Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

INTELLIGENT ANALYSIS OF SCIENTIFIC AND TECHNOLOGICAL LITERATURE: A NEW PARADIGM FOR RESEARCH EFFICIENCY AND INSIGHT DISCOVERY

ZAINAB KAREEM ABDULLAH[*1]

[1]*Ministry of Education, Iraq, almwswymhmd125@gmail.com*

ABSTRACT

This paper proposes a new scientific research paradigm, intelligent analysis of scientific and technological literature. By comparing traditional literature analysis methods, it emphasizes the significant advantages of intelligent analysis of scientific and technological literature in improving research efficiency and depth. The article elaborates on the concept of intelligent analysis of scientific and technological literature and its great role in scientific research, and looks forward to the theoretical basis of natural language processing, machine learning and other technologies in realizing intelligent analysis of scientific and technological literature. A proof-of-concept system is designed, and some core functions are tested and analyzed using some random paper data. Intelligent analysis of scientific and technological literature is expected to become an indispensable research tool for researchers and promote scientific research to a new level.

Keywords: Intelligent Literature Analysis, Natural Language Processing (NLP), Machine Learning in Research, Scientific Knowledge Discovery, Bibliometric Tools

1. INTRODUCTION

As an important means to extract valuable information from massive literature, understand and gain insights into research trends, and assist in scientific research decision-making, the development of scientific and technological literature analysis is closely related to the development of information technology. In the traditional period, it mainly relied on manual reading, manual sorting, and simple statistical analysis. It was inefficient and easily affected by subjective factors. Entering the computer-assisted stage, with the popularization of computers, various literature management tools such as EndNote and Zotero [1] emerged to assist researchers in collecting, sorting, and annotating literature. Then entered the data mining era. The introduction of data mining technology made it possible to conduct more in-depth mining of literature. Researchers began to use clustering, classification, association rules and other technologies to develop tools such as CiteSpace, VOSviewer, Bibliometrix and SciMAT [2] to extract implicit knowledge from literature. With the rise of artificial intelligence, we have gradually entered the era of artificial intelligence. Especially with the progress in the fields of big data and natural language processing, scientific and technological literature analysis has entered a new stage. Machine learning, deep learning and other technologies are widely used in tasks such as literature summarization, sentiment analysis, and topic modeling, which will greatly improve the efficiency and accuracy of analysis. Based on this, this paper proposes a new intelligent analysis of scientific and technological literature (Intelligent Insights) for scientific and technological literature analysis, and designs a corresponding proof of concept (Proof of Concept) system.

2. LITERATURE REVIEW

2.1 Functions and features of scientific literature quantitative analysis software

Existing bibliometric analysis tools [2] CiteSpace, VOSviewer, Bibliometrix and SciMAT are designed to help researchers deeply explore the huge scientific literature database and reveal the complex relationships therein. Their core function is to visualize the co-citation network. Through this intuitive way, researchers can clearly observe the relationship between different research fields, the evolution path of knowledge, and the distribution of key authors, papers and topics. The generated co-citation

network not only presents a static knowledge graph, but also provides some interactive functions. Researchers can delve into co-citation papers of interest and understand the development status of research frontiers in different fields. This function is valuable for conducting a comprehensive literature review and identifying key topics and influential works in a field. By analyzing the citation activity of a specific research topic, emerging research trends can be identified, helping researchers to gain insight into future research directions in advance. These quantitative software can also analyze the time development of citations and track the changes and development trends of literature, so as to have a deeper understanding of the knowledge evolution process in a certain research field. In addition, the software can also detect and visualize the cooperation relationship between authors, so as to understand the collaboration mode between different research teams and the impact of cooperation on research output and innovation. In addition, the software can track the emergence of keywords, conduct in-depth analysis of topics in the literature, and identify emerging research areas by tracking the frequency and changes of keywords.

These analytical software can help researchers keep up with the latest research trends and provide researchers with a diverse toolbox to help them better understand the generation, dissemination and evolution of scientific knowledge. Through visualization, interactive and predictive analysis functions, researchers can position their research in the current academic environment.

