Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2628–2635, DOI 10.12694/scpe.v25i4.2920

# ARTIFICIAL INTELLIGENCE RETRIEVAL SYSTEM BASED ON COMPUTER BIG DATA TECHNOLOGY

HONGHUA YU*

**Abstract.** In order to achieve automatic retrieval of library literature, the author proposes a robot retrieval method based on intention recognition for library literature automatic retrieval. Firstly, propose a robot retrieval approach; Then, the keywords of the user's question are automatically extracted, and the intention of the user's question is identified. Finally, combined with intention recognition, the semantic similarity between the intention recognition literature and the search literature is calculated. The latest results indicate that: The accuracy of keyword automatic extraction based on KL and TF-IDF for each sentence keyword is 98.3%, indicating that this extraction method performs better. The retrieval accuracy of the automatic retrieval robot reaches 98%, indicating that this method can accurately achieve accurate retrieval of user intentions and candidate literature.

**Key words:** Artificial intelligence, Automatic retrieval, Intentional identification, Semantic similarity calculation

**1. Introduction.** Currently, as we enter a new stage of social development, people can search through the internet to obtain the information they need [1]. Although online information retrieval has enormous advantages compared to traditional paper-based information retrieval methods, there are still problems such as information classification defects, partial information invalidity, huge amount of information, and difficulty in distinguishing true from false [2,3]. On this basis, using artificial intelligence for information classification can provide true and accurate information according to users' habits, thereby promoting information retrieval. Network information retrieval is actually based on the network as a platform, allowing users to search for relevant information using network search engines.

By using allocation storage technology, massive amounts of data information can be distributed to corresponding servers. For users, they can use terminals to query and view pre stored data. Therefore, all information can be retrieved and utilized on the Internet, and providing information search methods or means for network users can be called network information retrieval. Artificial intelligence is a type of technology that allows robots to simulate and perceive people's feelings during the process of doing things through machinery, and then make the right decisions, thereby giving them a specialized technology to solve problems. Therefore, the essence of artificial intelligence also revolves around the human mind. Its biggest feature is to reconstruct and apply knowledge in a relatively complete logical system based on a correct understanding of the problem.

At present, artificial intelligence can be divided into the following schools: distributed school, cognitive school, connected school, logical school, and knowledge engineering school. Although the research of different schools is different, the goal in the construction and function of artificial intelligence is the same, that is, artificial intelligence should be composed of an intermediate database, interpreter, knowledge collector, user interface, knowledge base, and inference engine. Taking the knowledge base as an example, it is an important component of the artificial intelligence system and an important storage technology. There are facts, information, common sense, and rules in the knowledge base. Some specific systems also include databases. Next is the inference engine, also known as the inference engine, which includes control strategies and various types of task searches. As a special database, it plays a very important role in providing query support. The content of the user interface includes the transmission of the system and related information. The inference engine is a bridge connecting external and internal information, which can not only display the final processing effect to the user, but also transmit the user's wishes to the computer. In this case, the use of non natural language can alleviate the psychological pressure of users and play a role in storing intermediate results and data during work and

* Jilin University, Changchun, 130000, China. (yuhonghua_main@163.com)

