. Reinforcement learning is one of the three learning techniques of machine learning, which helps determine behavior based on context to achieve the most benefit (maximizing the performance). Research results show that reinforcement learning techniques for text classification are equally effective as supervised and unsupervised learning [16, 17].

Approaches based on deep learning techniques are also implemented in many studies. A group of authors has proposed three basic architectures of deep learning models for text classification, including Deep Belief Neural Network (DBN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) [18]. The study has concluded that deep learning models are potential techniques that can be used for text classification. However, it depends on the data set to decide which technique to use for a specific classification model. In addition, a comparison between traditional machine learning techniques and deep learning techniques is needed to recommend which technique is suitable for the actual data.

Document search is essentially checking the similarity of documents to recommend suitable documents. Therefore, measuring text similarity between words, sentences, paragraphs, and texts plays an important role in text-related research and applications such as information retrieval, text classification, etc. There are many studies on text similarity, which have been systematized into three main methods, which include string-based, corpus-based, and knowledge-based [19]. The string-based similarity is used to calculate lexical similarity, while corpus-based similarity and knowledge-based similarity are used for semantic similarity.

A proposed algorithm calculates text similarity based on the combination of semantic information of sentences and word order in sentences [20]. First, the semantic similarity between two sentences determined by the lexical structure is calculated. Then, the similarity of word order due to the position of the word in the sentence is also calculated. The authors have combined these two similarities to calculate the sentence similarity, thereby calculating the text similarity. Experiments show that this algorithm can be applied in the conversation processing system quite effectively. However, this algorithm only stops at the English language. An algorithm to measure the similarity of sentences is proposed based on measuring similarity in terms of semantics and syntax of sentences, using the vector space model [21]. There are two relationships in this algorithm including the relationship between verbs and sentence pairs and the relationship between nouns and sentence pairs. One advantage of this method is that it can be used for variable-length sentences. Another study proposes to check text similarity based on semantics by using synonyms to replace the original words [22]. This study has pre-processed the words by using the word separation method and removing the stop words and then checking with the data set to detect semantic similarity through the WordNet dictionary. Other researchers [23] have proposed a method to measure semantic similarity between documents by mapping keywords such as verbs, adverbs, and adjectives to nouns, followed then by finding similarities between the mapped words. The experimental results show that the proposed algorithm gives fairly accurate results in detecting semantic similarity between documents. A similarity-checking technique based on semantic knowledge has been proposed [24]. This technique analyzes and compares the text based on the semantic allocation for each word or term in the sentence. Semantic knowledge generates semantic arguments for each sentence. The experimental results on the dataset show a significant increase, surpassing previous plagiarism detection methods in terms of precision and recall.

The above studies show that most researchers believe that the similarity in the semantic representation of sentences and the word order in sentences determines the similarity of sentences and texts [25]. Inheriting these studies can be applied to solving the problem of document search, specifically searching for learning resources, based on text similarity. In particular, for Vietnamese documents, it is necessary to perform pre-processing techniques such as word separation, stop word removal, etc., especially query classification before searching learning resources so that the search process is faster and more efficient.

Predicting learning performance becomes an important need for universities to support learners or students to achieve the best academic results. The results of many researches show that there are several different techniques applied to predicting learning performance such as artificial intelligence, machine learning, collaborative filtering, and artificial neural networks [26]. Predicting learning performance is an important task in educational data mining. From the idea that learners' knowledge can be improved and accumulated over time, an approach using a three-dimensional matrix decomposition technique including learners, subjects, and time factors is proposed to predict student learning performance [27]. With this approach, authors have personalized predictions for each specific learner. Experimental results on large data sets show that combining matrix decomposition techniques for prediction is an effective approach.

Most of the reviews have biases about users and items, which means some users are easy-going or grumpy. Sometimes, some items are highly rated by users because they have followed others. Therefore, the authors [28] have developed a recommender system using Biased Matrix Factorization (BMF) to predict student learning performance, thereby helping students choose more appropriate subjects. Experimental results using the open-source of MyMediaLite show that the BMF technique gives improved results compared to the standard matrix decomposition technique by solving the bias issue. The ability to combine prediction methods is also used by researchers. A research team has built a model to predict student learning performance based on a combination of the Taylor approximation method with gray models to obtain the most optimal predictive values [29]. The research results help teachers and educational administrators with appropriate solutions to improve the learning performance of students. Meanwhile, other authors use collaborative filtering techniques, standard matrix decomposition techniques, and restricted Boltzmann machine techniques to systematically analyze the collected data from a university [30]. The results show that the restricted Boltzmann machine's techniques predict students’ learning performance better than the other techniques.

Collaborative filtering techniques are often used in recommender systems due to their simplicity. However, when the data is sparse, this imposes a limitation of the algorithm’s effectiveness. Therefore, models that combine collaborative filtering algorithms with deep learning techniques are of more interest. A study has proposed a model based on a quadratic polynomial regression model to obtain more accurate latent features by improving the traditional matrix decomposition algorithm [31].

Then, the latent features are the input data of the deep neural network model. Experiments show that this model improves quite well prediction efficiency compared to the model using the standard matrix decomposition technique. Some other approaches that combine the collaborative filtering model with deep learning are also mentioned [32]. With this approach, in the prediction phase, a feed-forward neural network is used to simulate the interaction between the user and the item, where the feature vectors at the pre-processing stage are used as input to the neural network. Experiments based on two data sets (MovieLens 1M and MovieLens 10M) have demonstrated the effectiveness of this method and have given completely feasible results.