2.2. Deficiencies of existing scientific literature quantitative analysis software

In today's era of rapid development and widespread application of artificial intelligence, existing scientific literature analysis software has highlighted some functional deficiencies in the following aspects. The specific manifestations are as follows, such as the lack of real-time data integration. Existing software usually requires users to manually import paper data in a specified format, cannot import full-text files, and cannot achieve real-time connection with the literature database, which greatly reduces efficiency; without machine learning capabilities, the analysis results of existing software are often static and cannot be automatically updated as new literature emerges; lack of context analysis, the analysis of existing software on literature mostly stays at the level of keywords and abstracts, and cannot deeply understand the relationship between literature and citation context; weak natural language processing function, analysis software overly relies on article abstracts and cited literature, and cannot extract more valuable information from massive text. In addition, due to the lack of natural language processing function, analysis software cannot analyze subjective emotions in literature and cannot understand the author's attitude and views on research results. Faced with today's information explosion, scientific and technological literature has shown an exponential growth. The massive data generated by scientific research results worldwide requires us to no longer stay on the visualization function of existing software, but to deeply understand the data and content of literature. Obviously, traditional retrieval, export methods, and quantitative analysis software are no longer sufficient to meet the needs of researchers.

2.3. What is "intelligent analysis of scientific literature"?

Traditionally, researchers often complete such work by combing literature, including manual collection and combing, or using bibliometric analysis tools to process the collected data in a certain format. The intelligent analysis of scientific and technological literature proposed by us refers to the use of artificial intelligence technology to deeply mine, analyze and understand massive amounts of unstructured scientific and technological literature data, so as to quickly and accurately obtain key information and discover potential knowledge associations, thereby helping researchers, engineers, etc. to quickly obtain the required information and promote scientific research innovation. The text of scientific and technological literature contains rich implicit knowledge and potential value. The intelligent analysis of scientific and technological literature also represents a paradigm shift in research. From manual processing, importing data into quantitative analysis software to generate visualization, to extracting deeper and more usable knowledge information from complex data sets (unstructured data). Such a system integrates artificial intelligence, natural language processing, machine learning, data analysis, visualization and cognitive reasoning technology. It can unlock deeper information in the literature, predict future scientific research development trends, and thus promote the rapid development of scientific research. The intelligent analysis system of scientific and technological literature also uses machine learning functions to train local data models and automatically update and iterate, thereby providing forward-looking guidance for researchers' decision-making. The system is also expected to better achieve cross-disciplinary and cross-professional collaboration and complementarity, and improve scientific research prediction capabilities.

3. INTELLIGENT ANALYSIS SYSTEM OF SCIENTIFIC AND TECHNOLOGICAL LITERATURE AND ITS IMPLEMENTATION

3.1. Overview of the Intelligent Analysis System for Scientific and Technological Literature

Based on the above ideas, we designed a proof of concept system for intelligent analysis of scientific literature.

First, the support for real-time collection and input of unstructured data is a significant difference between the scientific and technological literature intelligent analysis system and traditional literature quantitative analysis tools. Traditional software often relies on pre-defined structured

data and requires input data to have clear formats and labels. However, there is a large amount of unstructured text data in the real world. For example, academic papers are often file data in various formats, such as Word, PDF, or HTML documents. The scientific and technological literature intelligent analysis system can support these documents without the need for user conversion or export.

As mentioned above, traditional quantitative analysis software can usually only provide objective information such as the frequency of citations and publication year of the literature, and cannot conduct in-depth semantic analysis of the text content. Therefore, it cannot accurately judge the author's attitude towards the research results. Traditional literature analysis software lacks semantic understanding and has limited ability to understand text. It cannot accurately identify key information in the text, nor can it provide users with an overview of key articles. At the same time, it cannot provide fully flexible customized visualization.

