Abstract: To explore the construction of the ideal paradigm of colleges and universities with artificial intelligence technology in the process of higher personnel training. Using the literature research method, this paper expounds from three aspects: origin, positioning and path. Origin: Artificial intelligence is the reality of technological development in the cultivation of higher talents in the new era, and it is a realistic demand in the cultivation of higher talents, which requires political and ideological empowerment. Positioning: Determining the role of artificial intelligence as an "education assistant", and then clarifying it as a "quasi-subject object", is the basic positioning for the introduction of higher talents at this stage. Paths: 1) Build a teacher-student interconnection model supported by "human-machine collaboration"; 2) Build a smart learning and smart teaching model supported by an "algorithm database"; 3) Establish a "people-oriented" assessment and purpose-led mechanism principle, Deeply improve the quality of the cultivation of higher education personnel in our country's high-level colleges and universities, and promote the final formation of its training pattern.

Keywords: formation, training pattern, people-oriented, education assistant

I. INTRODUCTION

Higher talents themselves have the characteristics of relatively low cognitive ability, poor discipline, and emphasis on skill training rather than theoretical study. Therefore, in order to improve the quality of higher talents in our country[1]. In the new era, in the process of cultivating higher talents in colleges and universities in our country, strengthening the fundamental task of cultivating people with morality. Grinding effect. With the rapid development of science and technology today, artificial intelligence has become an irreplaceable and important part of social life at this stage, especially in promoting the development of modern teaching, improving educational methods, integrating educational resources, and innovating educational concepts[2][3][4]. Educating people, educating people in the whole process, and educating people in all directions are the realistic requirements for the cultivation of higher talents in colleges and universities[5][6]. It is even more crucial to pay attention to the development requirements and development opportunities of the new era brought by artificial intelligence technology. The fundamental task is to find artificial intelligence. The breakthrough and main attack direction of deep integration[7]. In the new era, under the new requirements of my country's higher personnel training, clarify the positioning of artificial intelligence, explore the combination mechanism of artificial intelligence as a technological product in higher personnel training, and apply the new achievements of artificial intelligence in the thinking of my country's higher personnel training. It has important practical significance to be applied in the process of political education[8].
Since the 21st century, in the face of the ever-increasing evolution of technological innovation and international technological competition, how to effectively identify the research frontiers in the field of science and predict the future development of science and technology is crucial. The research frontier hotspot detection based on scientific and technological literature has always been the key content of scientific and technological strategic intelligence[9]. At the macro level, it can provide decision support for national discipline structure planning, think tank construction and fund planning. From the perspective of individual microscopic, grasping and trend tracking of research hotspots is of great significance for scientific researchers to improve scientific research efficiency and scientific research output. Therefore, it has been the focus of attention to effectively capture the information of current active scientific research activities and track the trend of research hotspots[10].

At present, many scholars have carried out effective research on the frontiers of scientific research and hotspots in the field, and have achieved fruitful results. However, at the same time, the detection of research hotspots is mostly based on citation analysis methods, which has the problem of time lag[11]. At the same time, there are many problems such as insufficient word frequency statistical analysis semantics, sufficient evolutionary analysis but insufficient predictive analysis, and coarse-grained time slices cannot effectively reveal the development and evolution of hot topics. The problem makes the research hotspot detection and trend analysis scientifically insufficient[12].

II. METHOD FRAMEWORK

In order to effectively predict and analyze the trend of research hotspots, this paper proposes a research hotspot prediction model based on machine learning algorithm. First, to obtain the WOS core collection abstract data, and then the LDA topic model is used to achieve information extraction from scientific and technological literature[16,17,18]. And topic intensity is used to characterize hotness, and then the cosine similarity theorem is used to establish topic association construction. Finally, machine learning algorithm is used to predict and analyze its future development trend, and the prediction ability of different machine learning algorithms is evaluated and verified. The experimental framework is shown in Figure 1.

![Figure 1 Framework of research hotspot prediction](image)

2.1 Research hot topic detection

The LDA model is proposed to be used as a research hot topic detection tool. The DA topic model can express the
three-layer semantic structure of topic, topic word and document. It uses unsupervised machine learning to extract
hidden topic information and express the weight of the topic word. Research hotspots represent the research
enthusiasm and intensity of a certain subject within a certain period of time, and can be expressed by the weight of
keywords or subject words. The larger the weight, the more popular the research topic. Using machine learning
algorithms to predict research popularity first requires obtaining the topic intensity value of each sub-period[19,20].
This paper proposes the following topic strength indicators for hot research fronts:

\[ TII_i^Z = \sum_{i=1}^{n} \text{weight}(k_i) \] (1)

Among them, \( \text{weight}(k_i) \) represents the weight ratio of the subject words; \( \sum_{i=1}^{n} \text{weight}(k_i) \) reflects the
cumulative weight value of the subject words, and \( TII_i^Z \) is the weight value of the subject. The effective detection
and prediction of this indicator can accurately grasp and locate the development context of popular research fronts.