The problem of rating prediction and recommendation are inseparable. In order to recommend, it is necessary to have ranking prediction results, and then choose the results with the top ranking for the recommendation. The results of the prediction as a premise can be used for more effective and accurate recommendations. Currently, there are many ratings prediction and recommendation systems with different approaches. However, the systems in the field of education, especially using real data on students' learning performance have not received much attention.

Based on previous studies, this study proposes different techniques, especially deep learning ones to build models of classification, learning resource search, learning performance prediction, and learning resource recommendation to solve existing problems.

3. Proposed approaches

In this study, we propose models for searching and recommending learning resources to meet the needs of learners, helping to achieve better student performance results. The specific tasks of this study include building models for searching learning resources with attention to semantic issues to improve the search effectiveness to meet the needs of learners and building models for predicting student performance and recommending appropriate learning resources for each learner. Learning resources are diverse, including lectures, course books, books, articles, theses, dissertations, images, videos, and other digital learning resources. However, the scope of this study focuses on texts or documents. The general architecture of the proposed models is presented in Fig. 1.

Firstly, an approach for building a learning resource classification model based on deep learning techniques such as Multi-Layer Perceptron (MLP) is proposed. This approach can be compared with other machine learning techniques; however, it can work better in case of non-linear data. The main reason for this classification approach is that it aims to narrow the search space, making the search process more effective. This is shown in the task of ① in Fig. 1.

Secondly, an approach for searching learning resources, which considers the semantics, based on ensemble the similarities of the Cosine and word order is proposed. It inherits the results of the learning resources classification above mentioned. Both approaches pre-process and classify queries and learning resources to determine the respective domain or topic to narrow the search space. This is shown in task of ② of Fig. 1.

Thirdly, models for rating prediction, specifically predicting student performance are proposed. The models use different approaches based on deep learning techniques, including a model that predicts learning overall performance on data using a CNN, a predictive model according to learning ability group using MLP, and per student prediction model using Long Short-Term Memory (LSTM). The reason for using CNN and MLP is that they can deal with non-linear data; while the LSTM can deal with sequential data (temporal effect) since the students' performance can improve/change over time. These models are shown in task ③ of Fig. 1 as a premise for recommending learning resources suitable for learners' abilities.

Finally, a Deep Matrix Factorization (DMF) model is proposed for recommending learning resources that are suitable for the learner's abilities, thereby improving learning performance. This DMF is extended from the standard Matrix Factorization (MF). It replaces the standard DOT product in the matrix factorization with a non-linear function such as a MLP so that it can improve the prediction results. This model is shown in task ④ of Fig. 1.

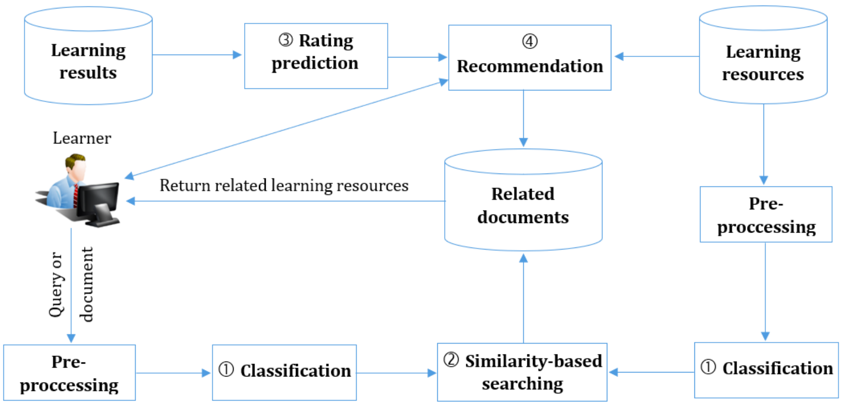

Fig. 1. The architecture of the learning resource search and recommendation system

The proposed techniques for the models in Fig. 1 are based on a previous research [33] and the experimental results. In specific, the MLP is proposed because this technique is suitable for classification prediction and the experimental data in this study in tabular format. In fact, the experiments show that MLP obtains the best results compared to other machine learning techniques. For learning performance prediction model on all students' data, CNN is used because it is a good technique for predictive model with One-Dimensional (1D) and time-sequential data. In the performance prediction model on learning ability group, MLP is proposed because it is suitable for this classification prediction for tabular format data as mentioned. However, CNN does not work well due to the non-sequential data after grouping data according to the learning ability group. LSTM is proposed with a learning performance prediction model per student, since this technique has proven to be very successful for predictive models with data in sequence or time series.

For searching learning resources with an interest in semantics, this study proposes approaches based on an ensemble of the similarities of cosine and word order. This is the model that is based on computing text similarity in a conversational processing system [20]. This model is perfectly suitable to apply for searching learning resources.

These four sub-systems will be presented in the following section including the models and experimental results.

4. Methods and experimental results

4.1. Learning resource classification model

The learning resources are normally very large, searching in the whole resources is not effective work because it takes a lot of time to respond. In the first task of our study, we propose a learning resource classification model using MLP technique. The general idea is that documents are classified into small topics before search, thus time and computer memory can be reduced. The general system of the resource classification model is presented in Fig. 2. When a learning resource, for example in this case as an article, is submitted to the system, it will be classified into a topic based on trained models of machine learning and deep learning.