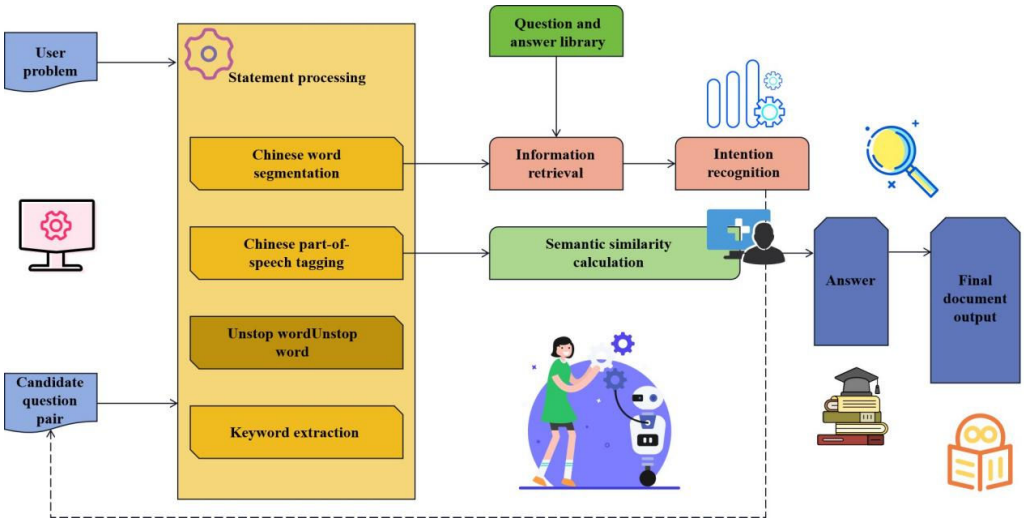


Fig. 2.1: Retrieval Ideas of Library Document Automatic Retrieval Robot

logical operations. In practical use, the system will display the problem on the blackboard and present the initial status of the problem. Then, the expert system will analyze the information in the knowledge base and analyze the information on the blackboard. If necessary, we will also seek advice from customers to supplement and address knowledge gaps. Therefore, in a sense, the blackboard can also be seen as a dynamic knowledge foundation, playing a crucial role in the operational process. Finally, there is a knowledge collector, whose main function is to update the actual operation of the knowledge base to ensure the stable operation of the entire system. The task of the analyzer is to answer users' questions and provide them with operational paths related to the system's results. In short, artificial intelligence is the use of scientific and virtual methods to solve human problems.

With the rapid development of natural language, utilizing natural language to achieve autonomous interaction between humans and robots has become a trend in artificial intelligence [4]. Some scholars have also conducted extensive research on the application of natural language in robot interaction scenarios. For example, Pang Hui proposed using natural language understanding technology to achieve interaction with customer service robots, thereby achieving control of customer service robots; Miao Haifeng proposed the use of natural language object retrieval technology to meet the requirements of different users for robots [5]. In the field of literature retrieval and intelligent robots, combining natural language with different retrieval methods is an important way to improve their application scenarios. However, in library robot retrieval, although natural language and literature retrieval methods are used, the user's intentions are ignored. Therefore, adding more dimensional factors to improve the accuracy of library robots in literature retrieval has become a hot topic and an innovative point of this study. Therefore, based on the above analysis, the author proposes an intention recognition library literature automatic retrieval robot to better promote the intelligence of university libraries [6,7].

## 2. Methods.

**2.1. Automatic Retrieval Robot Retrieval Ideas for Image and Literature.** The robot literature automatic retrieval constructed in this study is based on the FAQ question answering system, and intention recognition is added to this foundation to better improve the accuracy of literature automatic retrieval. The specific idea is shown in Figure 2.1 [8].

As shown in Figure 2.1, this robot retrieval mainly includes three modules: statement processing, information retrieval, and literature push [9]. The main function of sentence processing is to segment students' search questions and candidate question and answer pairs, label part of speech, remove stop words, and extract key-

words; The main function of the information retrieval module is to retrieve literature categories that are highly relevant to student search questions from the stored literature dataset, and output candidate literature; The literature push module includes intention recognition and semantic similarity calculation. Intention recognition filters candidate literature that has different intentions from the student's search content; Semantic similarity calculation calculates the similarity between candidate literature and student search content to obtain candidate literature with high similarity, and ultimately outputs the final literature by the library's automatic retrieval robot.

**2.2. Design of automatic literature retrieval module for intention recognition.**

*(1) The process of extracting keywords from user questions.* In order to achieve automatic retrieval of literature by robots, it is first necessary to extract keywords from user questions in order to correctly understand user needs. Therefore, various keyword automatic extraction methods such as KL divergence, TF-IDF, word length and part of speech are introduced [10]. KL divergence value is a relative entropy that reflects the degree of deviation between two different probability distributions P and Q, and its calculation formula can be expressed as:

$$D_{KL}(P|Q) = \sum_i P(i) \log \frac{Q(i)}{P(i)} \tag{2.1}$$

If P and Q represent the distribution of continuous random variables, the calculation formula for KL divergence can be expressed as:

$$D_{KL}(P|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx \tag{2.2}$$

Consider whether two texts are similar as two different probability distributions, and label similar ones as 1 and dissimilar ones as 0. Calculate the degree of deviation between the two probability distributions using KL divergence. If each word in the text is treated as a separate one-dimensional feature, the KL divergence calculation result represents the importance score of each feature [11].