The intelligent analysis system of scientific and technological literature automatically discovers implicit topics in the literature and classifies the literature through concept extraction and nomenclature recognition. By analyzing the distribution of topics in different periods, we can understand the changes and development trends of hot spots in the research field. The fully customized visualization function can flexibly use the latest Python call function library to generate visualizations such as word clouds and theme maps to intuitively display the distribution and evolution of topics. Based on the function of opinion mining (also known as sentiment analysis) of the system, the system can analyze the opinions in the literature and judge whether the author's attitude towards the research results is supportive, opposed or neutral. The sentiment polarity analysis of the opinions can be performed to determine whether the opinions are positive, negative or neutral. Therefore, relevant high-quality literature and the latest research results can be recommended to users. Literature similar to the target literature can also be recommended based on the content similarity of the literature. With the help of the overview extraction function, the system can extract relevant wonderful paragraphs in the original text according to user needs, saving users time reading the original text, and also obtaining valuable paragraphs far higher in quantity and quality than the abstract. Based on intelligent question and answer of generative AI, users can obtain knowledge in specific fields from trained big data models by asking questions, such as "Who is the authoritative scholar in this field?", "Which institutions are most active in research in this field?", "What is the main contribution of this article?", etc.

Compared with the function set of the above-mentioned intelligent analysis of scientific and technological literature, the functional deficiencies of the existing software are obvious. Table 1 compares the function sets of the two.

3.2. Overview of the implementation methods of intelligent analysis of scientific literature

The user input, data collection and preprocessing module is the foundation of the system and also the input module. This module is responsible for importing unstructured text data such as PDF, Word, TXT, HTML and other files from various channels (such as user-uploaded documents, network, document library real-time download, etc.), and performing data cleaning, denoising, word segmentation and other preprocessing to provide high-quality data for subsequent analysis.

Table 1. Comparison of features between traditional tools and intelligent Insights system

Functional Classification	**Traditional Bibliometric Analysis Tools**	**Intelligent Analysis System for Scientific and Technological Literature**
Topic network generation	Support	Support
Author network generation	Support	Support
Co-citation network generation	Support	Support
Evolution timeline generation	Support	Support
Emergence Detection	Support	Support
Unstructured Data	Not supported	Support
Full-text-based nomenclature recognition	Not supported	Support
Concept extraction based on full text	Not supported	Support
Full-text based text classification	Not supported	Support
Opinion mining based on full text	Not supported	Support
Text profile based on the full text	Not supported	Support
Smart Question and Answer	Not supported	Support
Visual customization	Partial support	Support

The system’s large language model interface, LLM API (Large Language Model API), provides users with a means to quickly access the most advanced large language models [3] (such as GPT-4, Google Bard, Meta LLaMA, etc.) to achieve human-computer dialogue. The LLM API is highly flexible and can customize the model’s output style, content scope, and interaction method according to different application scenarios, and develop customized intelligent question and answer systems.

The NLP module is at the center of the scientific and technological literature intelligent analysis system, including the data mining engine (Text Mining Engine), machine learning and model training, and internal model submodules. Based on preprocessing, the NLP module extracts valuable features from the text, such as keywords,

word frequency, sentiment tendency, etc. These features will be used as input for machine learning and model training to build and train the internal model. This internal model is different from the LLM from the outside. It is the basis of the entire scientific and technological literature intelligent analysis system. The extracted features are used to build and train internal models based on deep learning, such as classification models, clustering models, and generation models. Based on the internal model that is continuously iteratively trained, users can extract named entities and concepts, classify them, and compress long texts to generate a refined overview of the core information. We will discuss these core functions of the scientific and technological literature intelligent analysis system in detail in the next chapter.

The report generation and visualization module of the system presents the analysis results to the user in a concise and visual manner, such as in the form of graphs, tables, text overviews, etc., to generate reports with clear structure and rich content.

In general, the scientific literature intelligent analysis system is a highly integrated system, with each module interdependent and mutually reinforcing. From data collection to report generation, the entire process is a continuous process of intelligent knowledge processing. Through such a system, we can more deeply explore the value contained in the literature and provide support for scientific research decision-making.

4. KEY FUNCTIONS AND IMPLEMENTATION OF INTELLIGENT ANALYSIS OF SCIENTIFIC LITERATURE

To realize the intelligent analysis system of scientific and technological literature, it is necessary to integrate the research results of multiple interdisciplinary fields, including computer science, artificial intelligence, natural language processing, data mining, etc. Its theoretical basis mainly comes from the following disciplines: knowledge graph, data mining, natural language processing, and machine learning. Among them, knowledge graph and data mining based on statistical results have been widely used in existing methods and will not be repeated here. How to solve the functional deficiencies of existing software tools in Table 1 technically? In theory, these problems can be answered by natural language processing (NLP) technology [4]. We discuss them in the following sections and use our proof-of-concept system to give practical applications and demonstration results.