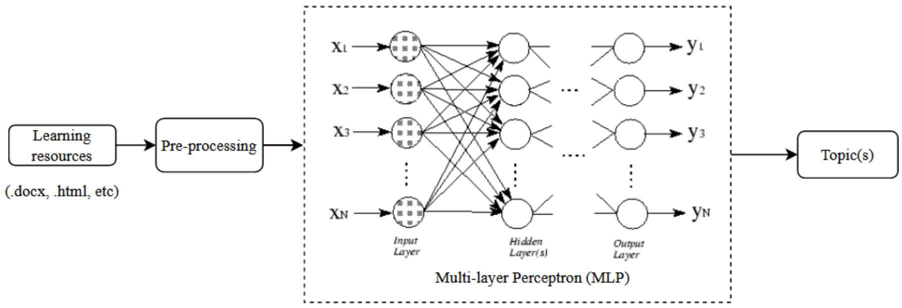

Fig. 2. The proposed architecture for pre-processing and classification of learning resources

We select the parameters for MLP model by searching hyper-parameters. The results show that the MLP architecture consisting of one hidden layer with 16 neurons has the best performance on considered five datasets as described in Table 1. In addition, to minimize overfitting problems, the early stopping technique is used if, during the learning process, the result is not improved for 5 consecutive epochs, up to 10 epochs. The network is deployed with the Adam optimization function; the default learning rate is 0.001.

Table 1. The experimental datasets

Dataset	Number of instances	Number of attributes	Number of classes
Reuters_Newswire	2158	1503	2
School_Text_Books	1786	2566	4
Turkish_News_Articles	3600	5693	6
Scientific_Articles	650	3431	9
VnExpress_Newsletters	10,000	3266	10

To evaluate the model, the Area Under the Curve (AUC) measure with a cross-validation of 3-fold is used. The reason for using AUC is that it is a reliable metric for evaluating classifiers for unbalanced data. The Support Vector Machine (SVM) algorithm and decision tree are used as a baseline for comparison with MLP method. The experimental result shows that MLP has superior classification results compared to the other two algorithms described in Table 2.

Table 2. AUC measure with learning resource classification techniques

Dataset	MLP	SVM	Decision tree
Reuters_Newswire	**0.991**	0.811	0.813
School_Text_Books	**0.999**	0.991	0.928
Turkish_News_Articles	**0.962**	0.949	0.871
Scientific_Articles	**0.977**	0.965	0.819
VnExpress_Newsletters	**0.990**	0.985	0.876

4.2. Learning resource search model

The learning resource documents are usually stored in terms of text (PDF, Word, PowerPoint, etc.); searching in the content of these resources is much more effective than searching the meta-data (attributes) of those resources. For example, when a document is stored as "Chapter 3.pdf", it cannot be searched by the normal query; this needs the content-based searching approach.

An approach for learning resource search based on text similarity is proposed in the second one of the study. In this work, to search for learning resources based on text similarity, the text similarity calculation method is applied based on the conversational processing system [20]. It is suggested to combine the semantic similarity of the document with the similarity of word order in the text. The search model is briefly described by Algorithm 1. The input data is pre-processed, extracted, vectorized, and presented as Term Frequency-Inverse Document Frequency (TF-IDF) and word order. Then, the cosine similarity and word-order similarity of the document are calculated. Finally, these two similarities are combined to calculate the document similarity to apply to learning resource search.

Algorithm 1. SimilarityDetection

Input: Document d, corpus-of-Pre-processed-Documents D, float α, float SimThreshold

Step 1. Conversion(d) // convert the input document (word/pdf) to text

Step 2. WordSegmention(d) // seperate document to words

Step 3. WordNormalization(d) // change to lower cases, remove blanks

Step 4. RemovingStopWords(d)

Step 5. VectorizationTF-IDF(d)

Step 6. VectorizationOrder(d) // the word-order in the sentences

Step 7. sim $\leftarrow \alpha \times$ CosineSimilarity(d, D) + $(1 - \alpha) \times$ OrderSimilarity(d, D)

Step 8. Return sets of documents in D that have sim > SimThreshold

Here: CosineSimilarity(d, D) is semantic similarity (cosine similarity) between d and D; OrderSimilarity(d, D) is word order similarity between d and D. Hyper-parameter of $\alpha \leq 1$ is the importance of semantic similarity and word order similarity of the document. In this study, semantic similarity and word-order similarity are considered to be equally important, $\alpha = 0.5$.

The experimental dataset includes 680 scientific articles in Vietnamese covering 10 fields (topics), published in Can Tho University Journal of Science from 2016 to 2018. The dataset of articles is randomly separated; 90% is used as a training dataset, and the remaining 10% is used as a testing dataset. The document classification model is built using SVM algorithm. To experiment with searching learning resources based on document similarity, a system to check document (scientific articles) similarity is built.

There are two search methods based on document similarity. The first method is to search for the similarity of a document on the entire available corpus (unclassified). The second method is to search for the similarity of a document on each field after classification. For searching on the entire unclassified corpus, the SIMilarity threshold (SIM) is set to SIM > 20% to perform the search. The results show that measures of Precision and Recall are quite low (resulting in a low F1 measure). The reason could be that the search result depends on a given similarity threshold; the search is performed on the entire dataset instead of on the same field as the query to be searched. To overcome this problem, it is necessary to classify the article to be searched (query) before performing a search on the corresponding field.

Table 3. The experimental results in checking the similarity of articles

No	Articles	Results
Field: Technology; SIM threshold > 20%		
1	Development of mix proportion for self-compacting concrete based on optimal dense packing of aggregates and paste content	**Article 1**. Study on reuse of plastic waste to produce light concrete as construction materials. SIM = 0.274 **Article 2**. Developing computer vision algorithm for ripe tomato localization and estimation of the distance from the camera system to the centre of the ripe tomato on the tree. SIM = 0.210

Table 4. The experimental result of checking the similarity of two given articles

No	Article 1	Article 2	SIM threshold	Result
1	Biomass of Melaleuca forest at the U Minh Thuong National Park, Kien Giang Province	Biomass and CO_2 absorption of Melaleuca forest in Lung Ngoc Hoang Natural Reserve	> 30%	SIM=0.556

For the search method on classified datasets, there are 10 fields (10 classes). When uploading a query that is an article to be searched for similarity, the search system will classify the query based on the built classification model to determine the field of the article. Then, the system performs a search on the corresponding field of the article (query) with a given threshold and returns a list of articles with a similarity matching the threshold.

