

# Fusion mode of multi-type scientific and technological information and its application<sup>①</sup>

ZENG Wen(曾文)<sup>②</sup>, LIU Xiaolin, MA Hongyan

(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

## Abstract

The development of network and information technology has brought changes to the production environment of scientific and technological information, leading to the integration of multi-type scientific and technological information, which has become one of the primary research focuses in the current field of scientific and technological information analysis. This article proposes a basic mode to realize the fusion of multi-type scientific and technological information, expounds the corresponding basic construction method, and applies it to the scientific and technological topics identification in the field of artificial intelligence (AI). The research results show that the multi-type scientific and technological information fusion mode proposed in this article has certain feasibility in specific application scenarios, which lays a foundation for the subsequent research work.

**Key words:** information fusion, scientific and technological information, fusion mode, fusion method

## 0 Introduction

The development of network and information technology brings great convenience to the acquisition of scientific and technological information, but the increasingly large scale and different types of scientific and technological data bring challenges to the analysis of scientific and technological information. In recent years, the analysis object of scientific and technological information has changed from a single-type scientific and technological information to a multi-type scientific and technological information. Because multi-type scientific and technological information have their own characteristics in terms of data and content, it is necessary to realize the multi-type scientific and technological information fusion in specific application scenarios. The research and application of information fusion method in the field of sensor and automation are relatively mature, but research and application of scientific and technological information analysis is still in the exploration and research stage. In recent years, some scholars have classified the types of scientific and technological information and put forward the application direction of scientific and technological information fusion<sup>[1-2]</sup>; some scholars try to combine data and geospace for fusion of scientific and technological data<sup>[3]</sup>,

or explore the fusion of multi-source heterogeneous information from the level of data features<sup>[4]</sup>. However, there is no consensus on how to conduct the mode and methods of scientific and technological information fusion. This article analyzes and summarizes the existing problems in the existing scientific and technological information fusion research, combined with the specific application scenarios, puts forward a three-layer basic mode of multi-type scientific and technological information fusion, and carries out its application of relevant methods.

## 1 About the information fusion research

Information fusion is a technology that integrates information from the same types and structure or different types and structure to obtain unified results. It originates from multi-sensor applications in the field of military and automation. Based on multi-sensor data, scholars have developed a series of information fusion methods to achieve tasks such as situation recognition. Subsequently, the concept and method of information fusion have been continuously expanded to remote sensing, transportation and other fields, and the ‘information’ sources used for fusion have also been extended from multi-sensor data to automated systems, data warehouses. Taking the algorithm concept as the

<sup>①</sup> Supported by the National Natural Science Foundation of China (No. 72074201).

<sup>②</sup> To whom correspondence should be addressed. E-mail: zengw@istic.ac.cn.

Received on July 23, 2024

standard, this article divides the information fusion method into the following three types. (1) Method based on the physical model. This method is based on the physical features of the identified object, which is usually of high complexity. It is mostly used for sensing and detection research<sup>[5]</sup>, and it is difficult to be applied to the field of scientific and technological information analysis. (2) Method based on the parameter classification. The method integrates the input information based on statistical theory and information theory. The fusion algorithm based on statistical theory includes classical reasoning method, Bayesian algorithm, Dempster-Shafer (D-S) evidence theory algorithm, etc. Classical reasoning method is only applicable to the judgment between two assumptions. Bayesian algorithm has difficult prior probability definition<sup>[6]</sup>. D-S evidence theory algorithm as a generalized expansion of Bayes theory, considers the overall uncertainty, but there are also ‘exponential explosion’ and ‘evidence conflict’ problems<sup>[7]</sup>. (3) Method based on the cognitive model. The method mainly includes fuzzy set theory, knowledge system. Fuzzy set theory can solve the problems of imprecise, incomplete and unreliable fusion results under the condition of complex environment, noise interference and unstable recognition system. In the process of fusion, the importance of the information source is considered, which is highly practical, and is mostly modeled through ontology in practical

application. In the field of scientific and technological information analysis research, at present multi-types scientific and technological information fusion is still lack of unified mode and construction method, most related researches focus on the level of relationship fusion. Some scholars integrate references, cited relationship of patent and paper data to evaluate emerging technology<sup>[8]</sup>, and in other levels of information fusion research are in the exploration stage, in addition to specific information fusion methods need to be combined with specific application. Therefore, it is very necessary to carry out the fusion mode and application research of multi-type scientific and technological information.

