Annals of Emerging Technologies in Computing (AETiC)
Vol. 9, No. 4, 2025

Research Article

Optimization of University Library Services through Big Data and Multi-source Data Fusion

DiYin Zhu

Shaoxing University Yuanpei College, Shaoxing, 312000, China
zdy124561@163.com

Received: 16 September 2024; Accepted: 23 September 2025; Published: 1 October 2025

Abstract: The advent of the big data era has not only advanced the informatization of libraries but also opened unprecedented opportunities for their sustainable development. Libraries are no longer limited to traditional resource management; instead, they have embraced emerging technologies such as Web 2.0, mobile solutions, cloud computing, resource discovery systems, and big data platforms. While these developments provide a solid technological foundation, libraries must further enhance their ability to conduct data analysis, semantic processing, decision-making, and visualization in order to respond effectively to evolving user demands and complex information environments. This study contributes to that goal by discussing the application of multi-source data fusion in science and technology decision-making. It presents a comprehensive decision support framework that integrates semantic preprocessing techniques—including data cleaning, partition segmentation, and synonym merging—supported by Python's Pandas library and Jieba's text-cutting functions. Through this approach, the research successfully identified six science and technology text clusters and three mass technology-related clusters, thereby providing a refined view of user information needs and thematic structures within large-scale datasets. The findings demonstrate that a decision support framework based on multi-source data fusion can proactively detect and respond to user needs, moving libraries from passive service providers to active, intelligent participants in knowledge dissemination. This proactive transformation enriches the quality of information services, enables accurate and personalized decision support, and aligns with the demands of the new era defined by innovation-driven and intelligence-first strategies. Ultimately, this work highlights the value of integrating big data technologies into library management and decision-making systems. By bridging semantic analysis with multi-source data fusion, libraries can evolve into dynamic hubs of innovation, offering precise, context-aware services that not only enhance user satisfaction but also strengthen their role in supporting scientific research, technological advancement, and informed decision-making in the digital age.

Keywords: ***Big data integration; Decision support systems; Intelligent library systems; Multi-source data fusion; University library services***

1. Introduction

The development of internet technology and computers has allowed extensive data technology to be exploited in all aspects of life. To empower college learners to enhance their knowledge levels and expand their fields of knowledge, information construction in university libraries is important for college students' abilities and the sustainable growth of libraries. The advent of the extensive data era not only creates conditions for the informatization of libraries but also presents significant opportunities for their advancement. The idea of "All information for all people at all times" has become a reality with advancements in information technology. With over 6 billion mobile phone users having immense computing power, libraries have rapidly evolved, with recent changes surpassing those of the past century and predicting even faster future advancements. Libraries have embraced technological developments like

DiYin Zhu, "Optimization of University Library Services through Big Data and Multi-source Data Fusion", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 1-16, Vol. 9, No. 4, 1 October 2025, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2025.04.001, Available: http://aetic.theiaer.org/archive/v9/v9n4/p1.html.

Web 2.0, mobile library solutions, cloud computing, resource discovery, and big data. Big data, in particular, brings challenges in extracting valuable knowledge, necessitating enhancements in data analysis, decision-making, semantic analysis, and visualization for better library management. Digital libraries, evolving from the concept of electronic libraries proposed in 1975, use modern IT to digitize resources and services. They perform data resource processing, storage management, and access services, utilizing servers, cloud platforms, linked data, semantic analysis, personalization, and visualization technologies [1].

Garoufallou and Gaitanou [2] focused on data acquisition, storage, search, organization, and visualization as they evaluated the benefits and problems that "Big Data" presents for libraries. The actual difficulty is extracting value from this data to enhance services. Libraries, traditionally information handlers and technology adopters, must determine their role in managing Big Data to develop new services, often requiring external expertise. In order to examine librarians' roles in the Big Data age and potential directions for future study, the paper conducts a thorough analysis of the literature, encompassing publications from 2012 to 2018. It categorizes the literature based on library types, identifying new roles and providing tables to help librarians find relevant articles for practice and service development.

The core elements of current university libraries are users and resources, and traditional libraries will gradually transform into smarter ones. Intelligent libraries can provide users with autonomous services, such as using intelligent awareness technology to analyze user behavior characteristics and provide them with intelligent guidance. Additionally, libraries in the Big Data era have expanded active services, such as analyzing the paths of user knowledge activities to supply users with precise information, including the latest hotspots of recent interest or potential hotspot events. The knowledge and information data involved in university libraries are complex, diverse, and changeable. Library services supported by Big Data are increasingly rich in application content, diverse in data types, and processed in various ways [3].

Hussain [4] explored the emerging role of artificial intelligence (AI) in library services, emphasizing its capability to improve intelligent decision-making and transform services in the information technology age. The paper reviewed the body of research using qualitative content analysis, highlighting the advantages and difficulties of using AI in libraries. Findings indicate that while AI can significantly improve library operations, obstacles such as funding, librarian attitudes, and technical skills need to be addressed. The paper suggests that AI applications can be deployed cost-effectively and emphasizes the need for librarians to embrace AI to accelerate and improve library services. The study aims to inform policy stakeholders, librarians, and scholars about the practical and social implications of AI in library services.

Umeozor and Ibegwam. [5] assessed the roles of Nigerian federal university libraries using a survey of 43 libraries. Traditional roles focused on acquiring, preserving, and providing access to information. Only 7 out of 18 potential emerging roles related to the internet and ICT were recognized, indicating slow adaptation to new technologies. Administrative and personnel issues were major challenges. The study concluded that while libraries are aware of their traditional roles, they need to better utilize modern technologies. It recommended adopting internet, web, and ICT technologies and providing relevant training to improve service delivery. According to Hamad *et al.* [6], academic libraries are being forced to embrace new technology in order to satisfy the changing demands of tech-savvy patrons who need quick and convenient access to information from a distance. The application of smart technology in Jordanian academic libraries and the difficulties in creating and offering these services were the main topics of the study. Data were collected via a questionnaire from 246 library staff members out of 340 surveyed. The findings showed a middling degree of smart service implementation and challenges, with resistance to change and issues related to privacy and confidentiality being the primary obstacles. Financial constraints, insufficient personnel training, and subpar facilities were also significant difficulties. The adoption and delivery of smart services in Jordanian academic libraries were severely harmed by these difficulties. The research provides guidance on how to plan and get past roadblocks when implementing smart services for librarians and policymakers.

A rising topic called "Big Data" presents information technology issues related to data collection, archiving, searching, organizing, and visualization. Organizations must find ways to extract value from Big Data to improve services. Large-scale, complicated data is produced by academic and other organizations, which presents new difficulties for libraries. Libraries have to decide how to handle Big Data and provide services related to it, which frequently calls for outside expertise. Libraries have historically managed information and adopted new technologies, so Big Data will impact their context. Future research paths and the duties of librarians in the Big Data age are only two of the topics covered in some of the studies that have used systematic literature reviews to investigate these concerns [7], [8].

Additionally, the development of higher education in China is vibrant, with a steady increase in the number and size of universities and colleges, along with a growing student population. Libraries, as key institutions in these academic settings, play a crucial role. Enhancing library services and advancing their information capabilities are essential tasks. However, many university libraries struggle with data service innovation, limiting their overall service improvement. There are significant weaknesses in the comprehensive utilization of big data technology, with inadequate integration of data service resources, informatization, networking, and intelligence. Consequently, this hampers the ability to provide targeted and effective data services for faculty and students. Naeem *et al.* [9] focused on the five qualities of big data—volume, velocity, veracity, variety, and value—while analyzing the potential and difficulties it presents. While significant research has addressed the issues of volume and velocity, a comprehensive solution for the variety of data types remains elusive. Traditional DBMS solutions often struggle with new, incompatible data types. Big data analytics, which involves analyzing large datasets to gain insights and identify patterns, is crucial for modern organizations to make informed decisions. This paper discusses the current issues, trends, and challenges in big data, with a particular emphasis on addressing a variety of problems.