The experiment to check the similarity of the article is presented as an example in Table 3. When the user uploads an article to check the similarity and selects the similarity threshold, then the checking performance will result in articles that are similar to the article being considered with the given threshold.

Experimentally checking the similarity of two certain articles is performed; the result is described as an example in Table 4.

4.3. Learning performance prediction models

For recommending the learning resources, first, we need to predict which resources are suitable for the learners, then, recommendations are generated by sorting the prediction scores. However, the recommendation task is too easy after having the prediction scores. In this work, we focus on the first phase of recommendation systems, which is finding the best prediction models. This part proposes the models to predict learning performance with three approaches based on deep learning techniques, including building a prediction model for all students using CNN, a prediction model of learning ability using MLP, and a prediction model for per student using LSTM and MLP.

4.3.1. Learning performance prediction model on all student data

In the first approach, the learning performance prediction model on all students' data uses CNN architecture on one-dimensional data illustrated in Fig. 3. The proposed CNN network takes as input a data sequence with 21 attributes passing through the first convolutional layer using 64 kernels of size 3 with a stride of 1. The experimental data that is collected related to students, courses, marks, and other information from 2007 to 2019 contain more than 3.8 million records. The data is phased from 2007 to 2016 for training and from 2017 to 2019 for testing.

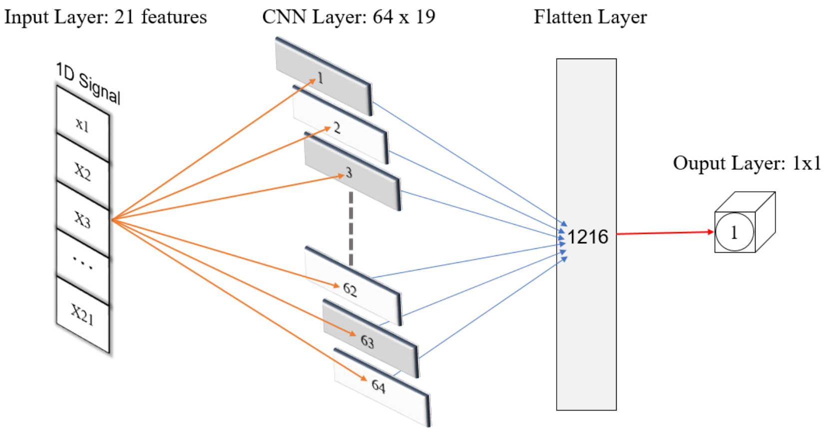

Fig. 3. The proposed CNN architecture

Two optimization functions of RMSprop and Adam are compared and used. After five epochs consecutively, if the result does not improve, the learning process stops, running up to 500 epochs. The Mean Absolute Error (MAE) measure is used. Large input values slow down the learning and convergence process, and the training time is large. Therefore, it is necessary to be able to scale the values of the attribute to a certain range of values. In this study, Quantile TransFormation (QTF) is suggested as a data transformation, helping deep learning algorithms to converge better. In the experiment, using CNN, QTF, and Adam optimization function show that the performance prediction results have a quite good error when there are 16 considered datasets whose MAE measures are all less than 0.8 (prediction on a scale of 4), some of them are less than 0.5. Besides using CNN, QTF, and Adam optimization function, RMSprop optimization function is used to compare and evaluate the proposed model more objectively.

The experimental results in Table 5 show that with the prediction model using CNN, the RMSprop optimization function gives better prediction results than Adam on most of the datasets being considered (13 out of 16 datasets), when using QTF. This result shows that the RMSprop optimization function may be suitable when using One-Dimensional (1D) and sequence time data.

Table 5. The results of learning performance prediction with MAE measure using CNN, QTF, Adam and RMSprop optimization function

Dataset	CNN-RMSprop	CNN-Adam
Education	**0.5733**	0.5847
Environment and Natural Resources	**0.5989**	0.6130
Economics	**0.5922**	0.6098
Foreign Languages	**0.4853**	0.4961
Social Sciences and Humanities	0.5920	**0.5793**
Aquaculture and Fisheries	**0.5918**	0.6471
Law	**0.5546**	0.5675
Political Sciences	0.5765	**0.5547**
Mekong Delta Development Research	**0.5678**	0.5684
Agriculture	**0.5806**	0.5828
Biotechnology R&D	**0.5330**	0.5980
Physical Education	**0.6762**	0.6853
Engineering Technology	**0.7454**	0.7487
Information & Communications Technology	**0.6903**	0.7285
Natural Sciences	**0.6725**	0.7989
Rural Development	0.7134	**0.6936**

With QTF, RMSprop and Adam optimization functions, the model also uses CNN to predict learning performance on the entire datasets containing more than 3.8 million records collected from all academic units of Can Tho University. The results show that using the Adam optimization function is better than the RMSprop optimization function when using the prediction model with CNN architecture. This can be explained that when the entire datasets are used, the sequence nature of the data is limited, so the RMSprop function may not promote its strengths.

4.3.2. Performance prediction model on learning ability group

For this approach, four prediction models are proposed for four groups of students with different academic abilities, using MLP techniques shown in Fig. 4.