## 2 Basic mode and construction method of scientific and technological information fusion

This article suggests that scientific and technological information fusion can be divided into three essential phase: data level, relationship level and content level according to the different levels of scientific and technological information fusion. Based on these three levels, this article proposes a three-layer basic mode of scientific and technological information fusion, which includes feature integration, relationship fusion and cluster fusion, as shown in Fig. 1.

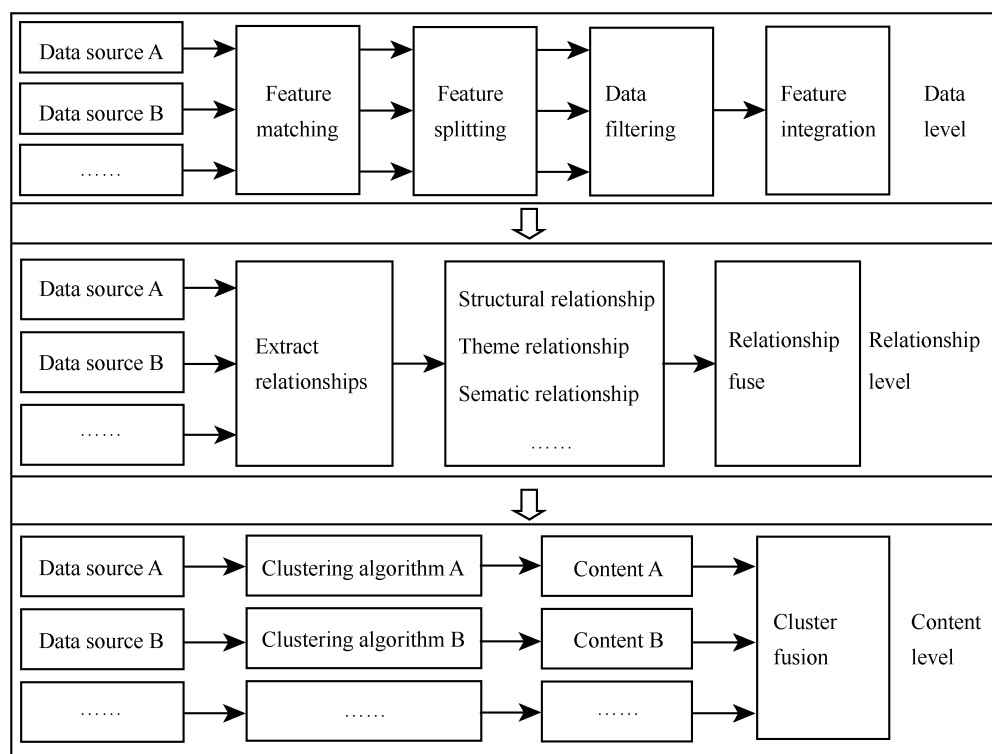


Fig. 1 Three-layer basic mode of multi-type scientific and technological information fusion

## 2.1 Construction of feature integration

Feature integration refers to the integration of the internal features (such as title, abstract, key words, etc.) and the external features (such as author, public time, time cited, etc.) of multi-types scientific and technological information into a unified framework, as shown in Fig. 2. The construction of feature integration includes three steps: feature matching, feature splitting, and data filtering. (1) Feature matching is the analysis and feature mapping of data with the same feature but different identities. For example, the title field of a certain paper is characterized as ‘PIANMING’ in CNKI (China National Knowledge Infrastructure) database and ‘TIMING’ in WANFANG database. For such paper data, it is necessary to unify the title field. It means that the title fields are unified named ‘title’, and then researcher adopts a way of heterogeneous weighted to achieve linear fusion of the papers.