2.2 Topic Association Construction

After fine-grained identification of sub-period research topics, it is necessary to explore the internal and external
correlations of topic clusters in different time segments, so as to obtain the precursor and successor relationships of
different topics in different periods to form a dynamic topic chain. In this paper, the similarity of time series topics
is calculated based on the method of cosine similarity, and the development context of topics based on time series
development is constructed, which provides experimental preparation and foundation for subsequent prediction
analysis based on machine learning algorithms. A fixed threshold is set to determine the similarity. If the similarity
is greater than the threshold, it means that the hot topics in the two time dimensions are the evolution and changes
of the same topic. The formula is as follows:

\[ \text{Sim}(\text{Topic}_i, \text{Topic}_j) = \cos \theta = \frac{\sum_{k=1}^{n} w_k(\text{Topic}_i) \times w_k(\text{Topic}_j)}{\sqrt{\sum_{k=1}^{n} w_k^2(\text{Topic}_i) \times \sum_{k=1}^{n} w_k^2(\text{Topic}_j)}} \] (2)

Among them, the numerator represents the dot product of the two subject vectors, and the denominator represents
the product of the modulus of the two subject vectors.

III. PREDICTION MODELS

Three machine learning algorithms commonly used in time series forecasting research with high accuracy are
selected as the hotspot trend forecasting models in this experimental research, namely BP neural network, SVM and
LSTM model[21].

3.1 BP neural network structure prediction model analysis.

The corresponding weight value used for propagation parameter adjustment. neural node i :
The model input obtained through the hidden layer node is:

\[ O_i^w = g \left( H_i^w \right) = g \left[ \sum_{j=1}^{J} W_{ij} g \left( \sum_{k=1}^{K} W_{jk} X_k \right) \right] \]  \hspace{1cm} (4)

The error function is:

\[ E^w(w) = \frac{1}{2} \left( x_i^w - y_i^w \right)^2 = \frac{1}{2} \left[ x_i^w - g \left( \sum_{j=1}^{J} W_{ij} \right) g \left( \sum_{k=1}^{K} W_{jk} X_k \right) \right]^2 \]  \hspace{1cm} (5)

when the BP neural network faces the complex optimization objective function, when the neuron output approaches the real value, the training effect is poor and it is easy to fall into the local optimum.

3.2 SVM Prediction Model

Compared with traditional neural network algorithms such as BP neural network, the SVM model adopts the optimal structure risk and its generalization ability has always been one of the advantages of this model. For a given sample \((x_i, y_i) (i = 1, 2, 3, \ldots N)\), \(N\) is the sample size, \(x_i\) is the input vector, and \(y_i\) is the output target. The SVM model uses a high-dimensional mapping feature space \(R^n \rightarrow R^m\), and then uses a linear function to approximate the function in the feature space:

\[ y = f(X) = \langle W, \varphi(X) \rangle + b \]  \hspace{1cm} (6)

In the formula, \(W, f(X)\) is the m-dimensional vector data, \(b\) is the function threshold, and \(y\) is the function value after dot product processing. According to the statistical theory SVM minimizes the objective function to obtain the fitting regression function formula:

\[ \min W, b : \frac{1}{2} W^2 + c \sum_{i=1}^{n} \left| y_i - \langle W, \varphi(x_i) \rangle - b \right| \]  \hspace{1cm} (7)

In the formula, \(c\) represents the penalty coefficient of the control model loss \(1/2 w^2\) and the complexity of the training model, and \(i = 1, 2, \cdots, n\) represents the number of SVM points. The kernel function can be used to realize the high-dimensional mapping of the feature space of the data, and then output the model prediction time series results without affecting the computational complexity.

\[ g(x_i) = w^x \Phi(x_i) + b = \sum_{j=1}^{m} \alpha_j \left[ \Phi(x_i) \Phi(x) \right] + b = \sum \alpha_i k(x_i, x) + b \]  \hspace{1cm} (8)

SVM has strong small sample learning ability and nonlinear fitting ability, it is not easy to fall into local optimum.
and parameter setting is relatively simple, so it has been widely used in the field of machine learning time series prediction, but it is in the forefront of information science research. There are few applications in detection and predictive analysis, so this paper selects the SVM model as a class of machine learning algorithms to discuss and analyze[22,23].

