The Computer Intelligent Selection of Scientific Research Subjects Through Ensemble Learning for Large-Scale Data Sources and Deep Neural Network

Yan Ma1,2, Lida Zou3, Ke Liu1, Yingkun Han2 and Lei Ma2

1Shandong Electric Power Research Institute, Jinan 250002, China
2State Grid Shandong Electric Power Research Institute, Jinan 250002, China
3Shandong University of Finance and Economics, Jinan 250014, China

*Corresponding author: zoulida@sdufe.edu.cn

Abstract. Selecting a proper scientific research subject is critical for scientific researchers and managers. Scientific researching data are from massive sources and have various attributes. For the problem of subject selection, feature extraction and prediction model play important role in performance optimization. In the paper we introduce ensemble learning method to help find the best fit attributes describing data. Our ensemble learning models include random forests, support vector machine, Boltzmann machine and decision tree. Since the data are from many data sources, we adopt multiple models of deep neural network. An acceleration method is used to reduce the training time as well. Experiments shows that the proposed approach performs better than RNN algorithm both in accuracy ratio and recall ratio. The model selection module and acceleration method help optimize the time cost largely.

Keywords: Select scientific research subjects, ensemble learning, deep neural network, large-scale data sources.

1. Introduction

A scientific research subject is an overview on theoretical consciousness and practical tools for issues and researchers. The selection of scientific research subjects embodies thinking level, theoretical depth and practical ability of selector. Whether the selection is right or not matters in research subject application and researching development. The critical problem of selecting scientific subjects is whether they are important and advanced.

At present, the selection of scientific research subjects is a new applicable demand in the domain of scientific information. Scientific researchers and project managers should be proactive in selecting scientific research subjects. Most similar works are in the book industry. Some techniques such as time series analysis, neural network, collaborative filtering and personalized recommendation are applied. For the works of selection scientific research subjects, most works focus on document retrieval tools.

The scientific and technological data are from different data sources. Each data source has massive data attributes. The data of selecting subjects have high dimensions and large volumes. It is hard to finish accurate feature extraction and subject prediction when just using one kind of technology for
large-scale data sources. Moreover, the features of selecting subjects change largely with time. When
the variance is large and sudden, the false positive rate will increase rapidly. Therefore, we will
propose a Selection method based on Ensemble Learning and multiple structures of deep-learning
network (SEL) in the paper to solve the above problems.

2. Related Work
The selection of scientific research subjects becomes more and more important with the emphasis on
scientific and technological innovation. Book industry has most similar works. The operating progress
of selecting book topics includes information choice, topic design, topic assessment and optimization.
In order to solve the problem of short-term fluctuation and long-term period, data mining and web
mining are introduced.

Time series analysis model is first used in selecting book topic [1], which is based on the data of
book popularity and construct a series of mathematical models. An improved neural network model is
used [2] according to the characteristics of short-term fluctuation, which provides a reliable guarantee
for the profits of publishing units. Aiming for the problem of subjective experience and out of hotspot,
collaborative filtering algorithm is introduced into the selection system of publishing topics.

Many other works [4-6] also explore the application of selecting subjects. They help improve the
managing quality and efficiency of enterprises and universities. However, the above works focus on
function design, developing platform and system framework, which makes system optimization and
selection effectiveness weak. The works in literatures [7-9] research the personalized recommendation
technology and they improve the status of selecting subjects.

University, scientific institutes and intelligence agencies may all have the problem of selecting
scientific research subjects and it is the basis and starting point of scientific research and think-tank
decision. For selecting scientific research subjects, most researches are based on expert experience and
from the prospective of management. The work [10] enumerates dozens of selection cases of scientific
research subjects for ten years and provides the several patterns and characteristics. Scientific
document retrieval system is used as a tool for selection subjects [11]. It helps select the proper
subjects in the aspects of avoiding application of similar projects and determination of researching
ideas.

3. The Framework of Selecting Scientific Research Subjects
To solve the above problems, we propose a selection method of scientific research subjects based on
ensemble learning, which could accurately obtain the fine-grained subjects in a future period. The
method is fit for large-scale data sources. The ensemble learning model is used to abstract the features
and deep-learning technique is adopted to predict the hotspot words. We also introduce the model
selection module to optimize the proposed model. Our framework of selecting scientific research
subjects consists of six models, as shown in Fig. 1. They are data crawler, feature representation,
feature extraction, model training, model selection and prediction modules.