The MLP architecture consists of an input layer, an output layer, and five hidden layers. The input layer contains data attributes; the output layer has 1 neuron representing the mark to be predicted with a value from 0 to 4. The first of four hidden layers contains 256 neurons while the fifth hidden layer contains 8 neurons. The early stopping technique is used with 5 epochs, running up to 500 epochs; Adam optimization function is used; the default learning rate is 0.001.

Collected data concern students, courses, marks, and other information from 2007 to 2019 with more than 3.8 million records. The data are divided by time; the training dataset and the test dataset have a ratio of 2/3 and 1/3, respectively.

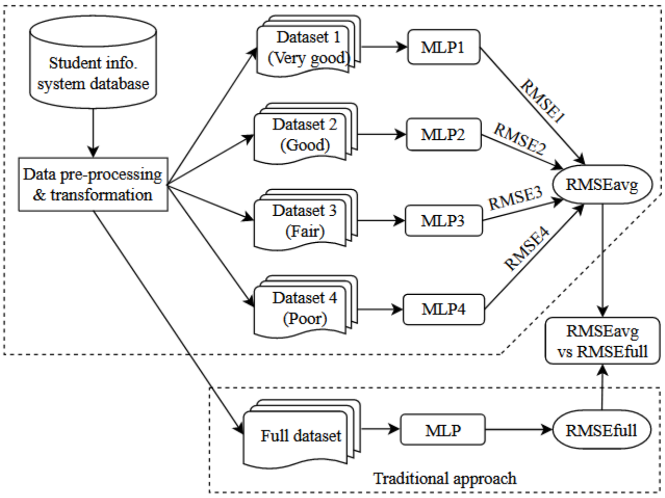

Fig. 4. The overall diagram of the approach

For comparison, baselines are used such as User Average (prediction based on the average results of students), and Item Average (prediction based on the average results of course). In addition, other methods of collaborative filtering are compared. In this study, two common measures, RMSE and MAE, are used to evaluate the models, averaging over 10 experimental runs. The experimental results with the two measures of RMSE and MAE are presented in Fig. 5. GroupMLP presents four models based on four groups of students learning abilities. MLP presents a model to predict the academic performance of all students.

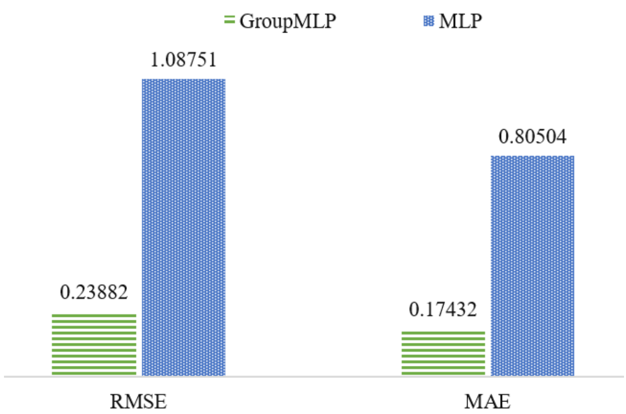

Fig. 5. Measure comparison between GroupMLP and MLP

The results show that GroupMLP performs better than other baselines of the recommender system with the two measures of MAE and RMSE, giving an improved result of over 70%. In addition to using MLP technique, another prediction model is proposed based on GPA to divide into four different models (including excellent, very good, good, and fair) using RF algorithm. The results show that this model also gives good prediction results according to each group of learning ability.

4.3.3. Learning performance prediction model on per student

In the third approach, prediction models are proposed to predict the learning performance of individual students using LSTM and MLP. The LSTM architecture takes as input sequences of time steps. The LSTM layer has 50 neurons, and a dense layer (hidden layer) with 1 neuron gives the result of the prediction a value between 0 and 4 as shown in Fig. 6.

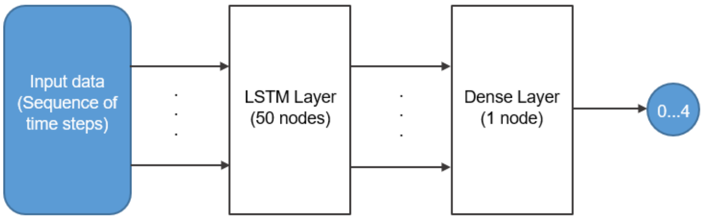

Fig. 6. Architecture of LSTM

Meanwhile, the MLP network architecture consists of an input layer, five hidden layers, and an output layer. The input layer contains the attributes of the input data. The first hidden layer has 9 neurons using the activation function of ReLU;

– the second and third hidden layers have 27 neurons using the activation function of the sigmoid;

– the fourth hidden layer has 9 neurons using the activation function of ReLU;

– the fifth hidden layer has 1 neuron for the output value between 0 and 4 as shown in Fig. 7.

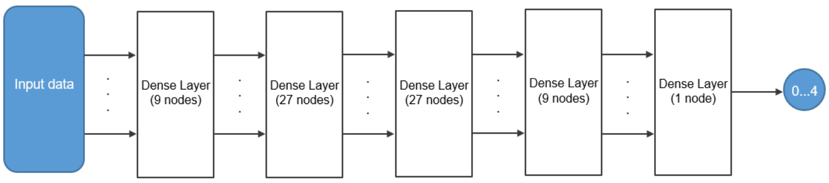

Fig. 7. Architecture of MLP network

For the experiment, a dataset of students' learning performance in some academic units (mainly in science and engineering technology) from a university academic performance data has been collected from 2017 to 2019 with more than one million records. To diversify the experimental data, the original dataset with more than 1 million records of students' learning performance for courses is divided into two new datasets that retain students with at least 10 records and 20 records of learning results. The prediction results using the RMSE measure with LSTM and MLP are shown in Table 6.