IV. EMPIRICAL RESEARCH

4.1 Experimental Platform

Hardware: Window10 operating system, Genuine Intel(R) CPU@1.70GHz, 8GRAM

Software: Anaconda, Keras deep learning framework based on Tensorflow (GPU version) backend, Rapidminer

4.2 Data set and preprocessing

Database: Wos core paper collection data of papers

Time span: 1965 to 2017

Search: Keyword="Genetic Engineering"

Search results: 2764 items.

4.3 Topic Recognition Experiment

In this paper, the LDA model is used for topic identification. LDA is a Bayesian probability distribution model with a three-layer structure including document set layer, topic layer and feature word layer. It simulates large-scale document generation through probability statistics and parameter fitting[22,23,24]. By extracting the subject words with practical significance in the scientific and technological literature, we can deeply mine the hidden subject information topology structure contained in the text data.

The number of topics in the document set is a hyperparameter, and the number of topics in multi-source information data needs to be determined before topic identification. Perplexity is a commonly used indicator to measure the pros and cons of a language model. [25] proposed the complexity indicator and defined the complexity of a topic model for a document set with M documents as:

$$\text{perplexity} (D_{\text{test}}) = \exp \left\{ \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} \log N_d} \right\} \quad (9)$$

Among them, M is the number of documents in the document set, \( P(W_d) \) is the probability that the PLDA model generates the d-th document, and \( N_d \) is the number of words. When \( \text{perplexity} (D_{\text{test}}) \) is the smallest, the topic has a better semantic expression effect, and the topic-document mapping is established to determine the document. The number of topics in the set. In this paper, experiments are carried out on the dynamic correspondence between the number of topics and the complexity, and the final number of topics is selected as 100 after the
experimental topic step size, as shown in Figure 2.

![Figure 2 Corresponding curve of number of topics and model complexity](image)

The cosine similarity is used to quantitatively establish the topic correlation of different sub-periods, and then establish a whole dynamic time series topic chain in the time series dimension, and the time span is from 2003 to 2017. When the similarity threshold is set to 0.5, the subject evolution and correlation are better, and the transition and evolution characteristics of different subject types in the time window can be more completely expressed. In the similarity calculation in Table 1, topic types with a similarity threshold greater than 0.5 can be selected to establish associations, that is, Topic_1 in 2014 developed into Topic_2 in 2015 and then became Topic_7 (2016), and in the same way, the topic development changes of 10 topics can be obtained.

The establishment of the topic dynamic time series chain provides a theoretical basis for the follow-up machine learning algorithm prediction research, and the specific topic intensity value can be obtained by using the LDA model experiment to obtain the weight of the topic word to represent the research popularity, and then detect the development and change rules of the research popularity of different topics in the time series.

### Table 1 Topic similarity calculation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic_1</td>
<td>Topic_2</td>
<td>0.6011</td>
<td>Topic_2</td>
<td>Topic_7</td>
<td>0.5211</td>
<td>……</td>
</tr>
<tr>
<td>Topic_1</td>
<td>Topic_9</td>
<td>0.3770</td>
<td>Topic_2</td>
<td>Topic_2</td>
<td>0.4343</td>
<td>……</td>
</tr>
<tr>
<td>Topic_1</td>
<td>Topic_4</td>
<td>0.3011</td>
<td>Topic_2</td>
<td>Topic_9</td>
<td>0.3411</td>
<td>……</td>
</tr>
<tr>
<td>Topic_1</td>
<td>Topic_0</td>
<td>0.1515</td>
<td>Topic_2</td>
<td>Topic_0</td>
<td>0.1515</td>
<td>……</td>
</tr>
<tr>
<td>Topic_1</td>
<td>Topic_2</td>
<td>0.0011</td>
<td>Topic_2</td>
<td>Topic_1</td>
<td>0.0155</td>
<td>……</td>
</tr>
<tr>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
</tr>
</tbody>
</table>
4.4 Research Hotspots