![Figure 1. The selection framework of scientific research subjects](image-url)
Data crawler module is to collect various kinds of data from different data sources, and do the initial pre-processing operations. Feature representation module is to provide a representing method for scientific and technological texts and the texts are periodical. Feature extraction module uses ensemble learning model to extract the critical features in a period and its output is the input of model training and prediction modules. Model training module is to train the selection model based on multiple deep neural networks and generate the fine-grained hotspot words. Model selection module adopts efficiency ratio algorithm to abandon badly-performed models. The badly-performed models have inaccurate prediction results or consume much resource cost. Prediction module is to do the selection of scientific subjects based the trained models and multiple data sources.

4. The Concrete Implementation

4.1. Data Crawler

The module uses web crawler technology to obtain scientific and technological information from technology news website, scientific and technical document database. We assume that the set of the collected data in a period is denoted as \( T_t \), where \( t \) is the number of periods.

4.2. Feature Representation

The weighted TF-IDF algorithm is used to obtain the keyword vectors of \( T_t \) and denoted as \( \alpha_i \), where \( Q \) is the number of keywords in word library, \( b_i \) is the weighted TF-IDF value of one keyword. The concrete computing steps are as follows.

1) Let \( t_i \) represent an article in \( T_t \). We get the TF-IDF value of the \( i \)-th word in \( t_i \) based on standard TF-IDF algorithm, which is denoted as \( b_i^t \). Let \( a_i^t \) denote a keyword of \( t_i \), where \( i \) is the identifier of corresponding keyword \( a_i \) in word library.

2) The downloading or reading volume of \( t_i \) is denoted as \( n_j \) and the citing volume is represented as \( m_j \). The parameter is computed as \( b_i = \sum_{t_i, a_i} (b_i^t + a_i^t \beta) \frac{n_j}{m_j} \), where \( \beta \) is the weight of keyword, \( \bar{n} \) and \( \bar{m} \) are separately the average value of \( n_j \) and \( m_j \) of all the articles in the period.

4.3. Feature Extraction

Since the sources of training data are wide, the inputs of prediction have high dimensions. If we do not select or delete, there exists the curse of dimensionality. For some domains, certain data have nothing to do with the domain. For example, the subject selection in water conservancy is weakly related with media data in physics. Therefore, only the strongly related data are needed for certain domains. Without expert experience, it is complicated to choose data attributes with strong relationship.

This module uses ensemble learning algorithm to do the extraction of features, which is the input of model training and prediction modules. Ensemble learning model is based on the algorithms including decision tree, random forests, support vector machine and deep Boltzmann machines. The module last chooses the best result of all the algorithms. In the paper we give the steps of deep Boltzmann machine. Its structure and parameters are set as follows.

1) Deep Boltzmann machines adopt three-layer restricted Boltzmann machine

2) The first layer is visible cell layer, which is a \( Q \times B \) binary matrix, which is denoted as \( B = \max \{b_i \mid b_i \in \alpha_i, t \in T \} \). In the equation, \( \tau \) is the set of all the periods, \( \eta \) is a positive integer with a large value, \([x]\) is rounding operation.

3) Assume \( v_{i,j} \) is a unit in visible layer. In period \( t \), if \( j = b_i \eta \) stands for \( b_i \in \alpha_i \), then we get \( v_{i,j} = 1 \).
4) The second layer is hidden layer, which is $\beta_Q \sqrt{Q} \times \beta_B \sqrt{B}$ binary matrix. Here $\beta_Q$ and $\beta_B$ are the coefficient larger than 1. The third layer is hidden layer, which is $\beta_Q^2 \sqrt{Q} \times \beta_B^2 \sqrt{B}$ binary matrix.

5) Our deep Boltzmann machine uses the training method of standard restricted Boltzmann machine Layer by Layer.

6) The output of deep Boltzmann machine is set as $X_i$ in period $t$.