Optimization pathways for university library data services encompass innovative approaches and strategic developments. These include the conceptualization of novel data services, the establishment of comprehensive data service platforms, and the expansion into diverse service domains. To enhance data services, university libraries should leverage big data technologies, embrace an innovative data service philosophy, and adopt a "person-centered" service approach. To meet user requirements, this entails encouraging the creation of personalized, distinctive, interactive, and interconnected data services. Such initiatives not only foster the innovation of the university library data service mechanism but also improve the overall data service system [10]. Currently, most university libraries utilize data service systems in their operations. It is imperative to enhance research on data warehousing, OLAP (Online Analytical Processing), and data control technologies, and to integrate these advancements into data service activities. Libraries can effectively handle the gathering, organizing, processing, and use of data by utilizing a large-scale data platform. Techniques such as data mining can be used to assess data services and generate analysis reports, facilitating the sharing of data resources across various platforms, including digital libraries, big data platforms, mobile applications, WeChat applets, campus networks, WeChat official accounts, and microblogs. The scope of application for extensive data techniques is broad. Big data not only creates favorable conditions for university library data service research but also provides support for data service model innovation. University libraries need to constantly expand their field and scope in the process of carrying out data services, especially to comprehensively apply big data technology, cloud technology, and AI technology for data service innovation [11].

The main research work of this study is divided into two parts. One part constructs a scientific and technological decision-support method framework, while the other conducts empirical research on these methods using the constructed framework. The focus is on institutional users of the Ministry of Science and Technology, identifying their scientific and technological decision-making needs and supporting intelligence. This research has developed an analysis framework for a "scientific and technological decision-making support method based on multi-source data fusion," which includes three aspects: multi-source data acquisition and extraction methods, detection methods for users' scientific and technological decision-making needs, and information mining and analysis methods for decision support.

The origin of the investigation's empirical data was the website data of the Ministry of Science and Technology of the People's Republic of China, mass media, national leaders' speeches, reports, meetings, policies, and scientific and technological literature. Using semantic association as a starting point, the research aims to analyze the decision-making needs of information users within the Ministry of Science and Technology. It employs data mining, natural language processing, and information analysis methods, along with scientific measurement and social network analysis techniques. The goal is to find information that supports users' scientific and technological decisions from relevant documents and to present a series of decision support methods.

Based on the theory of embodied cognition, this study systematically discusses the information services of university libraries using the following research methods:

1. **Literature Research Method**: Relevant studies on university library information services are sorted and analyzed, comprehensively collecting findings from the intersection of information science and psychology. This establishes the foundation for the study by reviewing existing research results.
2. **Theoretical Research Methods**: The study systematically researches theoretical knowledge related to precision information service and embodied cognition theory. It includes a thorough theoretical analysis of information service processes and the factors affecting embodied cognition, providing a theoretical basis for subsequent research.
3. **Case Analysis Method**: Typical cases in the construction of libraries and think tanks are selected for study. Examples include the Library of Congress, the Hoover Institution, the National Library, and the Shanghai Library. The analysis focuses on the successful construction and transformation of these libraries into think tanks, examining the directions and requirements for functional innovation and enhancement. This provides references and insights for other libraries seeking to transform their functions and develop into think tanks.
4. **Network Survey Method**: The study investigates the construction of university library think tanks by visiting 50 university library portals via the Internet. The survey examines the construction of institutional knowledge bases, embedded scientific research services, scientific research data management, information consulting, and decision-making services based on the library service content provided on the navigation and web pages of the university library portals.

2. Analysis of Library Data Resource Fusion Mode under the Background of Three Big Data

2.1. Overall analysis of library data resource integration under the background of big data

Within the field of big data, libraries are gradually revealing the properties of extensive data, thereby endowing their digital resources with new meanings. Not only is the volume of data growing explosively, but the variety of data types and sources is also expanding. Library digital resource integration involves the consolidation of library big data, emphasizing a platform-based integration that aims to merge technology, platforms, and services through data unification. The ultimate goal is to optimize the contribution and utilization of library data assets, providing users with high-quality services. The specific objectives include seamless linking, association analysis, knowledge discovery, and service innovation [2]. A key shift in thinking in the era of large-scale data is the focus on relevance rather than causation. The goal of association analysis is to find repeated themes, relationships, or causal structures in the set of items or objects found in relational data, transaction data, or other types of information carriers. This analysis, also known as association mining, aims to reveal the connections between different items in the transaction database. Since its inception in 2006, the importance of relevant data has been recognized by the library community. For example, the Library of Congress has fully integrated the Title List of Congress (LCSH) into a Simple Knowledge Organization System (SKOS) format, available for download, demonstrating the successful application of relevant data in the library [12]. The process of knowledge discovery involves identifying important, innovative, potentially valuable, and understandable patterns from large amounts of data. The process consists of three parts: preparing data, data mining, and expressing and interpreting findings. The basic step in this process is data mining. The survival and progress of the library are closely related to its service delivery. Expanding service function and improving service quality are the ultimate goal of all library activities [13].

2.2. Specific content of library digital resource integration in the context of big data

The integration of library digital resources involves combining extensive datasets into a cohesive and valuable whole. This comprehensive fusion includes both temporal and spatial dimensions, encompassing all data and information related to the library. By merging these elements, libraries can develop and utilize their resources to gain greater value. From a temporal perspective, this integration refers to the unification of data accumulated over time, including past, present, and future records. Spatially, it involves the consolidation of data generated by different regions and various types of libraries. This includes library resource data, business process data, user information, industry data, and external data relevant to the library. In a big data environment, this integration is characterized by "multi-source" data, meaning multiple sources record the same object from different angles, allowing for mutual confirmation of data. Therefore, the integration of library digital resources also includes the incorporation of data from the same subject from different sources, the combination of various types of library-centric data, and the integration of information from other institutions such as museums and archives [14].

2.3. Analysis of the level of library digital resource integration under the background of big data

"Big data acquisition, import preprocessing, statistical analysis, mining" the complete process mainly involves three aspects: data base (the original and complete collected data), facilities and technology (data storage, data processing, data mining, such as the required infrastructure and technology), data utilization (to solve specific problems and to achieve the goal). Based on this, the library digital resource fusion system in the era of massive data can be divided into three levels: data fusion layer, platform fusion layer and service fusion layer. Among them, data fusion is the foundation, platform fusion is the key, and service fusion is the purpose. Integrating large amounts of heterogeneous data on the platform is an innovation in itself. On this basis, the library digital resource integration and service platform will provide new vitality for libraries to provide information services in a wide range of data environment, and can improve the competitiveness of libraries in providing information services [15]. The complete library digital resource integration system in the big data environment is shown in Figure 1 [16].

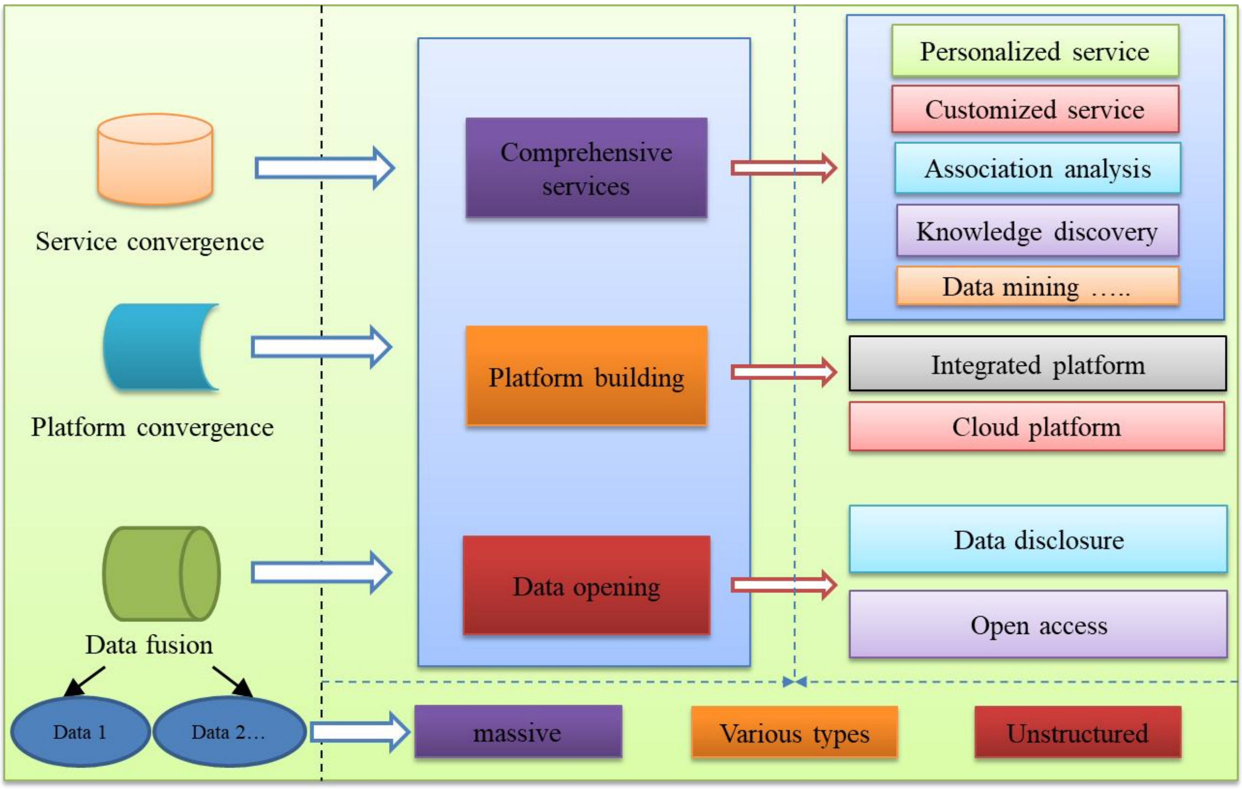

Figure 1. Data resource fusion system of the university library under the background of extensive data