TF IDF: TF represents the frequency of a word appearing in a given text, abbreviated as word frequency, and its statistical formula is:

$$TF_{i,j} = \frac{n_{i,j}}{\sum\limits_k n_{k,j}} \tag{2.3}$$

In the formula, $n_{i,j}$ represents the frequency of the i-th word appearing in the jth text; $\sum\limits_k n_{k,j}$ represents the total number of words contained in the jth text; The higher the frequency of a word, the greater the likelihood that it is a keyword in the jth text [12]. However, in long texts and documents, some words may have high frequency but are not keywords, which can easily lead to incorrect keyword extraction. Therefore, based on the overall corpus, inverse document frequency (IDF) is introduced to calculate the universal importance of each word. The calculation formula can be expressed as:

$$IDF_I = \log \frac{D+1}{\{j : t_i \epsilon d_j\} + 1} + 1 \tag{2.4}$$

In the formula, D represents the total number of texts; $t_i$ represents the i-th word in the text; $d_j$ represents text containing $t_i$ ; J represents the quantity containing $t_i$ text. The calculation formula for comprehensive TF-IDF keyword extraction can be expressed as:

$$TF - IDF_i = TF_{i,j} * IDF_i \tag{2.5}$$

TF-IDF combines the advantages of word frequency and inverse frequency, making it simple and effective for keyword extraction and suitable for library automatic retrieval robots.

Word length: The length of a word itself is called a word length feature, and the word length of a keyword is usually 2-3, which has good distinguishability compared to other words [13].

Part of speech: Based on the distribution of part of speech of keywords, the part of speech is transformed into multidimensional Boolean features and selectively combined. This method not only improves keyword distinguishability, but also solves the problem of high and sparse dimensions of part of speech features, thereby improving the performance of automatic extraction of text keywords. Automatic keyword extraction preprocesses user search problems, and extracts keyword features using KL divergence, TF-IDF, word length, etc. to obtain feature vectors and labels; Secondly, XGBOOST algorithm is used to construct multiple feature keyword automatic extraction models, and the scores of each word are calculated; Finally, output the words with the highest scores as keywords [14].

*(2) Intention recognition.* Intention recognition is a method of identifying requirements based on the content of a user's search question and dividing their types of requirements. Intention recognition is a multi classification model that extracts intent features. This model takes user search content as input and preprocesses student search content using Chinese word segmentation, part of speech tagging, and removal of stop words; Subsequently, domain word judgment and sentence Word2vec2 were used to extract features from the search content, and they were matched with the intentions in the intention classification template; Finally, Softmax classifier was used to achieve intention recognition [15].

*(3) Semantic similarity calculation.* Semantic similarity calculation mainly calculates the similarity score between the user's search content and the candidate literature after intention recognition and classification, and takes the highest score literature as the final output of the robot. Semantic similarity calculation is based on the input of two sentences of search content and candidate literature for users with labels 1 or 0, where 1 represents similarity between the two sentences and 0 represents dissimilarity. Firstly, two sentences are preprocessed, followed by feature extraction using methods such as domain word judgment and sentence Word2vec cosine distance; Finally, a semantic similarity model is constructed using the XGBOOST algorithm, and the similarity score between user search content and candidate literature is calculated to enable the library literature automatic retrieval robot to obtain the final output answer.

*(4) The overall process of robot literature automatic retrieval.* Based on the above implementation, the overall steps for designing automatic literature retrieval are as follows: Students input search questions into automatic search robots [16]; The automatic retrieval robot first performs word segmentation, part of speech tagging, removal of stop words, and keyword extraction preprocessing on student search questions; The information retrieval module retrieves candidate literature containing keywords in the database based on the extracted keywords; Identify the intentions of student search questions and filter out candidate literature that does not match the student's search intentions; Calculate the similarity between student search questions and candidate literature filtered through intention recognition through semantic similarity calculation, and obtain the literature with the highest similarity; Take the candidate literature with the highest similarity as the final output result.

### 2.3. Application Countermeasures of Artificial Intelligence in Network Information Retrieval in the Era of Big Data.

*(1) Network Intelligent Knowledge Service System.* Usually, a network intelligent knowledge service system consists of four parts, including a knowledge processing system, a knowledge collection system, a knowledge service system, and a knowledge base.