4.1. Named Entity Recognition for Intelligent Analysis of Scientific Literature

4.1.1. Theoretical basis of named entity recognition

Named Entity Recognition (NER) in intelligent analysis of scientific and technological literature [5] is a subtask in NLP. Its goal is to identify entities with specific meanings in text and classify them into predefined categories, such as names of people, places, and names of organizations. In the application of intelligent analysis of scientific and technological literature, NER plays a vital role. It provides a basis for subsequent tasks such as text analysis, information extraction, and knowledge graph construction. The theoretical basis of NER mainly comes from the following aspects: statistics and probability models. NER is essentially a classification by calculating the probability that a word belongs to a certain named entity. Commonly used probability models include hidden Markov model (HMM) and conditional random field (CRF). The text features are converted into numerical features that can be processed by the model, such as part of speech, word frequency, context, etc. NER usually adopts supervised learning, that is, the model is trained through labeled training data. Some classification algorithms, such as support vector machine (SVM), decision tree, random forest and other traditional machine learning algorithms have been widely used in NER tasks.

4.1.2. Experiments and results on the named entity recognition function

In the proof-of-concept system, we input a photo of a recently published article in the Journal of Intelligent Learning Systems and Applications [6] for named entity recognition and uploaded the article's PDF file "jilsa2024164_59601667.pdf". The system extracted the named entities shown in Table 1. This information is not obtained from a fixed-format file imported by the user like the trauma software, but from unstructured text such as PDF files. Table 2 includes geographic locations, names, and names of organizations. In total, 13 organizational names, 28 names, and 5 geographic locations were detected in the paper. Due to space constraints, not all information can be listed. Table 3 gives the geographic locations, and Figure 3 gives a bar chart of the distribution of detected organizations.

Table2. Named entity recognition result

Named Body	**quantity**
Organization Name	13
Name	28
Location	5

Table 3. Detection results of named geographic location

Place Names	**Confidence**
Madina	56.87%
Saudi Arabia	92.56%
Los Angeles, California	51.03%
New Orleans, Louisiana	88.38%

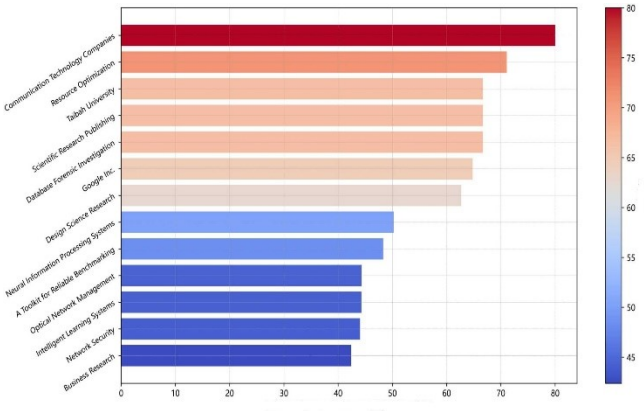

Figure 1. Named entity recognition demonstration and experimental results

As can be seen above, this proof-of-concept system can extract relevant real-name entities from randomly downloaded texts. Traditional software must import relevant data (such as abstracts and citations) in a certain format to extract relevant information such as research topics.

4.2. Concept extraction in intelligent analysis of scientific literature

4.2.1. Theoretical basis of concept extraction

Concept extraction [7] is an important branch of NLP. Different from named entity recognition, it aims to automatically identify and extract concepts with clear meanings from text. These concepts can be people, objects, organizations, events, and attributes, or more abstract concepts. Concept extraction provides a basis for many NLP tasks such as information extraction, knowledge graph construction, and text classification. The theoretical basis of its application is linguistic theory, statistics, and machine learning. Concept extraction is rooted in linguistic theory, especially semantics. Concepts in text can be identified by analyzing the semantic roles, semantic relationships, and contextual information of words. Statistical methods also provide powerful tools for concept extraction. For example, potential concepts can be discovered through word frequency statistics, co-occurrence analysis, and other methods. Machine learning algorithms, especially deep learning models, play an increasingly important role in concept extraction. By training a large amount of labeled data, the model can automatically learn complex feature representations and accurately identify concepts.