3. Multi-source data fusion based framework for scientific and technological decision-making support

This paper around science and technology information user "technology decision support" research purpose, put forward a technology decision support framework, from the "multi-source data acquisition and extraction method", "user decision demand detection method" and "science and technology decision support information mining and analysis method" three aspects to build the "based on multi-source data fusion of science and technology decision support method framework"

"Multi-source data collection and extraction method" is the premise and basis of "technology decision support method based on multi-source data fusion". To support technological decisions based on multi-source data, relevant multi-source data must be obtained from intelligence users to discover their decision needs and collect data to support their decisions. Therefore, obtaining multi-source data of demand sets and support sets is the basis of decision support based on multi-source data.

The "user scientific and technological decision-making demand detection method" is the core technical component of the "scientific and technological decision-making support method based on multi-source data fusion." To support the user's decision-making, it is crucial to understand their needs. This involves identifying the user's focus topics and thematic attributes through methods such as text mining and natural language processing. A correlation analysis of the user's decision-making topics and attributes is then conducted to detect their needs accurately.

The "information mining and analysis method for scientific and technological decision-making support information" is a vital part of the decision-making support method. By identifying user needs and combining the "acquisition and extraction methods of multi-source data" with scientific measurement and visualization methods, the needs of scientific and technological decision-making can be matched and correlated, thus fulfilling the investigation reason of "scientific and technological decision support" for intelligence users.

Point mutual information (PMI) is utilized for word segmentation optimization by computing the common occurrence data of adjacent words in the corpus. PMI is primarily used to determine the semantic similarity among words. It computes the likelihood of two words occurring in the text at the same time in order to function. The higher this probability, the stronger the correlation. The PMI value reflects the frequency of co-occurrence or the frequency of use between adjacent strings.

The PMI values of the two-word strings w_1 and w_2 are computed as shown in Equation (1):

$$PMI\ (word1,word2)\ = log_2\,(\frac{P(word1word2)}{P(word1)P(word2)}) \tag{1}$$

If PMI (word1, word2) > 0, the two words are connected. The higher the PMI value, the stronger the inter-word connection. For example, associated strings with PMI values greater than N can be selected for merging, a custom vocabulary can be added, and refraction can be optimized.

With the growing size of the Internet and the rapid expansion of various text information, it often consumes significant time and energy to screen and select large amounts of information. Keywords embody the core content of a text and concentrate the thematic information. By reading the keywords, one can quickly identify the theme of an article and extract useful information from numerous articles. Manually extracting keywords is time-consuming and laborious, with results varying from person to person. Therefore, achieving automatic extraction is important. Keyword extraction is a fundamental technology in the field of information processing, extensively utilized in information retrieval, automatic summarization, text clustering, classification, and other areas.

To detect the topics that users focus on and understand the requirements of information users, keywords can be extracted from the text to find the focus and particularity of the requirements in each corpus. Word frequency statistics is a research method for vocabulary analysis. It can describe vocabulary rules through the statistics and analysis of word frequency in a certain length of text. The word frequency in a text indicates its representativeness; the higher the word frequency, the more representative the text content. Term frequency-inverse document frequency, or TF-IDF, is a statistical measure used in information retrieval that attempts to capture a word's significance to documents within a corpus or collection.

In this paper, TF represents how often the words appear in the document. It is expressed by the following equation (2):

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{j,k}} \tag{2}$$

Equation (2) denotes that the denominator is the total number of times each word appears in the file, while $n_{i,j}$ is the number of times the word d_j appears in the file. The relevance of the words is gauged by the IDF (inverse document frequency). To find the IDF of a word, divide the total number of files by the number of files that include the word. Then, take the base-10 logarithm of the resulting quotient. The formula is as follows:

$$IDF_i= log\,\frac{|D|}{|\{j{:}t_i \in d_j\}|} \tag{3}$$

Here, $\left|\{j:t_i \in d_j\}\right|$ represents the number of files that include the word, and $|D|$ denotes the total number of files in the corpus. The denominator will be 0 if the word is absent from the data. As a result, the expression $1+\left|\{j:t_i \in d_j\}\right|$ is typically utilized and is as follows (4):

$$TFIDF_{i,\ j} = \frac{T}{F_{i,j}} \times IDF_i \tag{4}$$

A high-weight TF-IDF may be produced by combining a word's high frequency in one file with its low frequency in the entire collection of files. As a result, TF-IDF often eliminates unnecessary words while keeping crucial ones [16].

The science and technology decision support method in this study mainly focuses on institutional users, identifying the topics and subject attributes of user needs to detect decision-making requirements. For selecting the topics of user needs, after identifying the user requirements, it becomes essential to integrate the domain themes of other science and technology-related multi-source data, consolidate multi-source data, and conduct consistency and difference analysis. This helps identify the demand topics of the target users for the decision support service. The analysis of consistency and difference requires scoring the text for consistency after keyword identification, as shown in Formula 5:

$$F_m=100^* \textstyle\sum_{i=1}^{n} w_i \cdot Rank_i \cdot (TFIDF(m)) \tag{5}$$

Where F_m represents the consistency score of the keyword m, n represents the number of data sources, w_i signifies the weight of each data source, $TFIDF(m)$ is the TFIDF weight of keyword m in data source i, and $Rank_i$ represents the percentile score of the "TFIDF" value of keyword m sorted in order in data source i.

The total score of F_m is 100, with higher scores indicating better consistency. This formula measures the consistency score of keywords across the entire data source. In this study, different data sources are considered equally important. If the size of the data source text were considered, the sorting results would be too skewed due to the differences in the number of texts from different data sources. Therefore, the data source weight w_i is considered equally important and given an equal value. Since this study uses four data sources, the weight of each data source is 0.25 [17, 18].

4. Empirical research on the support of intelligence in big data technology decision-making

This section focuses on the textual data gathered from the Ministry of Science and Technology, science and technology conferences, and related social science and technology data. It undergoes semantic pre-processing, which includes data cleaning, word segmentation, synonym merging, and the removal of stop words. Using the TF-IDF keyword extraction approach on the basis of word frequency statistics, the extracted keywords are subjected to K-means clustering analysis [19]. The goal is to identify science and technology-related topics across various data sources and analyze these topics in detail at three different levels within the Ministry of Science and Technology institutions. This analysis serves as a critical tool for identifying the decision-making needs related to science and technology within the Ministry.

Subsequently, with the help of the Word2Vec word vector model, the context of the subject words in each attribute corpus is learned, and the words are represented as word vectors. The distance between word vectors in the vector space is calculated to determine the similarity between words and hence the thematic nature of the user's technological decisions. The paper analyzes the correlation between the themes and attributes of the specific decision needs of institutional users of SIT [20].

4.1. Data acquisition and preprocessing

The Ministry of Science and Technology's institutional users served as the study's science and technology decision support objects. Thus, in addition to other pertinent multi-source data, data from the Ministry of Science and Technology's institutional user website had to be crawled. These comprised proceedings of members of the Scientific and Technological Congress, national leadership presentations, conference and activity statistics, National Science and Technology Award conference materials, National Science and Technology Innovation conference materials, and Xinhua Net resources.

Given that the scientific and technological decision-making support objects of this study are the institutional users of the Ministry of Science and Technology, it was essential to gather data from their institutional user website and other relevant multi-source data. This included national leadership speeches, meeting and activity data, academician conference data from the science and technology conference, National Science and Technology Awards Conference data, National Science and Technology Innovation Conference data, and Xinhuanet data [21, 22].