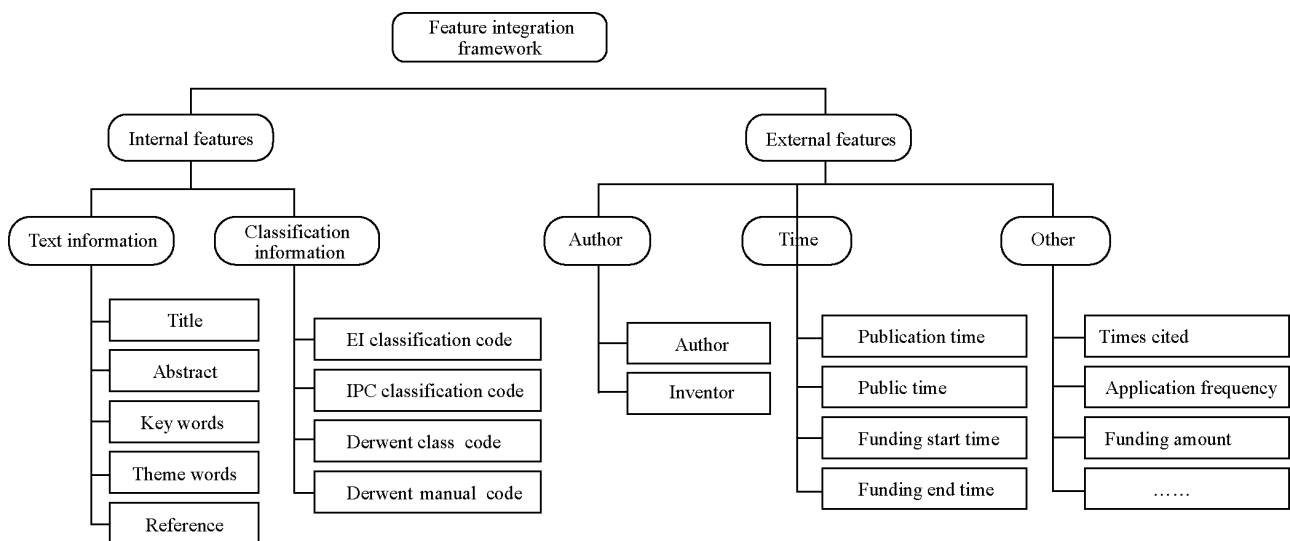


Fig. 2 Feature integration framework

## 2.2 Construction of relationship fusion

Relationship fusion aims to obtain the relationships of different types of data, such as structural relationship (such as citation, classification, co-citation, co-author, etc.), theme relationship (subject word co-occurrence), semantic relationship, etc., and then integrate multiple relationships into a new relationship. There are two main methods of relationship fusion. (1) Calculate the multiple distance matrices to obtain the relationship fusion matrix. Some scholars according to the different research scenarios of the author-literature matrix, literature-journal matrix, word co-occurrence matrix, word-literature matrix, references-literature matrix, author co-occurrence matrix, author-insti-

(2) Feature splitting. Feature split includes two cases such as multi-value same feature and multi-value different feature. Multi-value same feature such as author, organization, keywords and other features needs to be split. Multi-valued different feature such as the cooperative patent classification (CP) field in the Derwent database contains both the patent number and the author of the reference patent. This field contains different data features and needs to be split. (3) Data filtering. Different types of data are consistent in content, so many data are repeated. For example, many of the same journals are included in Science Citation Index (SCI) and Engineering Index (EI) databases. The key of data filtering is to determine the unique identification of data and remove duplicate data. For example, Digital Object Identifier (DOI) can be used as a unique identification for data. The unique identifiers for different types of data are different and need to be integrated according to the specific situation of the data.

tutions co-occurrence matrix and institutions-country matrix, respectively in the field of subject classification, knowledge creation and process analysis among scientific publications, summarizes the field of knowledge exchange mode in application scenarios<sup>[9-10]</sup>. (2) Use the theme model method as Latent Dirichlet Allocation (LDA) for relationship fusion. The method introduces two attributes of author and publication time based on LDA, which is used to explore the implicit themes in the scientific and technological literature and the law of the author’s research interest over time<sup>[11]</sup>. Integrate the author cooperative network and the patent-inventor network into a new heterogeneous network, and use the network to rank the importance of inventors<sup>[12]</sup>. Different types of data have different relation-

ships, so the relationship needs to be constructed according to the specific situation of the data. This article adopts a new method of relationship fusion, as shown in Section 3.

