Research on Online Learning Resource Recommendation 
Method Based on Wide & Deep and Elmo Model

Jie Liu\textsuperscript{a}, Hong Zhang\textsuperscript{a,b}, Zihui Liu\textsuperscript{c}
College Of Computer & Information Science, Southwest University, ChongQing, China
\textsuperscript{a}addressliujie$_{\text{ndx}}$@163.com, \textsuperscript{b}zhangh@swu.edu.cn, \textsuperscript{c}zihuil@163.com

ABSTRACT. In order to improve the accuracy of online learning platform recommendation for learner learning resources and to alleviate the cold start problem, this paper proposes an online learning resource recommendation method based on Wide&Deep and Elmo model. The method uses the real learning data set of a university network and educational learning platform, and uses Wide&Deep to deeply explore the deep features of learner characteristics and course content features under the condition of high-dimensional data sparseness, that is, automatic learning combination features, the learner-course feature vector is constructed as the input of the recommendation algorithm. In addition, for the learner's text feature, it will use the ELMo language model to pre-train the feature vector to improve the recommendation accuracy.

CCS Concepts
• Computing methodologies ➔ Cognitive science • Theory of computation ➔ Online learning algorithms

1. INTRODUCTION
With the close combination of Internet and education, online learning platforms emerge like bamboo shoots. Faced with the uneven learning courses on the learning platforms, how to let users find the most suitable courses in the massive curriculum information will become one of the most important problems to be solved urgently. To solve this problem, many scholars use recommendation system.

At present, there are many studies on the application of recommendation system to online learning platform. García [1] et al. introduced a data mining algorithm for collaborative filtering education based on Association rules, which provided suggestions for teachers to improve e-learning courses. Bousbahi [2] et al. put forward a case-based MOOC recommendation system, which uses Case-Based Reasoning (CBR) method and a special retrieval information technology to propose the most appropriate courses to learners according to their personal information, needs and knowledge. Vesin[3] et al. introduced a program teaching Protus system, The system can adapt learners' interest and knowledge level, cluster processing according to different learning styles, and then mining frequent sequences using a priori algorithm to analyze learners' habits and interests. Finally, personalized recommendation of learning content is completed according to the evaluation of these frequent sequences provided by Protus system. Li [4] et al. put forward a recommendation system based on relevant pattern recommendation, which can combine the advantages of user-based clustering and course-based clustering to measure the relevance between users and courses. Su [5] et al. proposed a big data analysis technology for learning portfolio, which studies learners' learning on the platform, recognizes the impact of large data analysis on students' learning behavior, and considers that learning behavior information can be used as an evaluation criterion for teachers to recommend similar courses.
to students. Huang [6] et al. proposed MCRS algorithm, studied curriculum recommendation algorithm based on distributed computing framework, and stored curriculum recommendation rules in database.

With the advent of artificial intelligence, deep learning has made breakthroughs in image processing, natural language understanding and speech recognition. It has become an upsurge of artificial intelligence and brought new opportunities for the research of recommendation systems. Oh [7] et al. proposed a news recommendation system model based on deep neural network user preference analysis, which extracts interest keywords from a group of news articles that specific users have read in the past to describe user preferences, and then makes news recommendation to users. Zhang [8] et al. and others proposed a personalized recommendation system based on DBN in MOOC environment. The system combines user-course feature vectors in learning platform to mine users' interest in courses. Huang [9] et al. proposed a high-precision resource recommendation model based on deep trust network (DBN) in MOOC environment. Stically mining of learner characteristics and course content attributes, combining learner behavior characteristics, user-course feature vectors are constructed as input of the deep model. So as to achieve the purpose of improving the learning efficiency and enthusiasm of learners.

These recommendation systems based on collaborative filtering algorithm, hybrid recommendation algorithm or big data technology can improve the accuracy of recommendation to a certain extent. However, it is difficult to learn the deep-seated features of learners and courses, and a lot of human and mental resources are needed to design the features artificially. These methods are rough in dealing with text features and do not accurately extract feature vectors from text features, which greatly restricts the performance of recommendation methods. In order to

Figure 1. Research on Online Learning Resource Recommendation Method Based on Wide & Deep and Elmo Model

solve the above problems, This paper proposes an online learning resource recommendation method based on Wide&Deep and Elmo models. As shown in Figure 1, this method can deeply mine the relationship between data features, and automatically learn feature combination in the case of high-dimensional sparse data without manual design features. In addition, in the past, learning resource recommendation tends to excavate the relationship between simple features for recommendation. This paper will train online learners on the language model of course commentary text, so as to improve the accuracy of recommendation. Experiments shows that text features are pre-trained by ELMo language model, and then trained by wide model for downstream tasks, which improves the generalization of the recommendation research and makes the research of the recommendation algorithm more accurate.