4.1.1. Data Cleaning

Using Python's Pandas package, Excel data were read and merged into rows by department to create four DataFrame datasets. The drop_duplicates function in Pandas was used to remove duplicates from the dataset. For the analysis of user intelligence needs, the title and content fields in the dataset were primarily utilized. These fields were extracted to generate two datasets. The first dataset comprised DataFrame structures with title and content fields by department. The second dataset concatenated content and titles into paragraphs, with each piece of content or title separated by a period. This resulted in a single piece of text for each department, and the second dataset was processed first.

4.1.2. Word Segmentation

The words in the text data were segmented using the precision option of the jieba.cut function.

4.1.3. Word Segmentation Optimization

For optimization, synonym synthesis with Synonym Forest and PMI mutual information was used to semantically merge words. The words for merging information are illustrated in Fig. 2.

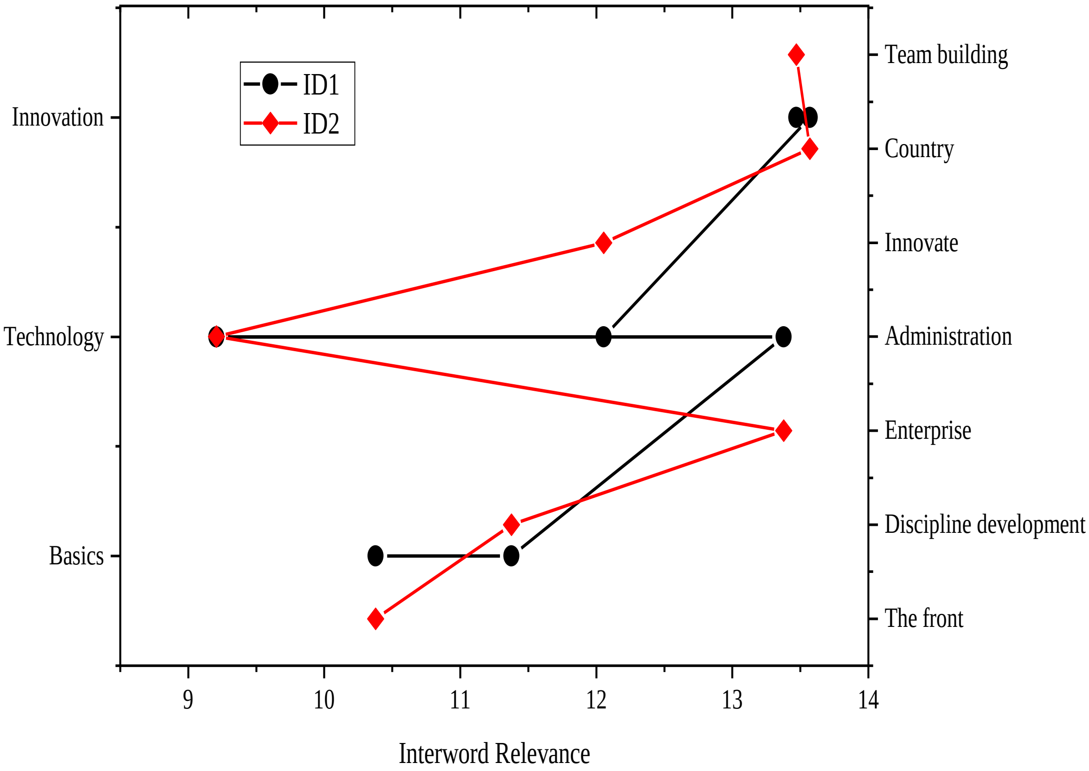

Figure 2. The correlation between big data words

4.1.4. Remove Stop Words

In vocabulary processing, word frequency and phrases are standardized, consisting of eliminating low-frequency and stop words. The Harbin Institute of Technology Stop Words Glossary is used to eliminate these low-frequency and stop words. Additionally, customizing the vocabulary involves adding other synonym thesauri, domain dictionaries, and official document subject word lists. After this step, the text is segmented again to obtain the final optimized segmentation results.

4.1.5. TF-IDF Keyword Extraction Based on Word Frequency Statistics

To understand the needs of information users, the focus and particularity of user needs at different levels can be identified by extracting keywords from the text. This study employs TF-IDF based on word frequency statistics to extract keywords. The TF>2 strategy is adopted for calculation. An example of TF-IDF keyword statistics and performance indicators based on word frequency statistics is shown in Fig. 3.

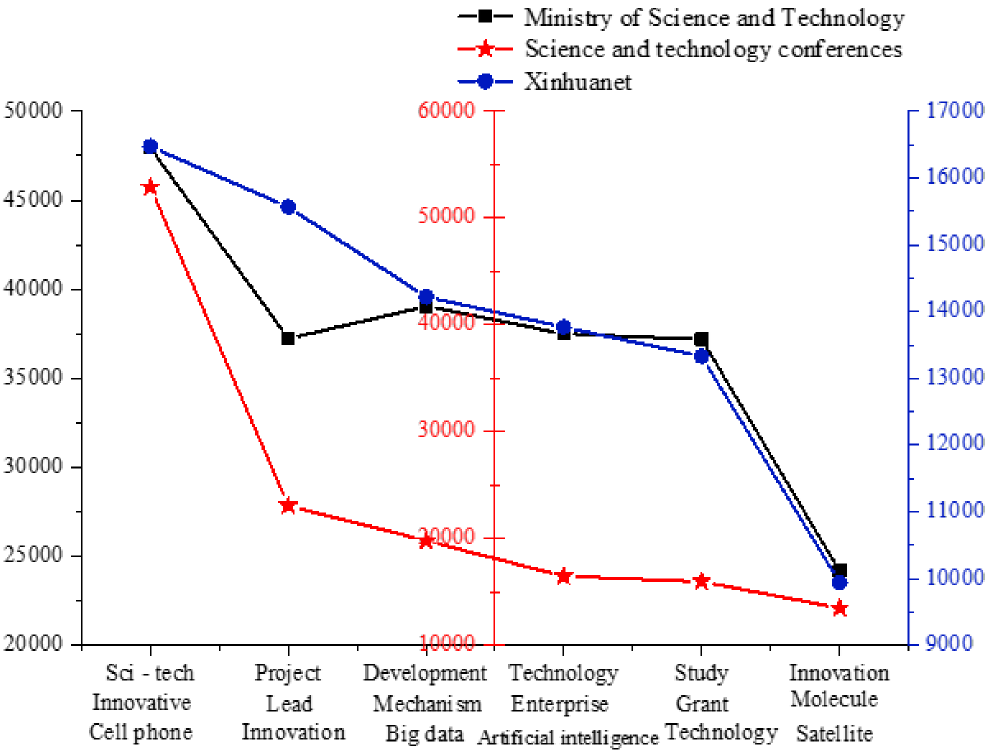

Figure 3. TF-IDF keyword statistics example and performance indicators based on word frequency statistics

4.2. Identification of user-focused topics based on multi-source data fusion in the age of big data

This section selects data from the Ministry of Science and Technology, Science and Technology conferences, and social science and technology related sources. It explores and analyzes topics of interest from these data sources. The integration of these relevant data provides a reference and supplement to identify topics of interest to users of Science and Technology.

Multi-source data fusion involves collecting information on different data structures obtained from various channels and multiple collection methods to form data sets with a uniform format. This allows the information from different sources and forms to complement each other. Therefore, this study selects data from national leadership activities, science and technology conference data and social science and technology related data to identify key technology related topics from different sources, which can supplement and guide the topics and needs of the users of the Ministry of Science and Technology in institutions [24-24].

This section connects the title and content fields of each source data and transforms it into numerical features that the model can handle using the TF-IDF keyword extraction based on word frequency statistics. It takes the silhouette coefficient to select the best k value and then applies the k-means method to cluster the text. The original index data must be standardized to ensure the validity of the conclusions. There are several techniques for data standardization, such as &quot;minimum max normalization&quot;, &quot;Z score normalization&quot; and &quot; decimal normalization&quot;. Standardizing dimensionless index evaluation values from the raw data can be used for thorough analysis and evaluation because all index values are at the same quantitative level. In this study, the raw data were transformed linearly, and then a maximum-minimum normalization method was applied to scale the study results to the [0,1] interval.