4.2.2. Functional Experiments and Results of Concept Extraction

We randomly selected an article from the open access journal “Open Journal of Social Sciences”, “Investigating the Impact of Conflict Management Approaches on Organizational Productivity in Healthcare Settings: A Qualitative Exploration” [8] as an example, and used our scientific literature intelligent analysis system to extract concepts, inputting the downloaded file jss20241211_231769282.pdf. The test results are shown in Figure 4. It is worth noting that the concepts extracted from the article are not limited to the information in the abstract, title, or keywords of the article, but the content of the full text of the paper is analyzed and processed. The system test results give the top ten complex concepts and simple concepts, and output the word cloud diagram shown in Figure 4. The system can generate different styles of word cloud diagrams according to different languages and fonts, and different mask images. This diagram uses a spherical mask as the mask image and the ERNHC.TTF font to generate a better visual effect. The system can also output detailed concept extraction results in the form of a bar heat map, as shown in Figure 5 .

Figure 2. Concept extraction word cloud graph

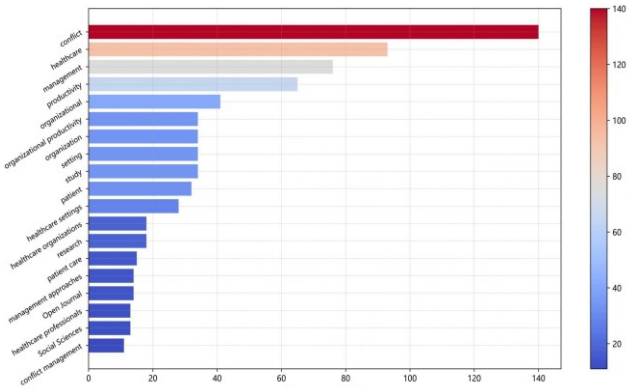

Figure 3. Concept extraction word frequency graph

4.3. Text Classification in Intelligent Analysis of Scientific and Technological Literature

4.3.1. Theoretical basis of text classification

Text classification in intelligent analysis of scientific and technological literature can be divided into different categories according to the content of the literature, such as research direction, theme, etc. As mentioned above, through NLP technology, we can analyze and classify any form of text, and are no longer limited to structured data. After text preprocessing and feature extraction, that is, through the bag-of-words model, TF-IDF, Word Embedding [9] and other technologies, the text is converted into a numerical feature vector for processing by the machine learning model. Then, the classification model uses machine learning

algorithms such as naive Bayes, support vector machine, deep learning model (such as RNN, CNN), etc. to classify the text. The advantage of NLP text classification is that it greatly improves the classification efficiency and reduces manual intervention. By training a large amount of labeled data or existing data models, a high classification accuracy can be achieved. It can classify various types of text and has strong adaptability. Moreover, with the continuous training and iteration of the model, the classification effect can be continuously improved.

4.3.2. Experiments and Results of Text Classification

Generally speaking, the dedicated local model in the intelligent analysis system of scientific and technological literature can accurately classify unstructured data. However, such a dedicated model requires a lot of machine training to improve the accuracy of classification. Due to time and space reasons, we adopted a shortcut mode to simplify the machine training process. In the test, we selected four journals from Hans Publishing House: "Material Science", "Frontiers in Sociology", "Statistics and Applications", and "Computer Science and Applications" for random downloads. The detailed file distribution in the test corpus is shown in Table 4 .

Table 3. Training corpus documents distribution

Journal Name	Number of papers	Text size
Materials Science	17	38,119,464
Frontiers of Social Sciences	19	11,375,037
Statistics and Applications	29	74,913,303
Computer Science and Applications	29	88,378,902

After learning the system machine learning, we selected another set of files for benchmarking. The file list is shown in Table 4 .