Table 6. The predictive results using RMSE with the architecture of LSTM and MLP

Dataset	LSTM	MLP	Description
StudentPerformance10	0.505	0.536	The dataset has 10 records per student
StudentPerformance 20	0.513	0.526	The dataset has 20 records per student

From the results, the LSTM model has better prediction performance than the MLP model on the same dataset. This shows that the LSTM network works quite well on data with sequence time. With the model using MLP architecture, the results are quite good compared to the model for all students in the previous section.

4.4. Learning resource recommendation model

In the last task of the study, a DMF model extended from standard MF to recommend learning resources suitable to learners' abilities is proposed. The recommendation model is detailed in Fig. 8.

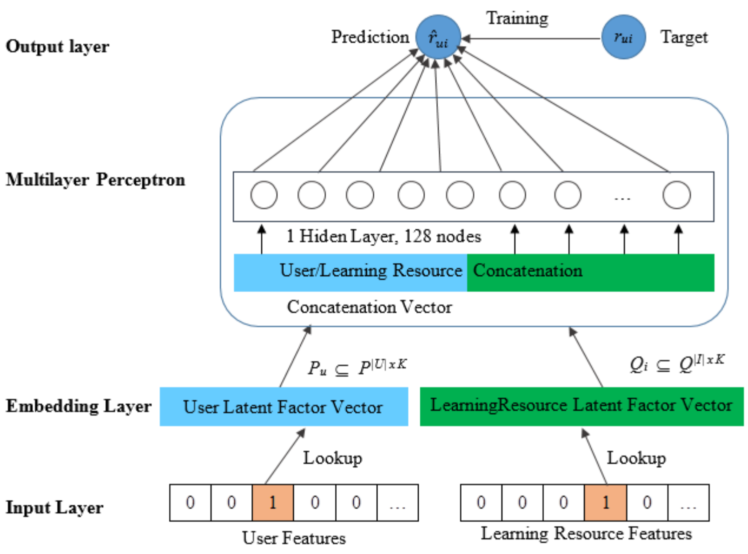

Fig. 8. Framework of DMF model

The DMF model being proposed has four layers. The input layer describes the current user or learning resource. Embedding layer to embed user and learning resources features (latent factors). Embedded features are concatenated as input to the hidden layer MLP. Finally, the output layer results in the predicted rating value. In this work, the hidden layer MLP has 128 neurons (the number of hidden layers and the number of neurons can be set depending on the dataset). The number of neurons is selected by the method of hyper-parameter searching. The network is deployed with the Adam optimization function, with a default learning rate of 0.001.

For experimental data, the proposed model is verified based on two groups of data including datasets of learning resources and datasets of students' learning performance at a university. The datasets of learning resources include five datasets describing the ratings of learning resources (items) of users. The number of users, learning resources, and ratings of these datasets are described in Table 7. These datasets are rather sparse, so they are filtered to retain users or learning resources having at least five ratings.

Table 7. Description of five datasets as learning resources

No	Dataset	Number of user	Number of item	Number of ratings
1	Ratings	53,424	10,000	981,756
2	Library things	70,618	385,251	1,387,125
3	BX-Book-ratings	105,283	340,556	1,149,780
4	Related-Article recommendation	2,663,825	7,224,279	48,879,167
5	Ratings-Books	8,026,324	2,330,066	22,507,155

Datasets of students' learning performance include three datasets. The first dataset is the students' learning performance in a university's academic units. The second dataset is the students' learning results, which retained at least 10 records (10 courses) for each student. Similarly, the third dataset retains at least 20 records for each student. The datasets are described in Table 8.

Table 8. Description of 3 datasets as students' learning performance

No	Dataset	Number of user	Number of item	Number of ratings
1	Student performance	94.087	4.836	1.046.515
2	Student performance 10	30.820	3.516	472.003
3	Student performance 20	1.182	485	16.590

Both experiments on the two data groups including datasets on learning resources and datasets on students' learning performance have quite similar results. For instance, in a learning resource dataset of dataset 1 (rating dataset), we find the number of the MLP layer's neurons is about 100; the number of latent factors K~10 described in Fig. 9; the number of epochs for the DMF model to converge is 2, comparing to the MF model that converges after 4 to 6 epochs described in Fig. 10. Similar to the data group of students' learning performance, the DMF model always converges earlier.

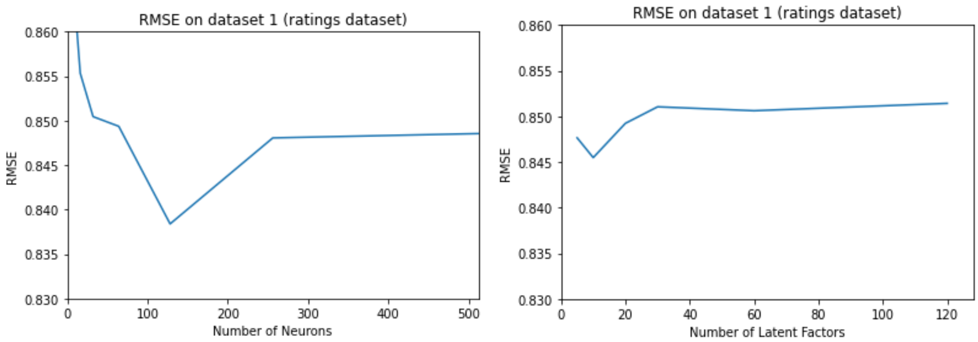

Fig. 9. The chart shows the relationship between the Number of Neurons and the Number of Latent Factors (features) for DMF model in RMSE performance