The cluster keyword tags and weights of science and technology conference texts are shown in the following figures: Cluster 0 in Fig. 4, Cluster 1 in Fig. 5, Cluster 2 in Fig. 6, Cluster 3 in Fig. 7, Cluster 4 in Fig. 8, and Cluster 5 in Fig. 9. According to the silhouette coefficient K=6, six clusters of scientific and technological texts were obtained after K-means clustering. High-frequency keywords such as "country," "innovation," and "research," which appeared together in each cluster, were removed, along with words not highly related to science and technology. Eight representative keywords were then selected as the cluster tag words for each cluster. An overview of the K-means clustering can be seen in Table 1. Additionally, the description of social mass technology-related text clustering is represented in Table 2.

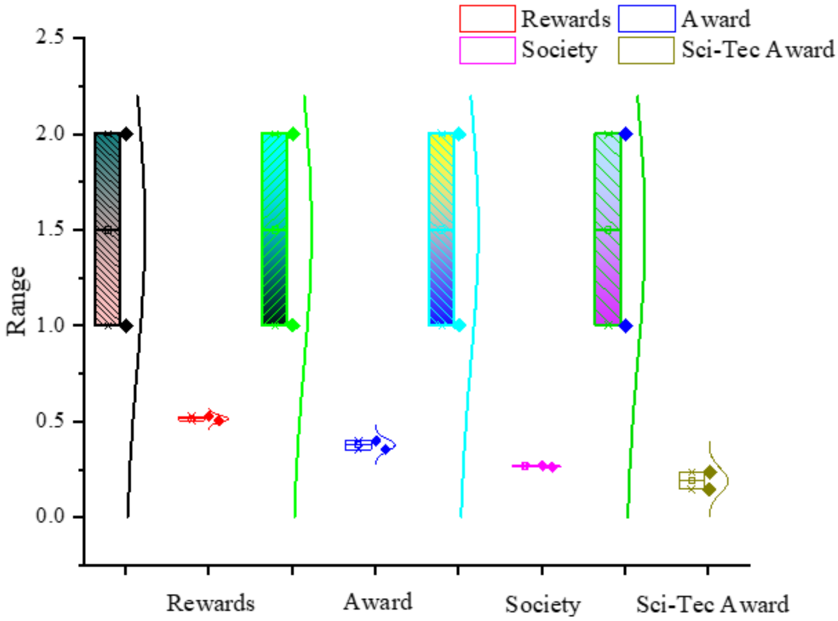

Figure 4. Science and technology conference text clustering keyword labels and weights (Cluster 0)

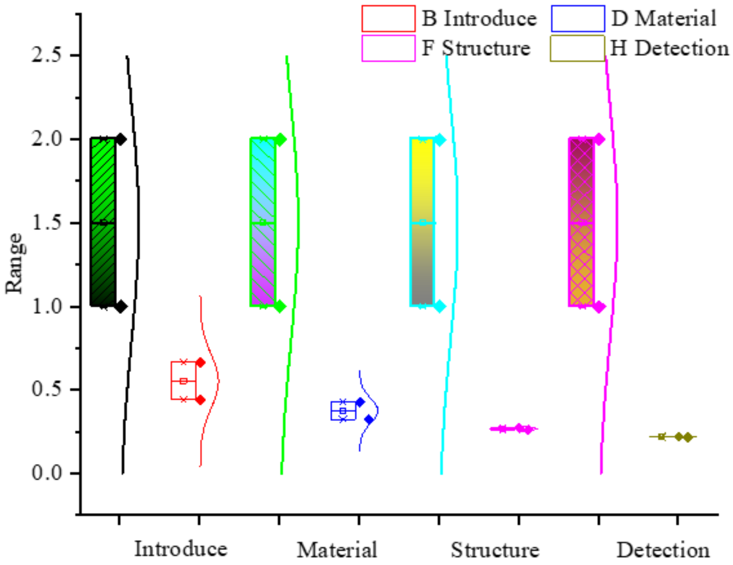

Figure 5. Science and technology conference text clustering keyword labels and weights (Cluster 1)

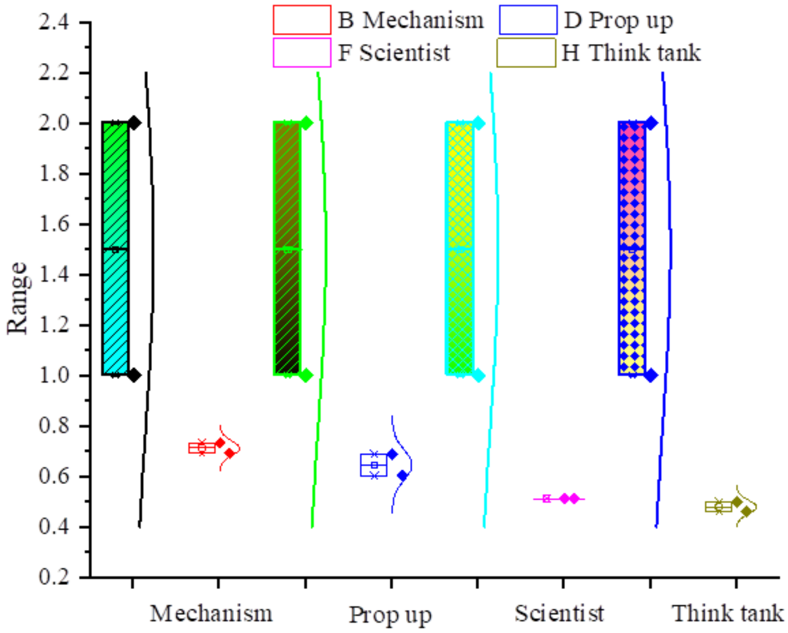

Figure 6. Science and technology conference text clustering keyword labels and weights (Cluster 2)

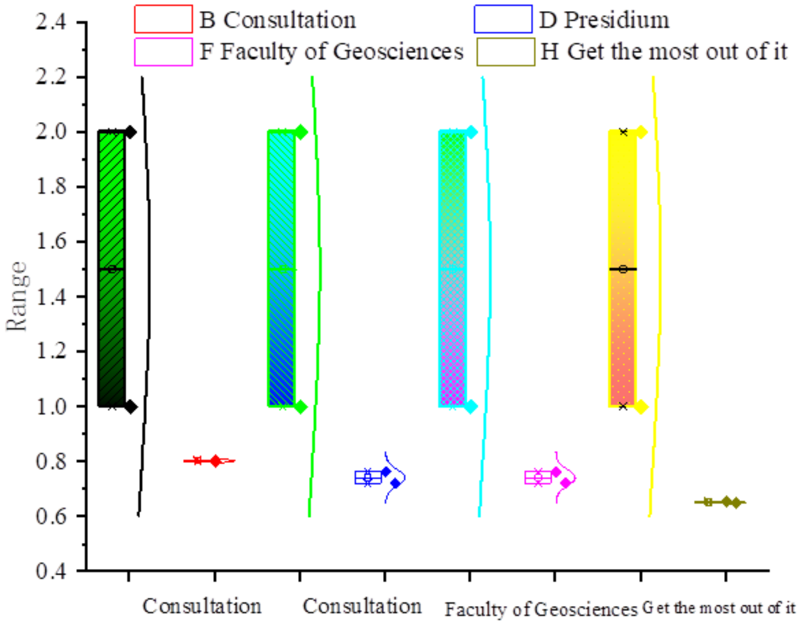

Figure 7. Science and technology conference text clustering keyword labels and weights (Cluster 3)