The knowledge processing system is to transfer the knowledge downloaded from the network to intelligent communication devices, classify the knowledge, search based on keywords, and finally transfer the approved knowledge to the database. Generally speaking, there are four points to streamline knowledge processing [17]. Firstly, intelligent knowledge classification. Implement classification based on relevant standards by combining data types and content, and then pass it into the intelligent matching process. Secondly, the intelligent knowledge matching process. In this process, knowledge in the database can be classified, and then downloaded using the internet. The downloaded knowledge can be compared and analyzed to avoid data overlap. The intelligent matching results can be transmitted to the intelligent update process. Once again, intelligent knowledge updates. Use the matching results to clarify which knowledge base the corresponding information is stored in, and then replace the original data or overlap the newly downloaded information with the original

information to form a new information concept. Finally, intelligent knowledge cleaning. The intelligent database needs to be cleaned regularly, and internal knowledge and network knowledge matching work should be done to achieve internal knowledge replacement and cleaning, so as to ensure the smooth operation of the entire system.

In order to enhance the richness of internal knowledge in the database, it is also necessary to do a good job in knowledge collection and updating. In other words, it means knowledge collection and processing, achieving the transformation and supplementation between knowledge and knowledge. Usually, the collection system involves two parts, one is the collection of printed knowledge, and the other is the collection of data knowledge. Among them, the collection of printed knowledge mainly involves scanning paper knowledge and using artificial intelligence technology to transform data, transforming text indices into digital form. The data knowledge system collects it, making network data resources more abundant. In the process of information collection, there are four processes that need to be followed. Firstly, site mirroring refers to downloading the internal information of a site like a mirror, usually copying some websites with abundant resources, and comprehensively copying the content into the system to improve information collection efficiency. Secondly, intelligent information monitoring. Intelligent monitoring of target information can be achieved to ensure the rationality of target information changes, and internal system information can be exchanged with the actual situation. Thirdly, intelligent resource discovery. The system can search for new resources and discover new data that meets network requirements. In the event of new data appearing, the system can automatically collect it. Fourthly, intelligent knowledge resource transformation. In this process, digital resource collection and allocation can be achieved, and a new meaning can be formed afterwards.

The preservation system is an indispensable part of the knowledge base and also serves as a basis for ensuring the quality of information retrieval. Usually divided into hardware retrieval, software retrieval, and system retrieval. Among them, hardware serves as the basis for storing data, while software is an information storage management system that can ensure efficient storage of various information. The retrieval system relies on hardware storage and software management, which is the foundation of intelligent retrieval systems. On this basis, the entire system can be redeveloped to make it more comprehensive and provide users with good services.

*(2) Intelligent Agent Technology.* In the context of big data, IA intelligent agent technology, as a major component in the field of intelligence, has emerged with the comprehensive development of the internet. IA intelligent agent technology has been widely applied in various fields and is also a key research topic in China's intelligent technology exploration. By applying IA intelligent agent technology to network information retrieval functions, various problems in traditional retrieval can be addressed, effectively improving the level of network information retrieval. The application structure of intelligent technology in information retrieval applies intelligent agent technology to information retrieval work, which can form a new retrieval agent tool that sets corresponding services according to the needs of each user. By combining the user's network application habits and needs, information retrieval methods can be classified. For example, if the user will frequently collect relevant information for a period of time, under the action of intelligent agent technology, the user's collection situation can be recorded, and the type of information that the user is interested in can be analyzed. At the same time, corresponding reference schemes can be set based on the key situation of other users in collecting this type of information, provide a basis for users to search for relevant information. For example, centralized push of relevant information, etc. Under the influence of intelligent agent technology, relevant information can be pushed according to user needs, which helps users better understand themselves. Assuming that the user is not satisfied with the pushed information, they can record the feedback results in the system based on the actual situation, and the system will automatically make changes to make the system more accurate in the next information matching process. The emergence of intelligent retrieval tools can better meet user needs and make network information retrieval work more rational and intelligent.