4.4. Model Training

Model training module initializes multiple deep neural networks to do the training. We use $V_j = v_{r, r \in [0, R - 1]}$ to denote its structure, where $r$ is an integer, $R$ is the layer of model $j$, and $v_r$ is the number of nodes in $r$-th layer. $R$ and $v_r$ are randomly generated in a certain range. The concrete training method is as follows.

1) First the general deep neural network is used to do the training.

2) For large-scale data sources, we generate a separate keyword set for a fine-grained domain. Each deep neural network corresponds to a fine-grained domain. The details are shown as follows.

For a model, we first delete the inaccurate words from test data. Let err to denote the prediction error. The algorithm deletes the first $n$ words with the largest err value.

For the words with small err value, we generate its new fine-grained hotspot keywords. These words would take part in the following training. The computing method of new hotspot keywords is as follows.

We first compute the parameter $cor(a, b)$, which is computed as $cor(a, b) = \frac{1}{(C(a) + C(b) - C(a, b))} \log \frac{C(a) * C(b)}{C^2(all)} \sum_{j \in J} \frac{a_j * b_j}{\log o_j}$. Here $cor(a, b)$ denotes the relevance of two keywords, where $a, b$ are two words, $C(a)$ is the number of articles including word $a$, $C(a, b)$ is the number of articles including words $a$ and $b$, $C(all)$ is the number of all articles, $J$ is the set of trained articles, $a_j$ is the times of word $a$ appeared in article $j, b_j$ is the times of word $b$ appeared in article $j$ and $o_j$ is the number of all the words in article $j$.

We compute the correlation coefficient of different words $l$. It is computed as $l = \frac{cor(a, b)}{err(a)}$, where $err(a)$ is prediction error of keywords $a$.

The first $n$ keywords with the largest $l$ are added into $W_j$.

4.5. Model Selection

About model evaluation, we use the average error $\overline{err}$ and performance ratio indicators. The performance ratio is denoted as $s$, which is computed as $s = \frac{\overline{err}}{t}$ and $t$ is the training time. To prevent the over-inflation of model structure, we introduce the parameter $\ell$.

When the $\overline{err}$ value of a model is smaller than $\eta$, the model is labelled as A model, where $\eta$ is a threshold. Its fine-grained keywords are deleted from $W$. When $\overline{err}$ is larger than $\eta$ and $s$ ranks in the top 80 percent, the model is labelled as B model. When $\overline{err}$ is larger than $\eta$ and $s$ ranks in the last 20 percent, the model is C model. The algorithm will directly delete the C model. Here we choose all the A models as our training model. For B models, algorithm would randomly select 50 percent of them.

4.6. Prediction

The module is to predict the selected subjects based on the trained models and multiple data sources. Our prediction method is as follows.
1) If the words are the well-trained ones, i.e., \( w_i \in W \), we use the model \( j \) to predict. The model is selected just as \( w_i \in W_j \).

2) If the words are the newly appeared words, i.e., \( w_i \notin W \), we do the prediction based on each model. If any model predicts it as a hotspot keyword, it is regarded as the selected subject.

3) If the number of selected models is massive, we need to compare all the results of each model. It is time-consuming. In the paper we propose an acceleration method. We compute the correlation coefficient between \( w_i \) and each model in \( W_j \). The model with larger correlation coefficient is used in preference. The equation of correlation coefficient between \( w_i \) and \( W_j \) is

\[
\text{corr}(w_i, W_j) = \sum_{w_j \in W_j} \text{corr}(w_i, w_j).
\]

5. Performance Evaluation

In the paper, we propose a Selection approach of scientific research subjects based on Ensemble Learning technique (SEL). The algorithm includes six modules, which are data crawler, feature representation, feature extraction, model training, model selection and prediction modules. In order to accelerate the prediction efficiency, we introduce the model selection module. We call the method without model selection module as SEL-NS algorithm. In the prediction module, we introduce an acceleration method. We call the algorithm without acceleration method as SEL-NA. The algorithm without model selection module and acceleration method is called SEL-NSNA.

In the section, we will evaluate the performance of SEL, SEL-NS, SEL-NA, SEL-NSNA. The comparison indicators are recall rate, accuracy rate and training time. About the benchmark algorithm, we use Recurrent Neural Network (RNN) algorithm [12].