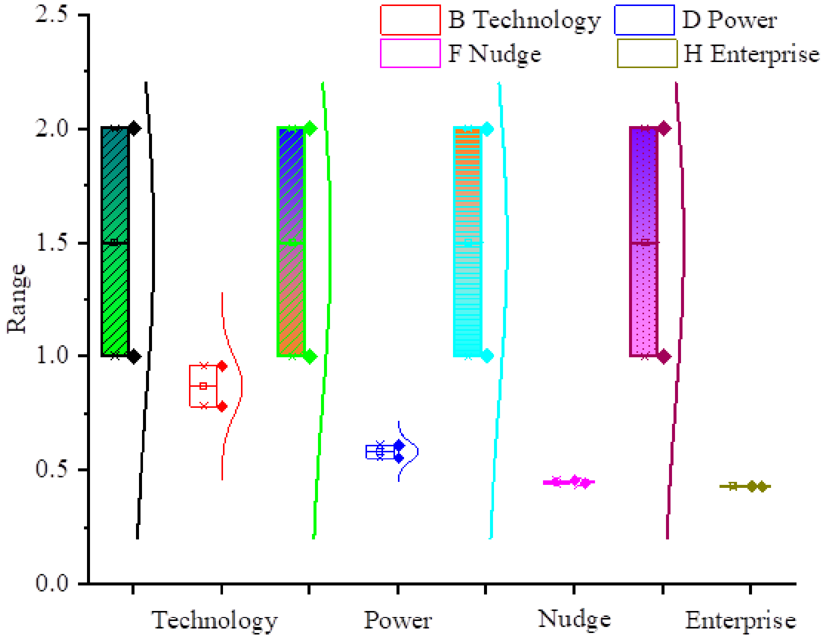

Figure 8. Science and technology conference text clustering keyword labels and weights (Cluster 4)

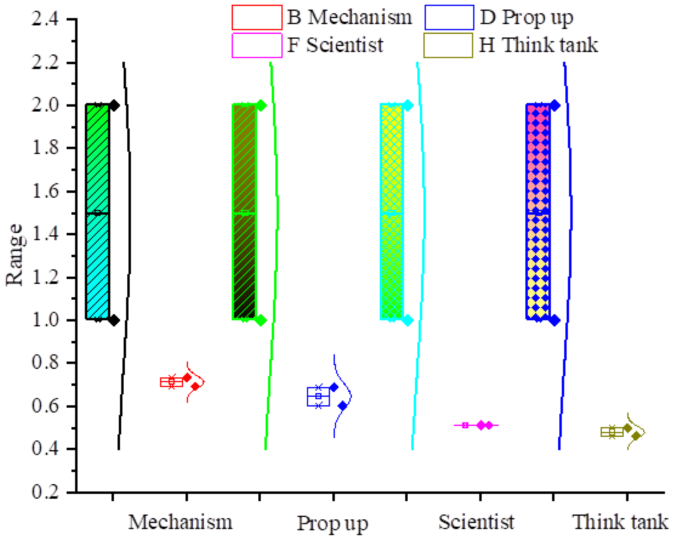

Figure 9. Science and technology conference text clustering keyword labels and weights (Cluster 5)

Table 1. K-means Clustering

Cluster	Description	Representative Keywords
Cluster 0	Reward status	Keywords related to China's science and technology award processes.
Cluster 1	Scientific research	Keywords focused on basic sciences within natural sciences and notable breakthroughs.
Cluster 2	Scientific research workers	Keywords emphasizing encouragement and rewarding of innovative achievements.
Cluster 3	Scientific and technological innovation and economic development	Keywords related to scientific and technological innovation and the knowledge economy.
Cluster 4	Discipline construction	Keywords highlighting continued strategic research in discipline development.
Cluster 5	Talent training	Keywords related to the training and development of talent.

Table 2. Social Mass Technology-Related Text Clustering

Cluster	Description	Representative Keywords
Cluster 0	New technologies	Keywords related to rapidly developing new technologies, such as shared bicycles.
Cluster 1	Technology-related data	Keywords like "Mars" and "rover," indicate focus on technology-related data.
Cluster 2	Consumption-related theme	Keywords such as "big data" and "development," related to e-commerce and Internet consumption processes.

Based on the keywords for each cluster label and combined with manual interpretation, these six categories are identified as reward status, scientific research, scientific research workers, scientific and technological innovation and economic development, discipline construction, and talent training. Since the text of the National Science and Technology Award conference comprises one-third of the total text, there is a distinct cluster (Cluster 0) dedicated to reward status. The keywords in this category indicate that China's science and technology award procedures have become more scientific and standardized, significantly encouraging scientific and technological workers to actively engage in innovation. Cluster 1 pertains to subjects related to scientific research, focusing on basic sciences within the natural sciences and notable breakthroughs in some areas. Cluster 2 encompasses groups or individuals related to scientific and technological workers, emphasizing the encouragement and rewarding of innovative achievements in scientific studies. Cluster 3 is associated with scientific and technological innovation and economic growth. Keywords, for instance, "economic system" and "science and technology" suggest that scientific and technological innovation and the knowledge economy have become the main themes of the new era, driving economic growth. Cluster 4 relates to discipline construction, highlighting the importance of continued strategic research in discipline development. This is crucial for popularizing scientific knowledge, disseminating scientific ideas, and promoting the scientific spirit.

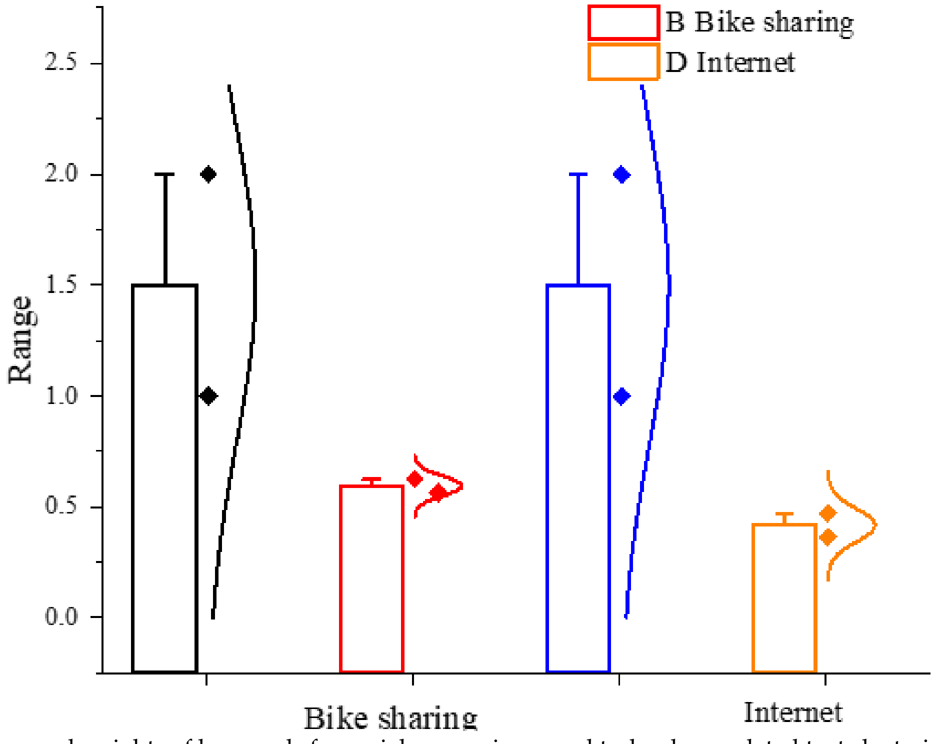

Figure 10. Tags and weights of keywords for social mass science and technology-related text clustering (Cluster 0)

The keyword tags and weights of social mass technology-related text clustering are shown in the following figures: Cluster 0 in Fig. 10, Cluster 1 in Fig. 11, and Cluster 2 in Fig. 12. According to the silhouette coefficient K=4, three clusters of scientific and technological texts were identified after K-means clustering. Remove "China", "innovation", "report" and "global", as well as words not closely related to technology. Four representative keywords were then selected as the cluster label words for each cluster. Four representative keywords were then selected as the clustering tag words for each cluster. As a mass media platform, Xinhua Network publishes content in rich language. Strong emotional and figurative words are often recognized, leading to a slightly more scattered clustering effect compared to other policy and conference texts. However, analyzing these words allows for a more accurate understanding of key science and technology issues at the public level.