### 2.3 Construction of the cluster fusion

Cluster fusion refers to the fusion of multi-type data content. The clustering algorithm can use the same clustering algorithm or different clustering algorithm, fusion method mainly uses fusion cluster to get the final clustering results, the basic construction process of cluster fusion can be described as follows: considering  $N$  data types of dataset,  $X = \{x_1, x_2, \dots, x_N\}$ ,  $x_i =$

$\{x_{i1}, x_{i2}, \dots, x_{iH}\}$ ,  $i$  represents the  $i$ th data type ( $i = 1, 2, \dots, N$ ),  $H$  is the number of  $H$ th data types. Dataset  $X$  yields a cluster membership set of  $N$  cluster results,  $\beta = \{\beta_1, \beta_2, \dots, \beta_N\}$ , where the cluster result of the  $i$ th data type,  $\beta_i = \{c_{i1}, c_{i2}, \dots, c_{ik}\}$  ( $i = 1, 2, \dots, N$ ),  $k$  is the number of clusters of cluster members. The purpose of cluster fusion is to fuse all the cluster members ( $\beta_1, \beta_2, \dots, \beta_N$ ) into a new cluster result  $\beta^\theta$  through the fusion function  $\theta i$ , as shown in Fig. 3. The content of different dataset is different, so it is necessary to choose the appropriate cluster fusion algorithm according to the specific application situation.

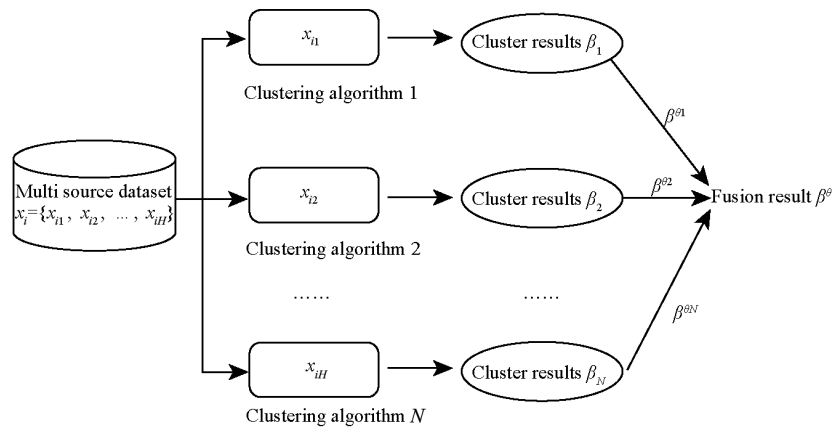


Fig. 3 Cluster fusion process

## 3 Construction example and application of multi-type scientific and technological information fusion mode

The construction of scientific and technological information fusion mode is closely related to the specific application. This article takes the study of topic identification in the scientific and technological field as an example to explain the specific construction process of this mode. Due to the increasingly abundant sources of scientific and technological information, the information available for topic identification is more diverse, so how to correctly identify topic from multi-type scientific and technological information is one of the research hotspot of scientific and technological information analysis, and it is also the basis of scientific and technological prediction. Therefore, this article uses the basic mode and construction method of multi-type scientific and technological information fusion mentioned above for the research of topic identification. The specific implementation process is shown in Fig. 4.

### 3.1 Realization of the construction process of scientific and technological information fusion mode

As shown in Fig. 2, this article selects scientific and technological papers (journal papers and conference papers), patents and science and technology projects as multi-type data, and expounds the specific implementation methods of information fusion from three levels of feature integration, relationship fusion and cluster fusion.