Fig. 2 shows the comparison results on recall rate and accuracy rate between our proposed SEL algorithm and RNN algorithm. We test their performance with the number of hotspot keywords. No matter for recall or accuracy rate, SEL performs better than RNN algorithm. That is because SEL uses multiple deep learning models to predict the selected subjects and RNN cannot suit for all kinds of fine-grained domains through one structure. SEL algorithm performs 65.1% better than RNN in recall rate, and 1.3 times better than RNN in accuracy rate.

Next, we evaluate the performance of model selection module and acceleration method. Fig. 3 gives the accuracy rate of SEL, SEL-NS, SEL-NA, SEL-NSNA algorithms with the number of hotspot keywords. In the figure, SEL and SEL-NA have similar accuracy rate, while SEL-NS and SEL-NSNA have lower accuracy rate. This means that model selection module plays an important role in optimization the accuracy rate. Model selection module could find the well performed models.

Fig. 4 shows the training time of SEL, SEL-NS, SEL-NA, SEL-NSNA algorithms with the number of hotspot keywords. In the figure, we see that the acceleration method is critical for reducing the
running time compared with model selection model. With the increasing number of hotspot keywords, their roles are becoming largely for optimizing time.

Figure 4. The training time of four algorithms

6. Conclusion
In order to solve the caused problem from large-scale data sources, we propose a selection approach of scientific research subjects based on ensemble learning and deep learning techniques. Since the data attributes are massive and variable, in the paper we adopt ensemble learning method including random forests, support vector machine, Boltzmann machine and decision tree, to do feature extraction. The different fine-grained domains have distinct hotspot keywords, and we generate multiple models of deep neural network. The most suitable model is chosen to do the prediction. We also consider the running efficiency of different models through an acceleration method. Our extensive experiments shows that our proposed algorithm performs 65.1% better than RNN in recall rate and 1.3 times better than RNN in accuracy rate.

Acknowledgments
This work was financially supported by Scientific Projects of Shandong Electric Power Research Institute (ZY-2021-16), Science and Technology Plan for Colleges and Universities in Shandong Province (KJ2018BAN046, KJ2018BAN058) and the 2020 annual project of the 13th Five-Year Plan for education science in Shandong Province (Grant No. 2020ZC312).

References
[1] Han Qin. The application of time series analysis on selecting book title [J]. Chinese Academic journal Electronic Publishing House, 2020, 357(3): 28-31.
[2] Kang Baozhong, Lin Hai. The prediction model of selecting publishing titles based on neural network [J]. Information Research, 2018, 12(10): 82-86.
[3] Bian Yuning, Li yeli, Sun Yanxiong. The design and implementation of publishing title selection system based on collaborative filtering algorithm [J]. Journal of Beijing Institute of Graphic Communication, 2018,26(12): 38-41.
[4] Sun Nanjun. The Application Research of Personalized Recommendation Technology in Graduation Project Selection System [D].Xi’an Shiyou university, 2015.
[5] Ji Wentiang, Guo Qingju. Research and application of personalized recommendation elective course in Higher Vocational School [J]. Software, 34(4): 40-41.
[6] Liu Yuju. The graduation design topic selection system based on web design and implementation [J].Computer CD Software and Applications, 2014(9): 261-261.
[7] Zhang Hua. E-commerce Travel Routes Recommended System Based on Data Mining Technology [J].Software, 2013, 34(3): 57-58.
[8] Fu Dajie, Li Jing. Design and Implementation Based on Web Mining for the Recommended Service System of the Educational Resources [J]. Software, 2014, 35(7): 33-35.
[9] Hu Jian, Wang Lijiang. Research on Data Mining in Course Recommendation [J]. Software, 2016, 37(4): 119-121.
[10] Wang Shiwei. The Thought and Selection of scientific subjects about think tank [J]. Journal of Documents and Data, 2019,1(4):11-23.
[11] Han Ling. The important role of scientific document retrieval in scientific subjects selection [J]. Jiangsu Science and Technology Information, 2019,33（11）:10-13.
[12] Y. Bengio, A. Courville and P. Vincent, Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1) 1798-1828.