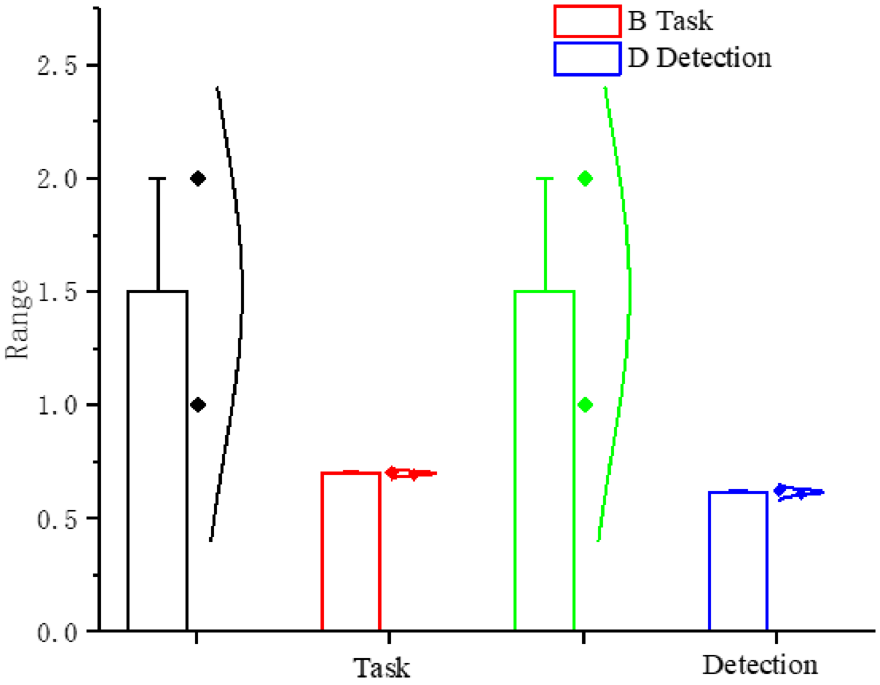

Figure 11. Tags and weights of keywords for social mass science and technology-related text clustering (Cluster 1)

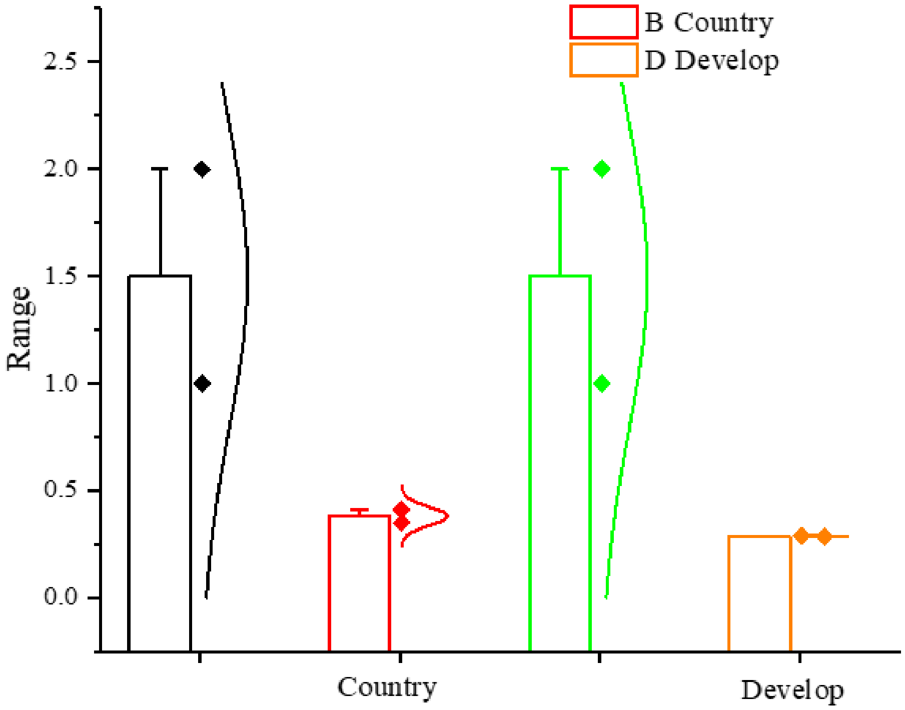

Figure 12. Tags and weights of keywords for social mass science and technology-related text clustering (Cluster 2)

Cluster 0 pertains to new technologies, such as shared bicycles, which have developed rapidly in recent years and are closely connected to people's lives, resulting in increased news coverage. Cluster 1 focuses on technology-related data, as indicated by keywords like "Mars" and "rover." Cluster 2 centers on a consumption-related theme. Keywords such as "big data" and "development" suggest that this cluster mostly pertains to the services and processes involved in the consumption process. Given that the original text is technology-related content from Xinhuanet, the consumption theme is closely tied to e-commerce and the Internet.

5. Conclusion

This study summarizes the current elements and processes of multi-source data fusion, scientific and technological decision support, and related methods at home and abroad. Constructed a comprehensive set of "science and technology decision support method based on multi-source data fusion", and established the framework of "science and technology decision support method based on multi-source data integration under the background of big data", including multi-source data acquisition and extraction, user technology decision-making demand detection, science and technology decision support intelligence information mining and analysis methods. In addition, this study determines the needs of institutional customers of the Ministry by analyzing website data and combining multi-source data such as mass media, reports, national policy data and science and technology conferences. Expanding the channels for collecting user needs from these different sources compensates for the limitations of relying on a single data source for scientific and technological decisions. It further reveals the scientific and intellectual value of the data. By determining the topics of users' concerns based on multi-source data fusion in the era of big data, the research found that the mention of big data at the technical level is mainly within the framework of introducing big data policies. Public attention to big data often involves explaining the large amount of data of public concern, and is often associated with artificial intelligence and the Internet. The converged analysis of these different data sources enables information to complement each other, providing crucial guidance for identifying topics of scientific and technological issues and decision needs. The integration of multi-source data in university libraries is complex, leading to increased processing time and potential inaccuracies, and scalability issues require substantial computational resources. Technical limitations such as funding, technical skills, and infrastructure limit the deployment of AI and edge computing, while privacy and security issues require strong data protection measures. Methodological challenges include the need for more longitudinal studies to enhance the robustness of the findings. Future research should develop sophisticated data fusion technologies, address scalability issues through advanced data processing technologies, and integrate artificial intelligence, machine learning, and the Internet of Things to enhance decision support systems. Ensuring strong privacy measures with an emphasis on longitudinal research and integrated assessment frameworks will improve the reliability and durability of big data utilization in libraries.

CRediT Author Contribution Statement

DiYin Zhu: Writing-Original draft preparation, Conceptualization, Project administration.

References

[1] Shuqing Li, Zhiyuan Hao, Li Ding and Xia Xu, “Research on the application of information technology of Big Data in Chinese digital library”, *Library Management*, Online ISSN: 0143-5124, Vol. 40, No. 8/9, pp. 518–531, 22 October 2019, Published by Emerald Publishing Limited, DOI: 10.1108/LM-04-2019-0021, Available: https://www.emerald.com/lm/article-abstract/40/8-9/518/264273/Research-on-the-application-of-information?.

[2] Emmanouel Garoufallou and Panorea Gaitanou, “Big data: opportunities and challenges in libraries, a systematic literature review”, *College & Research Libraries*, Online ISSN: 2150-6701, Vol. 82, No. 3, p. 410, 1 May 2021, Published by Association of College and Research Libraries(ACRL), DOI: 10.5860/crl.82.3.410, Available: https://crl.acrl.org/index.php/crl/article/view/24918.

[3] Feras M. Awaysheh, Mamoun Alazab, Sahil Garg, Dusit Niyato and Christos Verikoukis, “Big data resource management & networks: Taxonomy, survey, and future directions”, *IEEE Communications Surveys & Tutorials*, Online ISSN: 1553-877X, Vol. 23, No. 4, pp. 2098–2130, 28 July 2021, Published by IEEE, DOI: 10.1109/COMST.2021.3094993, Available: https://ieeexplore.ieee.org/document/9482525.

[4] Abid Hussain, “Use of artificial intelligence in the library services: prospects and challenges”, *Library Hi Tech News*, Online ISSN: 2054-1678, Vol. 40, No. 2, pp. 15–17, 10 January 2023, Published by Emerald Publishing Limited, DOI:

10.1108/lhtn-11-2022-0125, Available: https://www.emerald.com/insight/content/doi/10.1108/lhtn-11-2022-0125/full/html.

[5] Susan N. Umeozor and Ahiaoma Ibegwam, "Assessment of the traditional and emerging roles of university libraries in Nigeria", *Journal of ICT Development Applications and Research*, Online ISSN: 2636-7440, vol. 4, no. 1/2, pp. 85–95, 30 December 2022, Published by Credence Publishing Ltd, DOI: 10.47524/jictdar.v4i1.86, Available: https://credence-publishing.com/?val=publication_details&manuscript=20231683557959696969540948.

[6] Faten Hamad, Maha Al-Fadel and Hussam Fakhouri, "The provision of smart service at academic libraries and associated challenges", *Journal of Librarianship and Information Science*, Online ISSN: 1741-6477, Vol. 55, No. 4, pp. 960–971, 28 July 2022, Published by SAGE Publications, DOI: 10.1177/09610006221114173, Available: https://journals.sagepub.com/doi/abs/10.1177/09610006221114173.