(1) Feature integration. Select a scientific and technological field as the analysis object, and collect the data of conference papers, journal papers, patents and science and technology projects from various sources in the field. After feature matching, feature splitting and data filtering, the title, abstract, key unified words, classification code and other features are integrated into a consistent form.

(2) Relationship fusion. Relationship fusion needs to obtain different relationships according to different application scenarios. It mainly forms a text similarity matrix based on semantic, classification and citation relationship.

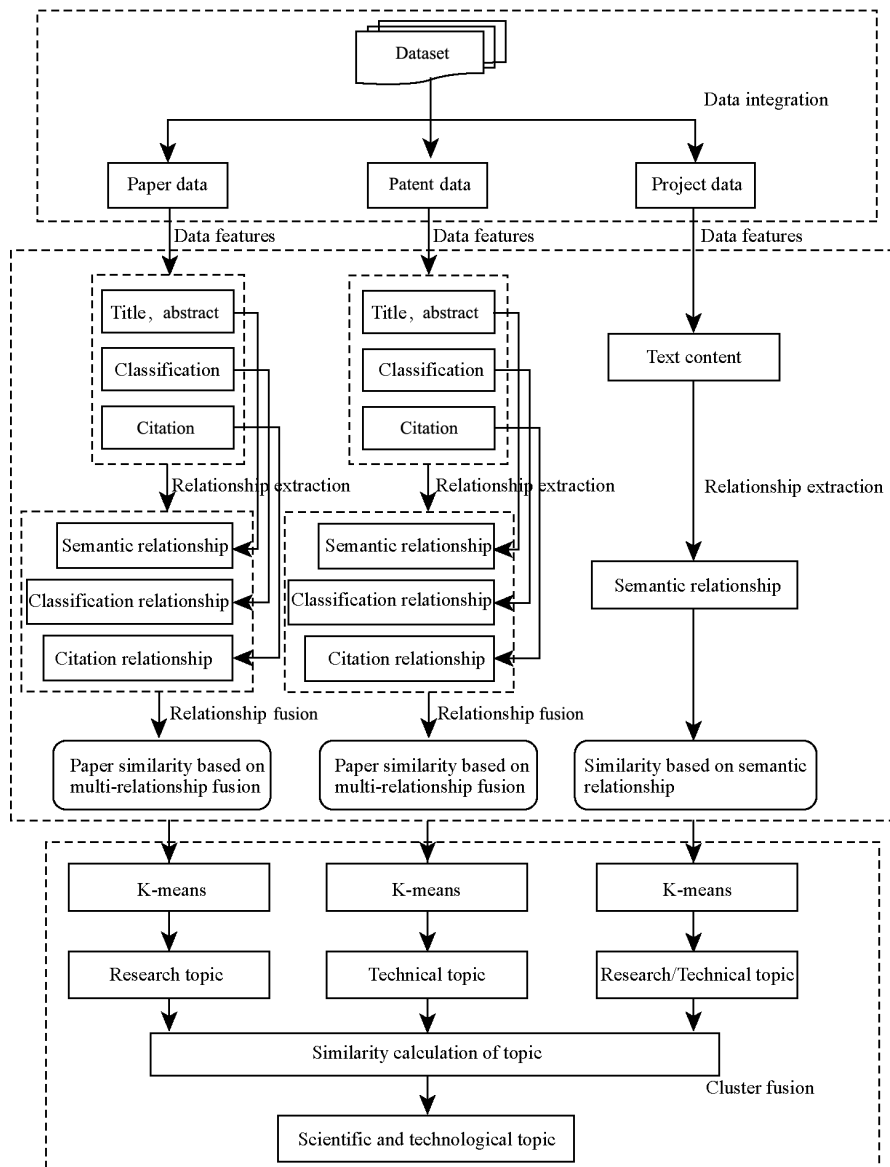


Fig. 4 Implementation process of scientific and technological topic identification based on multi-type information fusion mode

**Step 1** Build the text vector to obtain the semantic relationships. Firstly, extract the title and abstracts, classify, stop words, and select Doc2vec model<sup>[13]</sup> as the text vectorization technique to construct text vector.