The main contributions of the paper include:

This paper proposes an online learning resource recommendation method based on Wide&Deep and Elmo model. Firstly, the combination of features can be automatically learned in high-dimensional sparse data environment, which reduces the labor cost and alleviates the cold start problem to a certain
extent. Secondly, the ELMo language model pre-training for learners’ text features can extract more accurate feature information, which will improve the generalization of recommendation.

2. RELATED WORK

2.1 Wide&Deep Models
The Wide&Deep model of recommendation system is proposed by Heng-tze[10] et al. for Google Play. Its core idea is to optimize the parameters of two models simultaneously in the training process by combining the memory ability of linear model with the generalization ability of DNN model, so as to achieve the goal of optimizing the prediction ability of the whole model. As shown on the right side of Figure 1.

2.1.1 Wide Models
Wide Models is a generalized linear model:

\[ y = w^T x + b \]  \hspace{1cm} (1)

Where feature \( x = [x_1, x_2, ..., x_d] \) is a vector of d-dimension, \( w = [w_1, w_2, ..., w_d] \) is the parameter of the model. Finally, on the basis of \( y \), Sigmoid function is added as the final output. The feature set includes the original input feature and the transformation feature. One of the most important feature transformations is cross product transformation, which is formulated as follows:

\[ \phi_k = \prod_{i=1}^{d} X_i^{C_{ik}} \cdot C_{ik} \in \{0,1\} \]  \hspace{1cm} (2)

where \( C_{ik} \) is a boolean variable, if the \( i \)-th feature is part of the \( k \)-th transformation, \( \phi_k \) is 1, otherwise 0.

2.1.2 Deep models
Deep model is a feed forward neural network, where the vectors learned from the model are implicit, often without explicit explanatory, and are automatically generated without human intervention. Deep neural network models usually need continuous and dense input features. Therefore, for sparse and high-dimensional category features such as online learning platforms, it is necessary to first transform them into low-dimensional vectors. This process is also called embedding.

In training, embedding vectors are randomly initialized, and the values of the vectors are gradually modified during the training process of the model, that is, the vectors are used as parameters to participate in the training of the model.

The calculation method of hidden layer is as follows:

\[ \alpha^{(l+1)} = f(W^{(l)} \alpha^{(l)} + b^{(l)}) \]  \hspace{1cm} (3)

where \( l \) is the layer number and \( \alpha^{(l)}, b^{(l)} \) and \( W^{(l)} \) are the activations, bias, and model weights at \( l \)-th layer. \( f \) is the activation function, This paper uses ReLUs.

2.1.3 Wide&Deep Models
Wide &Deep model is to train Wide model and Deep model at the same time, and take the weighted sum of the results of the two models as the final prediction result:

\[ P(Y = 1 | x) = \delta(W_{wide}^T x, \phi(x)) + W_{deep}^T \alpha^{(l)} + b \]  \hspace{1cm} (4)

2.2 ELMo Models
In recent years, there are many studies on text context information analysis, hoping to get word vectors accurately and better. ELMo is a better method in natural language processing at present. It is proposed by Peters [11] et al. and learns from the internal state of the deep two-way language model. Its model structure is shown on the left side of Figure 1.
The model is mainly used in two training stages. First, the language model is used for pre-training. Then, Word Embedding is extracted from the pre-training network at all levels of the network to supplement the downstream task as a new feature.

In this model, its network structure adopts two-tier bidirectional LSTM. The author gives a sequence containing N tokens \( \{t_1, t_2, \ldots, t_N\} \) for pre-training. Language model calculates the probability of the occurrence of the k-th token by giving the token sequence of the previous k-1 position, also known as the forward representation:

\[
p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, \ldots, t_{k-1}) \tag{5}
\]

The reverse expression is:

\[
p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_{k+1}, t_{k+2}, \ldots, t_N | t_k) \tag{6}
\]

Forward is to use the above to predict the following, backward is to use the following to predict the above.

Assuming that the input token is \( x_{k}^{LM} \), in every position k, the corresponding context-dependent representation \( h_{k,j}^{LM} \) is output on each LSTM layer. \( L \) represents the number of layers of LSTM. Top-level LSTM output \( h_{k,j}^{LM} \). Predicting the next \( \text{token}_{k+1} \) through the softmax layer.