The results of the local model's benchmark proofreading are measured by the following indicators: Precision measures the accuracy of the model's prediction of the positive class, Recall measures the model's ability to identify all relevant instances, and F Score (F1 score) is the harmonic mean of Precision and Recall. Silence refers to the inability of the model to classify a text when doing benchmark verification. Silence = 1 − Precision. Noise refers to the interference information encountered by the model when processing the text, which causes the text to be incorrectly classified. Noise = 1 − Recall. It can be seen that the performance of the scientific and technological literature intelligent analysis proof of concept system can achieve a certain accuracy (~90%) even based on a small corpus. In actual operation, if you want to achieve higher precision and recall, you must expand the corpus and iterate machine learning.

Table 4. Journal paper file list for benchmarking

file name	Magazine	File size
ms20241000000_85110778.pdf	Materials Science	745,644
ms20241410_101281771.pdf	Materials Science	3,801,839
ms20241410_111281763.pdf	Materials Science	2,963,651
sa2024135_192581414.pdf	Statistics and Applications	2,365,446
sa2024135_202581421.pdf	Statistics and Applications	3,978,672
sa2024135_212581436.pdf	Statistics and Applications	438,700
ass20241311_392397511.pdf	Frontiers of Social Sciences	448,676
ass20241311_402397901.pdf	Frontiers of Social Sciences	533,220
ass20241311_412397940.pdf	Frontiers of Social Sciences	461,201
ms20240200000_45453600.pdf	Computer Science and Applications	5,023,856
ms20240200000_49175675.pdf	Computer Science and Applications	998,072
ms20240200000_71545278.pdf	Computer Science and Applications	7,702,578

4.4. Opinion Mining in Intelligent Analysis of Scientific Literature

4.4.1. Theoretical basis of opinion mining

Opinion mining, also known as sentiment analysis [10], aims to analyze subjective information such as emotions, opinions, and attitudes expressed in texts. For academic literature, sentiment analysis can help us understand the author's evaluation of the research results, whether it is positive affirmation or negative criticism, so as to have a deeper understanding of his academic views. Sentiment analysis technology can make up for the shortcomings of traditional methods by deeply understanding the semantics of the text. It can: For example, it can identify sentiment polarity: determine whether the text expresses positive, negative or neutral emotions, locate sentiment words to find keywords that express emotions, such as "good", "bad", "excellent", "bad", etc. Analyzing sentiment intensity can evaluate the intensity of emotions, such as "very satisfied", "average", "very dissatisfied". Understanding the reasons for emotions: analyzing the reasons for expressing emotions in the text, so as to have a deeper understanding of the author's views. The theoretical basis of sentiment analysis comes from statistics, and classification is performed by calculating the probability that the text belongs to different sentiment categories. Commonly used statistical models include naive Bayes and support vector machine discriminant models. Machine learning provides a large number of algorithms and models for learning rules from data and applying them to new data. For example, a classifier can be trained to map text features to sentiment

labels. Deep learning, especially neural network-based models, has achieved remarkable results in sentiment analysis tasks. Models such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and Transformers can automatically extract high-level features from text, thereby improving the accuracy of sentiment classification.

4.4.2. Experiments and results of opinion mining

Regarding the opinion mining test, we can enter two paragraphs of text in the scientific literature intelligent analysis system:

1. [John Mack] provides a thorough and insightful analysis of [1991], satisfactorily summarizing the existing problems and offering reasonable suggestions…
2. [Victor King]'s evaluation of [2] is too general and lacks in-depth analysis of the research details…

Table 6. Sentimental analysis test demonstration

Sentence	Positive Score	Negative Score	Overall Tone
John Mike conducted an in-depth and detailed analysis of the research in [1], and clearly sorted out the contributions and shortcomings of the research in this field. The author particularly emphasized the innovation of [1] and compared it with [related research] to highlight its uniqueness. In addition, the author also put forward constructive suggestions on the possible future development direction of [1], providing useful inspiration for subsequent research.	58.34	7.34%	Positive
Victor's evaluation of [2] is too general and lacks in-depth analysis of the research details. Although the author mentioned some of the advantages of [2], he did not fully discuss its limitations. In addition, when comparing [2] with other related studies, the author did not select clear comparison points, which affected the objectivity of the evaluation.	16.34%	53.8%	Negative