**Step 2** Construct the classification vector to obtain the classification relationship. Firstly, the classification information of the literature is obtained. For example, the papers are downloaded from EI database that contains the EI classification number, and the articles are downloaded from Web of Science that contains the research direction, so the classification information of the article can be obtained from the EI classification number and research direction from Web of Science. The classification information of the patent can be obtained from the International Patent Classification (IPC) number of Dewent database. The classification

vector is then constructed by using the acquired classification information. The element values of a vector are the common classification strength between this literature and other literature, that is, the ratio of the number of common classification number between the two literatures to the classification number of the two elements. The higher the number of common classification numbers in the two literatures, the more similar the research topics of the two literatures are. The calculation formula of the classification vector element value is

$$A_{ij} = \frac{B_{ij}}{C_{ij}} \quad (1)$$

where  $i, j$  represent the  $i$ th and  $j$ th literature respectively;  $A_{ij}$  represents the  $j$ th-element value of the  $i$  literature;  $B_{ij}$  represents the number of common classification numbers in the  $i$ th and  $j$ th literature; and  $C_{ij}$  represents the union number of classified numbers in the  $i$ th

and  $j$ th literature.

**Step 3** Build a reference vector to obtain reference relationships. The reference vector is constructed according to the reference of the literature, and the element value of the vector is the common citation intensity value of the literature and other literature, that is, the ratio of the number of common references to the union number of both references. The formula for calculating the vector element values can refer to Eq. (1).

**Step 4** Relationship fusion is the vector fusion. The constructed text vector, classification vector and reference vector are integrated to obtain the multi-relationship vector of the literature. If the text vector of a literature is  $[a_1, a_2, a_3, \dots, a_n]$ , the classification vector is  $[b_1, b_2, b_3, \dots, b_n]$ , and the reference vector is  $[c_1, c_2, c_3, \dots, c_n]$ , the multi-relationship fusion vector is  $[a_1, a_2, a_3, \dots, a_n, b_1, b_2, b_3, \dots, b_n, c_1, c_2, c_3, \dots, c_n]$

(3) Cluster fusion. Calculate the distance and similarity of the relationship fusion vector in space to form the text similarity matrix, and then select the clustering algorithm K-means<sup>[14]</sup> to cluster the text content of multi-type literature. After the clustering of text content, the cosine similarity calculation method is used to realize the fusion of multi-type scientific and technological topics.

### 3.2 Experiments and analysis

Due to the lack of comparable models and methods for the multi-type information fusion, this article validates the proposed mode and methods by analyzing historical data and comparing them with the conclusions of classic scientific and technological reports. Based on the considerations of data availability and completeness, this article selects scientific and technological papers, patent data, and project data in the field of artificial intelligence (AI) as examples to explore the feasibility of the methods described above. The data for the paper is sourced from international academic journals recommended by China Computer Federation (CCF) in the field of AI, which is included in the EI database. The conference paper data is obtained from the Conference Proceedings Citation Indices (CPCI) database and EI database, which is the core datasets of Web of Science. The total number of paper is 12 230. The patent data is taken from the Derwent Patent database, with a total of 11 680 patents. The project data is obtained from the National Science Foundation (NSF) database in the United States, with a total of 1 126 projects. The retrieval time of three types data is from January 2017 to January 2021.