Firstly, network management. This function allows users to monitor the distribution of network resource points and promptly handle resource failures when searching for corresponding site resources. When downloading the same data, it is convenient for users to choose sites with wide resource areas and ideal server operation, and select corresponding information transmission networks based on their actual situation, this can reduce the significant consumption of time resources by users due to network congestion. Secondly, information manage-

Table 3.1: Distribution of keyword automatic extraction training and testing sets

| Data set | Sentence | Terms | Keyword |
|---|---|---|---|
| Training Set | 4000 | 17000 | 3500 |
| Test Set | 1000 | 6000 | 1000 |

Table 3.2: Automatic Extraction Results of Two Kinds of Keywords

| Model | Accuracy (%) |
|---|---|
| KL+TF-IDF+XGBOOST | 98.325 |
| Logistic | 95.216 |

Table 3.3: Literature search results for semantic similarity calculation

| Model | Accuracy (%) |
|---|---|
| Semantic Similarity Model Test Results | 98 |

ment. This service function can clearly understand the distribution of user required data in the network, and provide a basis for keyword search by users. It can select target information and delete invalid information. In addition, based on user application habits, information push plans are set to promote interface optimization, allowing users to choose corresponding interfaces according to their own network application habits, making intelligent technology more personalized and targeted.

**3. Experimental verification.**

**3.1. Experimental Data and Evaluation Indicators.** Extract 5000 sentences from the user search question set, and after preprocessing with Chinese word segmentation, annotation, and removal of stop words, obtain 23000 words, including 4500 keywords. Divide the above data into training and testing sets, as shown in Table 3.1.

**3.2. Automatic keyword extraction results.** In order to verify the performance of the keyword automatic extraction method designed in this study, both KL+TF-IDF based keyword automatic extraction and logistic regression based keyword automatic extraction were trained and tested to obtain the accuracy of extracting N keywords from each sentence. The test results are shown in Table 3.2 [18].

From Table 3.2, it can be seen that the accuracy of keyword automatic extraction based on KL and TF-IDF for each sentence keyword is 98.3%, indicating better performance of this extraction method.

**3.3. Literature retrieval results based on intention recognition.** Classify the keywords extracted above and obtain a total of 175 different user intentions. In order to verify the performance of the intention recognition model, training and testing were conducted on the basis of the obtained data, and the test results were evaluated through evaluation indicators such as accuracy, recall, and F1 value. The evaluation results are shown in Figure 3.1.

Analyzing Figure 3.1, it can be seen that the accuracy of the original intention recognition model is close to 95%, indicating that it can improve the accuracy of library automatic retrieval robots in identifying user intentions.

**3.4. Accuracy of Literature Retrieval Based on Semantic Similarity Model.** Using accuracy as the evaluation indicator for model performance, semantic similarity was used to calculate the test results of 100 users on 50 candidate literature, as shown in Table 3.3 [19].

As shown in Table 3.3, the accuracy of semantic similarity calculation reaches 98%, indicating that this method is good and suitable for library literature automatic retrieval robot systems.

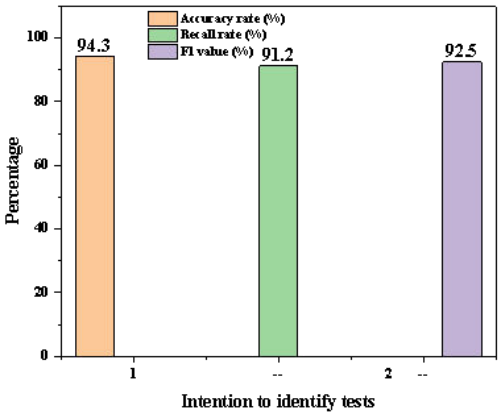


Fig. 3.1: Intention Recognition Test Results

Table 3.4: Automatic retrieval results of library literature retrieval robots based on intention recognition

| Model | Accuracy (%) |
|---|---|
| A Robot System for Library Document Automatic Retrieval Based on Intention Recognition | 98 |

**3.5. System Verification.** In order to further verify the retrieval effect of the automatic retrieval robot, the results of 100 students' automatic literature retrieval were collected, as shown in Table 3.4. As shown in Table 3.4, the designed automatic retrieval robot has a retrieval accuracy of 98%, indicating good performance [20].

**4. Conclusion.** In order to achieve automatic library literature retrieval based on artificial intelligence, the author designed a library literature automatic retrieval robot retrieval method based on intention recognition, and tested the retrieval method. The results show that the automatic retrieval robot has a retrieval accuracy of 98% for user search content. However, due to time and manager constraints, the classification of students' search intentions in this study is not detailed enough, and the design of the semantic similarity calculation module is not yet perfect, requiring further exploration and improvement.