In traditional bibliometric software, as long as two documents are in the references, they will be included in the calculation of co-citation documents and have the same weight. But the actual situation is not so. Not all cited articles play a positive role in the author's research. Sometimes, some authors try to show the uniqueness and innovation of their research by citing some non-frontier documents. Through the intelligent analysis of scientific and technological literature, we can identify them. The second case in this example is actually a negative evaluation, which is of general significance to both the author and the paper reader. Through the intelligent analysis system of scientific and technological literature, we can analyze the tone between the lines of the dialogue and mark the negative comments. The test results are shown in Table 6. In this example, the first document is judged as a positive evaluation, while the second document is listed as a negative evaluation. With such automatic marking, scientific researchers can choose documents for further reading.

4.5. Text overview in intelligent analysis of scientific literature

4.5.1. Theoretical basis of text profiles

Text summarization in intelligent analysis of scientific and technological literature is different from the abstract of scientific and technological literature. The abstract is a short and coherent text generated by the author from an entire article or document, extracting its core ideas, main arguments and conclusions. NLP-based text summarization [11] focuses more on extracting key information from the text and generating a shorter version than the original text while retaining the core meaning of the original article. It can be an abbreviated version of the entire article or a summary of a specific paragraph. Compared with the abstract, the text summary pays more attention to the compression and retention of information to ensure that the generated text can accurately reflect the main ideas and meaning of the original text; it can generate summary content of different lengths and styles according to different needs. Text summarization can be achieved through the following methods: converting the text into a representation that can be processed by a computer, such as word vectors, sentence vectors, etc.; sorting the sentences or words in the text by importance in order to extract key information; using various compression algorithms to compress the text into a shorter version; and converting the compressed information into natural language text. Common methods for text overview include statistical methods, such as algorithms centered on high-frequency words; TF-IDF methods to measure the importance of words in documents and consider the prevalence of words; and graph-based methods, which represent text as graphs, with nodes representing words and edges representing the relationship between words, and extracting key information through graph algorithms. In addition, machine learning-based methods use labeled data to train classifiers to determine whether sentences are important. Through a reward mechanism, the training model generates high-quality overviews. Deep learning-based methods use, for example, the Seq2Seq model: text is encoded into vectors and then decoded to generate overviews.

4.5.2. Experiments and results on text overview

We used an article by the author of this paper ("Trust Construction in Commercial Transactions in the Mobile Internet Era—Based on an Investigation of WeChat Group

Buying Groups in the J Community") [12] as a sample and generated a summary of the article. In the test, we chose to generate 8% of the full text as the proportion parameter. In actual applications, this percentage parameter is user-controllable and can generate appropriate summary content according to actual needs. The results show that the important points in the paper are well extracted, far exceeding the amount of information provided by the paper abstract. Compared with reading tens of thousands of words of the original text, reading the abbreviated summary of the text can indeed help researchers save a lot of time and energy and focus on innovation and practice.

4.6. Intelligent Question Answering in Intelligent Analysis of Scientific and Technological Literature

4.6.1. Theoretical basis of intelligent question answering

Intelligent question answering is an important function of the system. It uses the large language model (LLM) to allow users to ask questions to the system and obtain accurate and relevant answers [13]. Users do not need to read the literature one by one, but only need to ask questions to quickly obtain the required information. The system can answer various questions raised by users, including factual questions, conceptual questions, comparative questions, etc. through its deep understanding of the text. Personalized service: The system can provide personalized question-and-answer services based on the user's question history and interest preferences. The technical implementation of intelligent question answering benefits from natural language understanding (NLU) and uses knowledge graphs to extract entities and relationships from a large number of documents to build a knowledge graph. The user's question is converted into a query statement on the knowledge graph, and the answer is obtained from the knowledge graph. The machine learning behind intelligent question answering uses a large number of question-and-answer dialogues to train the machine learning model and improve the accuracy of the model. The user's question is input into the model to obtain the model's prediction result. The currently well-known ChatGPT is a conversational large language model developed by OpenAI that is good at generating human-level text. It can be used to build a more natural and fluent dialogue system to improve the user experience.