The experimental results obtained by the method described above are shown in Table 1. Table 1 shows the scientific and technological topics identified based on scientific and technological papers, patents, and project data respectively, as well as the scientific and technological topics identified by integrating three types data. It can be observed that the results of identifying scientific and technological topics using a single dimension of scientific and technological information are limited in both quantity and content, while the scientific and technological topics obtained through the use of scientific and technological information fusion methods have advantages in both quantity and content. In terms of the number of topics, 11 topics can be identified based on papers, 13 topics can be identified based on patents, 3 topics can be identified based on projects, and 15 topics can be identified using the scientific and technological information fusion method proposed in this article. From the perspective of topic content, compared with the papers-based method, the results of using the information fusion method proposed in this article for scientific and technological topic identification have added 6 topics including intrusion and anomaly detection technology, interpretability of deep learning, long short-term memory neural networks, deep learning clustering models, deep belief networks and human action recognition technology. Compared with the patents-based method, 7 topics have been added including target detection and recognition technology, intrusion and anomaly detection technology, interpretability of deep learning, deep neural networks, feature extraction techniques, deep transfer learning, and deep reinforcement learning. Compared with the projects-based method, 12 topics have been added including image segmentation technology, image enhancement technology, generative adversarial networks, image reconstruction technology, deep transfer learning, deep reinforcement learning, deep neural network, feature extraction technology, long short-term memory neural networks, deep learning clustering model, deep belief networks and human action recognition technology and so on.

It is found that the scientific and technological topics obtained by the use of information fusion method proposed in the article, such as reinforcement learning, interpretability of deep learning, feature extraction, image classification, object detection, generative adversarial networks, semantic segmentation, etc., all appear in the AI development reports issued by important research institutions. For example, Stanford University's 'Artificial Intelligence and Life 2030' points out that reinforcement learning will shift its focus to decision-making, but it has not achieved substantial

success in practice. The report points out that the rapid development of deep learning has provided a strong boost to reinforcement learning. The emergence of deep learning has promoted the combination of reinforcement learning and deep learning, giving rise to a new learning method, i. e. , deep reinforcement learning. The report points out that deep reinforcement learning is one of the research trends in the field of AI. In January 2021, the AI Research Institute of Tsinghua University and the Tsinghua Chinese Academy of Engineering Knowledge Intelligence Joint Research Center jointly released the ‘Artificial Intelligence Development Report 2020’, which mentioned that reinforcement learning and interpretability of AI were one of the key scientific and technological directions for future develop-

ment. The report points out that deep learning networks are in a period of expected inflation. After 2 – 5 a, deep learning networks would enter a mature stage. The report also pointed out that enhanced intelligence, including image enhancement, is a new direction of AI, which is currently in the research stage and is one of the research directions that urgently needed breakthroughs. The 6 hot topics identified in this article, such as deep neural networks, feature extraction, image classification, object detection, generative adversarial networks, semantic segmentation, etc. , all appear in the top 10 hot research topics of AI in this report. This indicates that the identifying topics method proposed in this study has certain feasibility and reliability.

Table 1 Comparison example of identifying topics before and after applying information fusion mode and method

| Topics obtained based on papers            | Topics obtained based on patents      | Topics obtained based on projects          | Topics obtained after applying information fusion mode and method |
|--|---------------------------------------|--|---|
| Deep neural network                        | Predictive algorithms and models      | Interpretability of deep learning          | Target detection and recognition technology                       |
| Classification algorithms and models       | Image classification technology       | Target recognition and detection algorithm | Intrusion and anomaly detection technology                        |
| Target recognition and detection algorithm | Object detection technology           | Intrusion detection model                  | Image segmentation technology                                     |
| Image segmentation technology              | Long short term memory neural network |  | Image enhancement technology                                      |
| Image enhancement technology               | Generative adversarial networks       |  | Generative adversarial networks                                   |
| Feature extraction technology              | Image segmentation technology         |  | Image reconstruction technology                                   |
| Generative adversarial networks            | Image reconstruction technology       |  | Interpretability of deep learning                                 |
| Image reconstruction technology            | Abnormal detection technology         |  | Deep neural network   |
| Abnormal detection technology              | Image enhancement technology          |  | Feature extraction technology                                     |
| Deep transfer learning                     | Deep learning clustering model        |  | Deep transfer learning  |
| Deep reinforcement learning                | Deep belief network                   |  | Deep reinforcement learning                                       |
|  | Human motion recognition technology   |  | Long short-term memory neural network                             |
|  | Text classification technology        |  | Deep learning clustering model                                    |
|  |                                       |  | Deep belief network   |
|  |                                       |  | Human motion recognition technology                               |