[7] Muhamad Khairulnizam Zaini, Wan Nor Haliza Wan Mokhtar and Irni Eliana Khairuddin, "Fostering Library Agility with Big Data", *Journal of Academic Library Management (AcLiM)*, Online ISSN: 2785-9185, Vol. 2, No. 1, pp. 33–45, 2022, 1 August 2022, Published by Perpustakaan Tun Abdul Razak, Universiti Teknologi MARA (UiTM), DOI: 10.24191/aclim.v2i1.19, Available: https://aclim.uitm.edu.my/article/article/view/19.

[8] Shan Liu and Xiao-Liang Shen, "Library management and innovation in the Big Data Era", *Library Hi Tech*, Online ISSN: 1758-0376, Vol. 36, No. 3, pp. 374–377, 4 January 2018, Published by Emerald Publishing Limited, DOI: 10.1108/LHT-09-2018-272, Available: https://www.emerald.com/insight/content/doi/10.1108/lht-09-2018-272/full/html.

[9] Muhammad Naeem, Tauseef Jamal, Jorge Diaz-Martinez, Shariq Aziz Butt, Nicolo Montesano *et al.*, "Trends and Future Perspective Challenges in Big Data", in *Proceedings of the Advances in Intelligent Data Analysis and Applications*, 26-28 April 2021, Porto, Portugal, Online ISBN: 13: 978-9811650352, DOI: 10.1007/978-3-031-01333-1_30, pp. 309–325, Published by Springer, Available: https://link.springer.com/chapter/10.1007/978-3-031-01333-1_30.

[10] Ebrahim A. A. Ghaleb, P. D. D. Dominic, Suliman Mohamed Fati, Amgad Muneer and Rao Faizan Ali, "The assessment of big data adoption readiness with a technology–organization–environment framework: a perspective towards healthcare employees", *Sustainability*, Online ISSN: 2071-1050, Vol. 13, No. 15, p. 8379, 27 July 2021, Published by MDPI (Multidisciplinary Digital Publishing Institute), DOI: 10.3390/su13158379, Available: https://www.mdpi.com/2071-1050/13/15/8379.

[11] Bin Hu, María-Manuela Moro-Cabero and Marta De-La-Mano, "Quality Management in Chinese Academic Libraries: A Systematic Review", *Sustainability*, Online ISSN: 2071-1050, Vol. 16, No. 7, p. 2700, 25 March 2024, Published by MDPI (Multidisciplinary Digital Publishing Institute), DOI: 10.3390/su16072700, Available: https://www.mdpi.com/2071-1050/16/7/2700.

[12] Marcia Lei Zeng and Philipp Mayr, "Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review", *International Journal on Digital Libraries*, Online ISSN: 1432-1300, Vol. 20, No. 3, pp. 209–230, 25 May 2018, Published by Springer Science+Business Media, DOI: 10.1007/s00799-018-0241-2, Available: https://link.springer.com/article/10.1007/s00799-018-0241-2.

[13] Osvaldo N. Oliveira and Maria Cristina F. Oliveira, "Materials Discovery With Machine Learning and Knowledge Discovery", *Frontiers in Chemistry*, Online ISSN: 2296-2646, Vol. 10, 7 July 2022, DOI: 10.3389/fchem.2022.930369, Available: https://www.frontiersin.org/journals/chemistry/articles/10.3389/fchem.2022.930369.

[14] Yong Chen, "Information integration in libraries", *Library High Technology*, Online ISSN : 0737-8831, Vol. 38, No. 1, pp. 210–219, 2 March 2020, Published by Emerald Publishing, DOI: 10.1108/LHT-11-2017-0232, Available: https://www.emerald.com/insight/content/doi/10.1108/lht-11-2017-0232/full/html.

[15] Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed *et al.*, "The role of big data analytics in Internet of Things", *Computer Networks*, Online ISSN: 1872-7069, Vol. 129, pp. 459–471, 24 December 2017, Published by Elsevier B.V., DOI: 10.1016/j.comnet.2017.06.013, Available: https://www.sciencedirect.com/science/article/abs/pii/S1389128617302591.

[16] Xin Li, Xiaoping Fan, Xilong Qu, Guang Sun, Chen Yang *et al.*, "Curriculum reform in big data education at applied technical colleges and universities in China", *IEEE Access*, Online ISSN: 3536-2169, Vol. 7, pp. 125511–125521, 3 September 2019, Published by IEEE (Institute of Electrical and Electronics Engineers), DOI: 10.1109/ACCESS.2019.2939196, 2019, Available: https://ieeexplore.ieee.org/document/8822937.

[17] Baeza-Yates, R. Ricardo and Berthier Ribeiro-Neto, *Modern Information Retrieval*, 1st ed. New York, USA: ACM Press, Harlow, UK: Addison-Wesley, 1999, Available: https://dlib.scu.ac.ir/handle/Hannan/325077.

[18] Jiping Xing, Wei Wu, Qixiu Cheng and Ronghui Liu, "Traffic state estimation of urban road networks by multi-source data fusion: Review and new insights", *Physica A: Statistical Mechanics and its Applications*, Online ISSN: 1873-2119, Vol. 595, p. 127079, 1 June 2022, Published by ScienceDirect, DOI: 10.1016/j.physa.2022.127079, Available: https://www.sciencedirect.com/science/article/pii/S037843712200125X.

[19] Xiao Wang, Yutong Wang, Jing Yang, Xiaofeng Jia, Lijun Li *et al.*, "The survey on multi-source data fusion in cyber-physical-social systems: Foundational infrastructure for industrial metaverses and industries 5.0", I*nformation Fusion*, Online ISSN: 1872-6305, Vol. 107, p. 102321, July 2024, Published by ScienceDirect, DOI: 10.1016/j.inffus.2024.102321, Available: https://www.sciencedirect.com/science/article/abs/pii/S156625352400099X.

[20] Weiquan Liu, Yu Zang, Zhangyue Xiong, Xuesheng Bian, Chenglu Wen *et al*., "3D building model generation from MLS point cloud and 3D mesh using multi-source data fusion", *International Journal of Applied Earth Observation and Geoinformation*, Online ISSN: 1872-826X, Vol. 116, p. 103171, February 2023, Published by ScienceDirect, DOI: 10.1016/j.jag.2022.103171, Available: https://www.sciencedirect.com/science/article/pii/S1569843222003594.

[21] Yaping Zhao, Jichang Zhao and Edmund Y. Lam, "House price prediction: A multi-source data fusion perspective", *Big Data Mining and Analytics*, Online ISSN: 2097-406X, Vol. 7, No. 3, pp. 603-620, September 2024, Published by TUP, DOI: 10.26599/BDMA.2024.9020019, Available: https://ieeexplore.ieee.org/abstract/document/10654670.

[22] Guangyao Chen, Shaofeng Wang, Yinsai Ran, Xiangpeng Cao and Zhuozhen Fang, "Intelligent monitoring and quantitative evaluation of fire risk in subway construction: Integration of multi-source data fusion, FTA, and deep learning", *Journal of Cleaner Production*, Online ISSN: 1879-1786, Vol. 478, p. 143832, 1 November 2024, DOI: 10.1016/j.jclepro.2024.143832, Available: https://www.sciencedirect.com/science/article/abs/pii/S0959652624032815.

[23] Hailin Feng, Qing Li, Wei Wang, Ali Kashif Bashir, Amit Kumar Singh *et al*., "Security of target recognition for UAV forestry remote sensing based on multi-source data fusion transformer framework", *Information Fusion*, Online ISSN: 1872-6305, Vol. 112, p. 102555, December 2024, Published by ScienceDirect, DOI: 10.1016/j.inffus.2024.102555, Available: https://www.sciencedirect.com/science/article/abs/pii/S1566253524003336.

[24] Bo Wang, Zengcong Li, Ziyu Xu, Zhiyong Sun and Kuo Tian, "Digital twin modeling for structural strength monitoring via transfer learning-based multi-source data fusion", *Mechanical Systems and Signal Processing*, Online ISSN: 1096-1216, Vol. 200, p. 110625, 1 October 2023, Published by ScienceDirect, DOI: 10.1016/j.ymssp.2023.110625, Available: https://www.sciencedirect.com/science/article/abs/pii/S0888327023005332.