The objective function is the maximum likelihood of the two directional language models:

\[
\sum_{k=1}^{N} \log p(t_k | t_1, t_2, \ldots, t_{k-1}); \Theta_L, \Theta_{\text{LSTM}}, \Theta_x
\]

After the language model is pre-trained, for each token, a L-level biLM calculates 2L+1 representation:

\[
R_k = \left\{ x_{k}^{LM}, \ldots, h_{k,1}^{LM}, \ldots, h_{k,j}^{LM}, \ldots, h_{k,L}^{LM} \right\} = \left\{ h_{k,j}^{LM} | j = 0, \ldots, L \right\} \tag{8}
\]

Where \( h_{k,j}^{LM} \) is the value of the token layer.

Finally, ELMo sums up each middle layer of the two-way language model as a word representation, and the formula is as follows:

\[
ELMo_{k,task}^{\text{task}} = E(R_k; \Theta_{\text{task}}) = \gamma_{\text{task}} \sum_{j=0}^{L} S_{j}^{\text{task}} h_{k,j}^{LM} \tag{9}
\]

Among them, \( \gamma_{\text{task}}^{\text{std}} \) is the standardized weight of soft-max and \( \gamma_{\text{std}}^{\text{std}} \) is the scaling coefficient, which allows the task model to scale the whole ELMo vector.

3. PROPOSED MODEL
Because Wide&Deep model has the ability to automatically combine features in high-dimensional sparse data and Elmo model has the ability to encode text information, this paper proposes an online learning resource recommendation method combining Wide&Deep and Elmo models, as shown in Figure 1. The problem we study is a binary classification problem. Through our proposed method, it can be applied to situations including continuous features, categorical features and text features. It is of great significance to improve the generalization of recommendation.

3.1 Data Sources
The data in this paper are from December 2016 to June 2018 of Southwest University. The platform provides various communication functions for learners, such as discussion, questions and answers. In this paper, 488 students with 25226 text characteristics are selected for experimental analysis.

3.2 Data Preprocessing
The data set adopted in this paper includes the following four parts:
the first part is the user's attribute characteristics, including studentID, gradation, major, Original_major, Graduation_type, that is, student ID, student level, specialty, original study specialty, Diploma type.

The second part is the user's behavior characteristics, including Login_count, question, Study_time, Learning_times, Detail, that is, login times, questions, learning time, learning times, discussion.

The third part is the user's scoring information, including Ogrades, Noachievement, Homework_achievement, Roll_score, Final_result, that is, normal grades, resource learning, homework and paper results.

The fourth part is learning resources information, mainly including Cour_Name, TeachID, teachTitle, teachSource, teachProfession, teachDegree, respectively indicating the curriculum name, teacher ID, teacher title, teacher source, teacher's profession, teacher's degree.

In data processing tasks, usually the data cannot be used directly, so it may contain more missing values. Our data set also has this situation. According to the characteristics of our data set, we adopt different processing strategies. If an attribute has too many missing values, it may contain little information. If the attribute is retained, there may be a serious over-fitting problem. Therefore, we abandon the attribute with too many missing values and set a threshold of 80% in our experiments. For the attributes with less missing values, there are three kinds of attributes: continuous feature, category feature and text feature. We use mean to fill in continuous feature and mode to fill in category feature and text feature. We use mean to fill in continuous feature and mode to fill in category feature and text feature.

In view of the fact that the original deep & wide model cannot deal with text features directly, the proposed model takes into account the information coding model in the field of natural processing to encode text features. In our experiments, we designed three experiments: 1) not using text features. 2) using text features as category features. 3) Encoding text features using Elmo model, and then using Deep&Wide model as the downstream task of the encoded text information vector. In the above three experiments, it is worth noting that the second experiment is meaningless if the textual features of each sample are different. At this time, the textual features can act as a unique identifier. We find that there are a lot of repetitions in text features, which means that many samples share the same text features, so we can use category features to deal with text features.

3.3 Experimental settings
In the study of this recommended method, three experiments were set up, and the learning rate of each experiment was 0.001. In this paper, the optimization algorithm used in Wide model is FTRL. AdaGrad algorithm is used in Deep model, which belongs to adaptive algorithm and can update the learning rate dynamically.

3.4 Evaluation Indicators
This study uses offline evaluation method to verify the performance of the recommendation method, which is based on whether it is necessary to provide the recommendation system for learners. Therefore, this paper mainly uses accuracy to calculate the difference between the predicted score and the actual score,
and uses it as the evaluation criterion of the accuracy of the recommendation. The formula is as follows:

\[
\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}
\]

Among them, TP refers to the number of positive samples, TN refers to the number of negative samples, FP refers to the number of positive samples and FN refers to the number of positive samples.