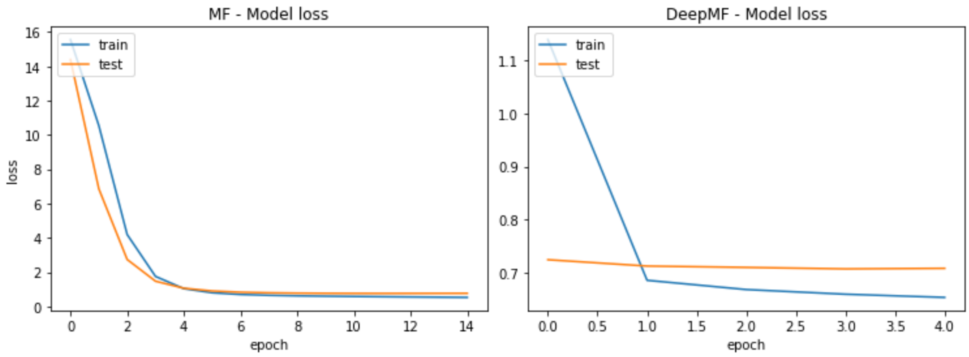

Fig. 10. Comparison of train loss and validation loss of the MF and DMF on Ratings dataset

In this work, the RMSE measure is used to evaluate the DMF model and compare it with other methods of the recommender system such as Global Average, User Average, Item Average, User kNN, and MF. An instance of the RMSE measure between DMF model and other methods in the recommender system on Dataset 1 is shown in Fig. 11. In general, DMF gives superior results compared to other methods of the recommender system. The datasets that overcome the sparse data situation have better results than the original dataset. From the results, the ratings can be used to recommend courses or learning resources suitable for learners. Similar results are also found in other datasets.

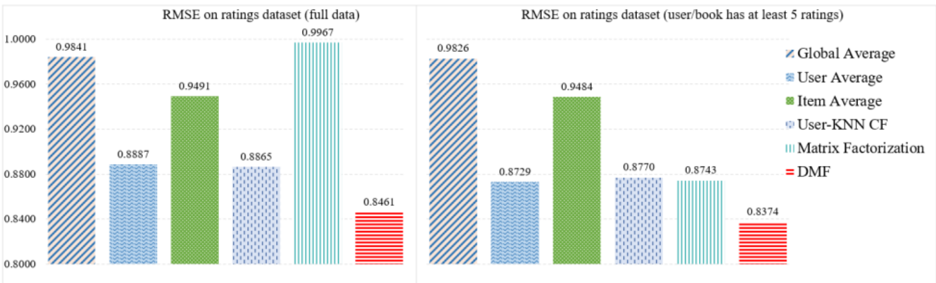

Fig. 11. Comparison of the RMSE measure between the methods on Dataset 1 (Ratings)

5. Conclusions

In this study, we propose state-of-the-art approaches for building models for searching and recommending learning resources. In order to achieve the expectation, the models of classification, learning resource search, learning performance prediction, and learning resource recommendation with various techniques are proposed to solve existing problems. The results of the study can be summarized as follows.

A learning resource classification model based on MLP is proposed. Besides, the results of comparing the deep learning techniques with other machine learning techniques show that this new approach gives a more feasible and effective performance of document classification.

The approach for searching learning resources is proposed based on document similarity. In this approach, queries and learning resources are classified to identify the topic to narrow the search space before searching on the corresponding topic of the built learning resources.

Models to predict learning performance are proposed using deep learning techniques including the learning performance prediction model on all student data using CNN, the learning performance prediction model on ability group using MLP, and the learning performance prediction model on per student using LSTM and MLP. The experimental results show that the three proposed models give ascendingly good predictive results, respectively. It can be seen that the proposed models and techniques, especially deep learning techniques are much potential to build prediction models of learning performance in particular or learning resources in general.

A learning resource recommendation model using the DMF, which is extended from the standard MF technique, is proposed. The model is validated with two groups of datasets including datasets of learning resources and datasets of students' learning performance at a university. The DMF model is also compared with other baselines of the recommender systems. The results show that the DMF model has good rating prediction performance compared to other techniques, thereby recommending suitable learning resources or courses for each learner.

In this study, the proposed models of classification, search, prediction, and recommendation focus on textual learning resources. Further research could be conducted on these models for other types of learning resources, like videos. In addition, it is possible to propose a solution to integrate the models of classification, search, rating prediction, and learning resource recommendation into a learning resource management system that can be applied to the context of educational institutions, especially higher education institutions.