## 4 Conclusion

At present, the relevant research of information fusion mode and method in the field of scientific and technological information analysis is not mature, and there are problems such as insufficient depth of information fusion and lack of unified fusion mode. Therefore, building upon the review of related research on information fusion, this article proposes a basic mode for the fusion of multi-type scientific and technological information from three levels: feature integration, relationship fusion, and cluster fusion. Empirical research on the problem of topic identification is conducted. The results show that the multi-type scientific and technological information fusion mode and method proposed in this article are more accurate than the topic identification methods based on a single-type of science and scientific information, which is conducive to the better scientific and technological prediction research in the future.

There are still some shortcomings in the research in this article as follows. (1) The data used in this article is relatively limited, so the relationship fusion of the data in the article is not sufficient. If future research can supplement the reference information of papers and patents, the identification results will be more accurate. (2) The calculation workload of data in this article is large, and the designed scientific and technological information fusion method needs to be further optimized to shorten the training time.

### Reference

- [ 1 ] HUA B L. Research on multi-source information fusion methods[J]. *Information Study:Theory and Application*, 2013, 36 (11): 16-19.
- [ 2 ] LI G J, YANG L. Intelligence analysis and intelligence technology in view of big data [J]. *Library and Information*, 2012, 33 (6): 1-8
- [ 3 ] BAI R J, JU Z H, ZHANG Y J, et al. Research on multi-source and multi-mode data fusion for intelligence perception [J]. *Journal of Intelligence*, 2023, 42 (10): 124-131.
- [ 4 ] LIU Y, HU Z P, HAN X Y, et al. Deep learning based fusion and recommendation of heterogeneous multi-source intelligence for emergency events [J]. *Information Science*, 2024, 42(4): 136-144.
- [ 5 ] MA H Y. Research on information fusion method orienting science and technology frontier exploration[D]. Beijing: Institute of Scientific and Technical Information of China, 2022: 35-37. (In Chinese)
- [ 6 ] YANG J B, LIU J, WANG J, et al. Belief rule-base inference methodology using the evidential reasoning approach——RIMER[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2006, 36 (2): 266-285.
- [ 7 ] DEMPSTER A P. Maximum likelihood from incomplete data via the EM algorithm[J]. *Journal of the Royal Statistical Society*, 1977, 39(1): 1-22.
- [ 8 ] TAN L, LI H. Research front recognition based on multi-source data knowledge fusion method [J]. *Journal of Modern Information*, 2019, 39(8): 29-36.
- [ 9 ] AMJAD T, DING Y, DAUD A, et al. Topic-based heterogeneous rank[J]. *Scientometrics*, 2015, 104(1): 313-334.
- [ 10 ] CALERO C M, NOYONS E C M. Combining mapping and citation network analysis for a better understanding of the scientific development: the case of the absorptive capacity field[J]. *Journal of Informetrics*, 2008, 2(4): 272-279.
- [ 11 ] XU S, SHI Q, QIAO X, et al. Author-topic over time (AToT): a dynamic users' interest model [J]. *Lecture Notes in Electrical Engineering*, 2014, 274: 1-7.
- [ 12 ] DU Y P, YAO C Q, LI N. Using heterogeneous patent network features to rank and discover influential inventors [J]. *Frontiers of Information Technology and Electronic Engineering*, 2015, 16(7): 568-578.
- [ 13 ] MIKOLOV T, CHENK K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. *Computer Science*, 2013, 43(7): 1-12.
- [ 14 ] WEN Z H. Research on weighted slope one algorithm based on project classification and K-means clustering [D]. Qinhuangdao: Yanshan University, 2017: 43-56. (In Chinese)

**ZENG Wen**, born in 1973. She received her Ph. D degree in Shenyang Institute of Automation Chinese Academy of Sciences in 2009. She also received M. S. degree from the College of Information Science and Engineering at Northeastern University in 2003. Her current research interests include scientific and technological information theory and method, scientific and technological frontier.