4. EXPERIMENT AND ANALYSIS

This paper divides the data set into two parts, and sets the ratio of training set to test set to 8:2. Firstly, the training is set to epoch to 70 and 1000 without using text features, as shown in the left figure of Figure 2, and then to embedding text features in Wide&Weep model, as shown in the right figure of Figure 2. Finally, epoch is set to 200, 500, 1000, and text features are pre-trained in the Elmo language model to get feature vectors, and then trained in the Wide&Deep model, as shown in Figure 1.

For each case, four training sessions were conducted, and the accuracy results are shown in Table 1.

Table 1. Three kinds of experimental accuracy results

|                  | No Using Text Features | Using Text Features | Using ELMo |
|------------------|------------------------|---------------------|------------|
| epoch            | 70                     | 1000                | 70         | 1000      | 200 | 500 | 1000 |
|                  | first                  | second              | third      | fourth    | mean value |
| accuracy         | 0.925 0.918            | 0.916 0.905         | 0.984 0.983 | 0.961     | 0.926 0.925 0.913 0.944 0.906 0.939 0.978 0.978 0.967 | 0.926 0.925 0.912 0.924 0.980 0.976 0.963 |

From the results of accuracy, we can see that the average accuracy is 0.926 without using text features, which indicates that the original deep&wide model has a certain effect on recommending online learning resources. When text features are used as categorical features, the accuracy is basically unchanged, which indicates that the information of text features is not effectively expressed. The experimental results of our proposed model show that the accuracy is much improved compared with the previous two cases. This shows that the proposed model adequately encodes the information of text features, which greatly improves the ability of the model.

5. Conclusions

With the rapid development of online learning platform, the data of online learners and learning resources are increasing at an alarming rate. Recommending learning resources of interest to learners is an important recommendation system at present. In order to improve the accuracy of online learning platform's recommendation for learners' learning resources and alleviate the cold start problem, this paper proposes an online learning resources recommendation method based on Wide&Deep and Elmo model, in the case of high-dimensional sparse data, Wide&Deep model can automatically combine features and mine deeper features, ELMo language model can mine feature information in text features and improve generalization of recommendation.

Experiments show that applying the two models together can greatly improve the accuracy of recommendation, which is of great significance to both learning platforms and learners.

References
[1] García E, Romero C, Ventura S, et al. 2011, A collaborative educational association rule mining tool. Internet High Educ. 14(2):77–88
[2] Bousbahi, F., & Chorfi, H. 2015, MOOC-Rec: A Case Based Recommender System for MOOCs. Procedia - Social and Behavioral Sciences, 195, 1813–1822.
[3] Klašnja-Milićević A, Vesin B, Ivanović M et al. 2011, E-Learning personalization based on hybrid recommendation strategy and learning style identification[J]. Computers in Education, 56(3):885–899

[4] Li Y, Li H. 2017, MOOC-FRS: A new fusion recommender system for MOOCs[C]. Ieee advanced information technology electronic and automation control conference, 1481-1488.

[5] Su Y, Ding T J, Lue J H, et al. 2017, Applying big data analysis technique to students’ learning behavior and learning resource recommendation in a MOOCs course[C]. international conference on applied system innovation, 1229-1230.

[6] Zhang H, Yang H, Huang T, et al. 2017, DBNCF: Personalized Courses Recommendation System Based on DBN in MOOC Environment[C]. international symposium on educational technology, 106-108.

[7] Oh K J, Lee W J, Lim C G, et al. Personalized news recommendation using classified keywords to capture user preference[J]. In: Advanced Communication Technology (ICACT), 2014 16th International Conference on. IEEE, 1283-1287.

[8] Hu T, Zhang T, et al. 2017, MCRS: A Course Recommendation System for MOOCs[J]. Proceedings of The 2017 International Conference on Advanced Technologies Enhancing Education, 68: 82-85.

[9] Zhang H, Huang T, Lv Z, et al. 2019, MOOCRC: A Highly Accurate Resource Recommendation Model for Use in MOOC Environments[J]. Mobile Networks and Applications, 24(1): 34-46.

[10] Cheng H, Koc L, 2016, Harmsen J, et al. Wide & Deep Learning for Recommender Systems[J]. conference on recommender systems, 7-10.

[11] Peters M E, Neumann M, Iyyer M, et al. 2018, DEEP CONTEXTUALIZED WORD REPRESENTATIONS[J]. north american chapter of the association for computational linguistics, 2227-2237.