References

1. S u s a n, D. A. Open Educational Resources: Reviewing Initiatives and Issues. – Open Learning: The Journal of Open, Distance and e-Learning, Vol. **24**, 2009, No 1, pp. 3-10.
2. C h a n g, B. Culture as a Tool: Facilitating Knowledge Construction in the Context of a Learning Community. – International Journal of Lifelong Education, Vol. **29**, 2010, No 6, pp. 705-722.
3. UNESCO. Recommendation on Open Educational Resources (OER). 2019.
4. S m i t h, M. C. What are learning resources? What are some examples? 2016.
5. B a r a j a s, M., G. J. G a n n a w a y. Implementing e-Learning in the Traditional Higher Education Institutions. – Higher Education in Europe, Vol. **32**, 2007, No 2-3, pp. 111-119.
6. M u ñ o z-M e r i n o, P. J., C. D. K l o o s, J. F. N a r a n j o. Enabling Interoperability for LMS Educational Services. – Computer Standards & Interfaces, Vol. **31**, 2009, No 2, pp. 484-498.
7. S a n t o s, O. C., J. G. B o t i c a r i o. Requirements for Semantic Educational Recommender Systems in Formal e-Learning Scenarios. – Algorithms, Vol. **4**, 2011, No 2, pp. 131-154.
8. L i, X., D. R o t h. Learning Question Classifiers. Association for Computational Linguistics. 2002.
9. H a c i o g l u, K., W. W a r d. Question Classification with Support Vector Machines and Error Correcting Codes. Association for Computational Linguistics. 2003.
10. Z h a n g, D., W. S. L e e. Question Classification Using Support Vector Machines. – In: Proc. of 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, Toronto, Canada, 2003 (Insert City of Publication; Insert 2003 of Publication).

11. B r o w n, J. Entity-Tagged Language Models for Question Classification in a QA System. 2004. **http://www-2.cs.cmu.edu/jonbrown/IRLab/Brown-IRLab.pdf**
12. M i s h r a, M., V. M i s h r a, H. R. S h a r m a. Question Classification Using Semantic, Syntactic and Lexical Features. – International Journal of Web & Semantic Technology, Vol. **4**, 1.07. 2013.
13. D u r m e, B., Y. C. A. H u a n g, E. N y b e r g. Towards Light Semantic Processing for Question Answering. 2.03. 2004.
14. D e, T. C., P. N. K h a n g. Text Document Classification with Support Vector Machine and Decision Tree. – Can Tho University Journal of Science, Vol. **21a**, 2012, pp. 52-63 (in Vietnamese).
15. T h a o, T. T. T., V. T. C h i n h. Building a Vietnamese Document Classification System. Hong Bang University, 2012 (in Vietnamese).
16. C h a i, D., W. W u, Q. H a n, F. W u, J. L i. Description Based Text Classification with Reinforcement Learning, 2020.
17. L i n, E., Q. C h e n, X. Q i. Deep Reinforcement Learning for Imbalanced Classification. – Applied Intelligence, Vol. **50**, No 8, 2020, pp. 2488-2502.
18. Z u l q a r n a i n, M., R. G h a z a l i, Y. M a z w i n, M. R e h a n. A Comparative Review on Deep Learning Models for Text Classification. 13.04.2020.
19. G o m a a, W., A. F a h m y. A Survey of Text Similarity Approaches. – International Journal of Computer Applications, Vol. **68**, 18.04.2013, No 13, pp. 13-18.
20. L i, Y., Z. B a n d a r, D. M c L e a n, J. O'S h e a. A Method for Measuring Sentence Similarity and Its Application to Conversational Agents, 2004.
21. G u n a s i n g h e, U. L. D. N., W. A. M. D. S i l v a, N. H. N. D. da S i l v a, A. S. P e r e r a, W. A. D. S a s h i k a, W. D. T. P. P r e m a s i r i. Sentence Similarity Measuring by Vector Space Model. 2014.
22. A l-S h a m e r y, E., H. Q. G h e n i. Plagiarism Detection Using Semantic Analysis. – Indian Journal of Science and Technology, Vol. **9**, 2016, No 1, pp. 1-8.
23. S h a r m a, K., S. S i n g h, I. M u k h e r j e e, B. K u m a r. Plagiarism Detection Based on Semantic Analysis. – International Journal of Knowledge and Learning, Vol. **12**, 1.01.2018, 242.
24. W a l i, W., B. G a r g o u r i, A. B e n H a m a d o u. An Enhanced Plagiarism Detection Based on Syntactico-Semantic Knowledge. Springer International Publishing. 2020.
25. F a r o u k, M. Measuring Text Similarity Based on Structure and Word Embedding. – Cognitive Systems Research, Vol. **63**, 1.10.2020, pp. 1-10.
26. R a s t r o l l o-G u e r r e r o, J. L., J. A. G ó m e z-P u l i d o, A. D u r á n-D o m í n g u e z. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. – Applied Sciences, Vol. **10**, 2020, No 3.
27. T h a i-N g h e, N., T. H o r v a t h, L. S c h m i d t-T h i e m e. Factorization Models for Forecasting Student Performance. 2011.
28. H u y n h-L y, T.-N., N. T h a i-N g h e. A System for Predicting Students's Course Result Using a Free Recommender System Library of MyMediaLite. 2013 (in Vietnamese).
29. H a i, N. P., T.-W. S h e u, M. N a g a i. Predicting the Student Learning Outcomes Based on the Combination of Taylor Approximation Method and Grey Models. – VNU Journal of Science: Education Research, Vol. **31**, 2015, No 2, pp. 70-83 (in Vietnamese).
30. I q b a l, Z., J. Q a d i r, A. M i a n, F. K a m i r a n. Machine Learning Based Student Grade Prediction: A Case Study. 17.08.2017.
31. Z h a n g, L., T. L u o, F. Z h a n g, Y. W u. A Recommendation Model Based on Deep Neural Network. – IEEE Access, Vol. **6**, 2018, pp. 9454-9463.
32. F u, M., H. Q u, Z. Y i, L. L u, Y. L i u. A Novel Deep Learning-Based Collaborative Filtering Model for Recommendation System. – IEEE Transactions on Cybernetics, Vol. **49**, 2019, No 3 pp. 1084-1096.
33. B r o w n l e e, J. When to Use MLP, CNN, and RNN Neural Networks. 2022.

Received: 14.11.2022; Second Version: 03.05.2023; Accepted: 12.05.2023