REFERENCES

[1] Tao, Z. , Jun, B. S. , & Bai, R. X. . (2021). Research on marketing management system based on independent erp and business bi using fuzzy topsis. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology0756(4), 40.
[2] Liu, H. , & Ko, Y. C. . (2021). Cross-media intelligent perception and retrieval analysis application technology based on deep learning education. International journal of pattern recognition and artificial intelligence123(15), 35.
[3] Cai, W. . (2023). Innovation and path of teacher literacy in basic education in the era of artificial intelligence. Region - Educational Research and Reviews, 5(3), 55-57.
[4] Bai, X. , & Li, J. . (2021). Personalized dynamic evaluation technology of online education quality management based on artificial intelligence big data. Journal of Intelligent and Fuzzy Systems5467(3), 1-10.
[5] Peng, J. . (2021). Oil painting material collection system based on artificial intelligence. Journal of Physics: Conference Series, 1852(2), 022029-.
[6] Yao, Q. , Liu, Y. , Guo, C. , Ao, C. , & Xu, Z. . (2021). Research on fault warning of marine diesel engine cooling system based on deep belief network. Journal of Physics Conference Series, 1750(890), 012066.
[7] Yin, Y. . (2021). Research on ideological and political evaluation model of university students based on data mining artificial intelligence technology. J. Intell. Fuzzy Syst., 40(2346), 3689-3698.

[8] Li, Y. . (2021). Research on the construction of tcfl resource database system based on artificial intelligence. Journal of Intelligent and Fuzzy Systems5468(6), 1-12.
[9] Zhang, K. , Chen, K. , & Fan, B. . (2021). Massive picture retrieval system based on big data image mining. Future Generation Computer Systems, 121(3),578.
[10] Chen, H. , Xie, J. , Wang, S. J. , Ramanathan, S. , & Mutegeki, R. . (2021). Research on intelligent management system of meteorological archives based on big data framework. Advances in Data Science and Adaptive Analysis: Theory and Applications146(3/4), 13.
[11] Huang, M. , Liu, S. , Zhang, Y. , Cui, K. , & Wen, Y. . (2021). Research on the university intelligent learning analysis system based on ai. Journal of Intelligent and Fuzzy Systems45(31), 1-10.
[12] Wei, Q. , & Qingna, L. . (2021). Construction of cultural industry development factor model based on factor analysis, artificial intelligence and big data. Microprocessors and Microsystems, 82(2), 103880.
[13] Jiang, S. . (2021). Research on big data audit based on financial sharing service model using fuzzy ahp. Journal of Intelligent and Fuzzy Systems, 40(4), 1-10.
[14] Wu, H. D. , & Han, L. . (2021). A novel reasoning model for credit investigation system based on fuzzy bayesian network. Procedia Computer Science, 183(19), 281-287.
[15] Dong, J. , Meng, W. , Liu, Y. , & Ti, J. . (2021). A framework of pavement management system based on iot and big data. Advanced Engineering Informatics, 47(2), 101226.
[16] Yunita, A. , Santoso, H. B. , & Hasibuan, Z. A. . (2021). Research review on big data usage for learning analytics and educational data mining: a way forward to develop an intelligent automation system. Journal of Physics: Conference Series, 1898(1), 012044 (13pp).
[17] Pence, H. E. . (2022). Future of artificial intelligence in libraries. The Reference Librarian, 63(3213), 133 - 143.
[18] Yao, W. , & Li, N. . (2021). Research on agricultural products logistics and supply chain system based on computer big data model. E3S Web of Conferences, 253(46), 02035.
[19] Liu, Y. , Wang, Z. , Pan, Y. , & Zuo, Y. . (2021). Research on intelligent monitoring and early warning of electric power safety based on artificial intelligence technology. Journal of Physics: Conference Series, 1748(5), 052046 (5pp).
[20] Tan, L. , & Ran, N. . (2023). Applying artificial intelligence technology to analyze the athletes' training under sports training monitoring system. International Journal of Humanoid Robotics, 20(06),56.

*Edited by:* Zhigao Zheng
*Special issue on:* Graph Powered Big Aerospace Data Processing
*Received:* Nov 28, 2023
*Accepted:* Dec 25, 2023