The relative error indicator (RE, Relative Error) is used to describe the prediction effect of the model. The formula is as follows, where represents the relative error, \( y_t \) represents the true value, and \( \hat{y}_t \) represents the model predicted value. The test set in this paper is divided into 10 topics and the topic strength value of each topic in the last five years. The relative error RE is obtained for each topic in different sub-periods and the mean value is processed to obtain the average prediction accuracy of the topic. Partly based on the comparison of prediction accuracy of different machine learning algorithms, see Table 2. Taking Topic0 as an example, the 2013-2017 topic heat value was predicted and analyzed by using BP neural network, SVM and LSTM model, and the average relative error of the topic was finally obtained. 15.69%, 12.98% and 10.75%, it can be seen that the prediction accuracy of the LSTM model for this topic is higher, the prediction effect of the SVM is lower than that of the LSTM model but higher than that of the BP neural network prediction model, but for Topic2, the highest prediction accuracy is the support vector. The prediction accuracy of the LSTM model and the BP neural network has a small difference of 13.08% and 14.25%, respectively.

\[
RE = \frac{|y_t - \hat{y}_t|}{y_t} \tag{10}
\]

<table>
<thead>
<tr>
<th>Theme</th>
<th>Actual value</th>
<th>BP neural network</th>
<th>SVMs</th>
<th>LSTM model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Predictive value</td>
<td>Predictive value</td>
<td>Predictive value</td>
</tr>
<tr>
<td>Topic0</td>
<td>2013</td>
<td>870</td>
<td>737.75</td>
<td>15.31</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>176</td>
<td>203.89</td>
<td>15.49</td>
</tr>
<tr>
<td></td>
<td>2015</td>
<td>411</td>
<td>364.01</td>
<td>11.21</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>639</td>
<td>757.39</td>
<td>18.33</td>
</tr>
<tr>
<td></td>
<td>2017</td>
<td>881</td>
<td>730.45</td>
<td>17.08</td>
</tr>
<tr>
<td></td>
<td>Topic Topic0 Average Relative Error</td>
<td></td>
<td></td>
<td>15.68</td>
</tr>
<tr>
<td>Topic1</td>
<td>2013</td>
<td>708</td>
<td>839.45</td>
<td>18.41</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>1.33</td>
<td>1031</td>
<td>1232.38</td>
</tr>
<tr>
<td></td>
<td>2015</td>
<td>691</td>
<td>546.61</td>
<td>20.79</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>990</td>
<td>841.65</td>
<td>15.08</td>
</tr>
<tr>
<td></td>
<td>2017</td>
<td>1079</td>
<td>1311.80</td>
<td>21.69</td>
</tr>
<tr>
<td></td>
<td>Topic Topic1 Average Relative Error</td>
<td></td>
<td></td>
<td>19.03</td>
</tr>
<tr>
<td>Topic2</td>
<td>2013</td>
<td>508</td>
<td>508</td>
<td>417.26</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>228</td>
<td>266.42</td>
<td>17.38</td>
</tr>
<tr>
<td></td>
<td>2015</td>
<td>141</td>
<td>124.25</td>
<td>12.50</td>
</tr>
</tbody>
</table>
In order to effectively express the prediction effect of machine learning algorithms, it is necessary to calculate the average relative error for 10 topics and obtain the model prediction accuracy, that is, the difference between 100% and relative error. The experimental analysis shows that the prediction accuracy of all topics based on the BP neural network model is 83.64%, and the prediction effect of the machine learning algorithm prediction model is the worst. The prediction accuracy of the SVM prediction model and the ISTM model are 88.28% and 89.10% respectively, the prediction effect is similar, in which the prediction accuracy based on the STM model is slightly higher, and the prediction stability based on the long short-term memory neural network and the SVM model is relatively good.

<table>
<thead>
<tr>
<th>Year</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Type 1</td>
<td>235.88</td>
<td>397.80</td>
<td>435.75</td>
<td>215.74</td>
<td>278.90</td>
</tr>
<tr>
<td>Average</td>
<td>14.24</td>
<td>10.97</td>
<td>13.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. CONCLUSION

Artificial intelligence as an "education assistant" is the proper meaning of the construction of the training of sports talents. However, we must also realize that the cultivation of talents is fundamentally value-oriented education, and artificial intelligence is the means to realize this value orientation. In the field of IPE and training of sports talents, ideological progress is the basic direction of talent cultivation. Moral progress is the proper purpose of social development. Therefore, artificial intelligence as a result of technological progress is to make up for the shortcomings of existing for the purpose of talent training, obey the overall situation, and promote the final formation of the training pattern for sports talents in colleges and universities.
REFERENCES