4.6.2. Experiments and Results of Intelligent Question Answering

An important module in the intelligent analysis system of scientific literature is the application interface module of the LLM model, so as to call the API in the LLM. The latest Google Gemini is a large language model application developed by Google, which has powerful text generation, translation, code writing and information retrieval capabilities. In the intelligent question-answering system, Gemini can be used to understand users' questions more accurately and generate more comprehensive and logical answers. use [14] as the theme (the development trend of foreign Internet trust research under the guidance of technical logic - recognition and visualization analysis based on CiteSpace) to let Gemini provide relevant information. The test results show that the answers obtained have considerable accuracy. The quality of the answers is also very high, covering the highlights of the paper, giving the paper ideas, research methods, research characteristics, and refining the research and development trends of the overview in the paper. The advantages of the large model are well reflected.

5. CONCLUSION

By reviewing traditional literature analysis tools, this paper proposes intelligent analysis of scientific and technological literature and discusses in detail this emerging research paradigm assisted by artificial intelligence. It reveals the great potential of intelligent analysis of scientific and technological literature in improving research efficiency and depth. The proposal of intelligent analysis of scientific and technological literature provides a new perspective for quantifying the correlation between literature and provides researchers with more precise guidance.

From the technical implementation level, the use of technologies such as natural language processing and machine learning provides favorable support for the intelligent analysis of scientific and technological literature. We have designed a proof-of-concept system, conducted an in-depth discussion on some core functions, and provided test and demonstration effects. In the future, with the continuous advancement of technology, intelligent analysis of scientific and technological literature is expected to achieve more complex and refined literature analysis, bringing revolutionary changes to scientific research. We need to continuously improve relevant technologies, strengthen human-computer collaboration, and actively respond to possible challenges. We believe that intelligent analysis of scientific and technological literature is a major change in the paradigm of scientific research. Through continuous exploration and innovation, we have reason to believe that intelligent analysis of scientific and technological literature will become an indispensable research tool for researchers and promote scientific research to new heights.

REFERENCES

[1]. George, P. M., & Robbins, K. (1994). Reference accuracy in the dermatologic literature. Journal of the American Academy of Dermatology, 31(1), 61-64.

[2]. Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El Profesional de la Información*, 29(1), e290103.

[3]. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arXiv preprint arXiv:2402.06196.

[4]. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 21-25) . IEEE.

[5]. Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. Computer Science Review, 29, 21-43.

[6]. Wang, D., & Zhang, M. (2021). Artificial intelligence in optical communications: from machine learning to deep learning. Frontiers in Communications and Networks, 2, 656786..

[7]. Chatterjee, N., & Mohan, S. (2008, February). Discovering word senses from text using random indexing. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 299-310). Berlin, Heidelberg: Springer Berlin Heidelberg.

[8]. Leon-Perez, J. M., Notelaers, G., & Leon-Rubio, J. M. (2016). Assessing the effectiveness of conflict management training in a health sector organization: evidence from subjective and objective indicators. European Journal of Work and Organizational Psychology, 25(1), 1-12.

[9]. Brown, PF, de Souza, PV, Mercer, RL, Della Pietra, VJ and Lai, JC (1992) Class-Based n- Gram Models of Natural Language. Computational Linguistics , 18, 467-479.

[10]. Keith, B., Fuentes, E. and Meneses, C. (2017) A Hybrid Approach for Sentiment Analysis Applied to Paper. Proceedings of ACM SIGKDD Conference , Halifax, 13-17 August 2017, 1-1011. Gupta, L., Jain, R., & Agrawal, S. (2020). A survey on 5G network: Architecture and emerging technologies. IEEE Access, 7, 75415-75443.

[11]. Knight, K. and Marcu, D. (2002) Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. Artificial Intelligence , 139, 91-107.

[12]. Huang Xiaoye. Trust construction in commercial transactions in the mobile Internet era: Based on an investigation of the J community WeChat group buying group[J]. Journal of China University of Mining and Technology (Social Sciences Edition), 2023, 25(3): 91-104.

[13]. Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, 1- 5 November 2016, 2383-2392

[14]. Huang Xiaoye. The development trend of foreign Internet trust research under the guidance of technical logic: Identification and visualization analysis based on CiteSpace[J]. Journal of Jiangnan University (Humanities and Social Sciences Edition), 2023, 22(2): 52-65.