

Novelty Assessment of Chinese Academic Articles in Information Resources Management: A Comparison of Knowledge Entity and Reference-Based Methods*

Yanqi Ren, Chen Yang, Yi Zhao, Heng Zhang and Chengzhi Zhang*

Nanjing University of Science and Technology, Nanjing 210094, China

Abstract

Novelty is a key factor in assessing research outcomes in a specific field. Measuring the novelty of articles in the field of Information Resources Management (IRM) helps researchers understand the current status of innovation and identify future development potential. Analyzing novelty across various themes within IRM offers insights for promoting innovation and ensuring balanced growth. Fine-grained knowledge entities encapsulate a paper's core knowledge, while references represent the flow of knowledge. Measuring article novelty from these two perspectives and comparing the results reveals thematic similarities and differences, providing a more comprehensive understanding. This study analyzes IRM-related research articles published in CSSCI-indexed journal from 2000 to 2022. After calculating article novelty using fine-grained research method entities and references, the BERTopic model identifies key themes in the field. The results indicate that novelty scores based on fine-grained knowledge entities are generally lower than those based on references, with both perspectives showing skewed distributions. Themes like University Libraries and Bibliometrics and Evaluation exhibit higher novelty scores from both perspectives.

Keywords

Novelty Assessment, Information Resources Management, Fine-Grained Knowledge Entities

1. Introduction

The term "Information Resources Management" (IRM) was first coined in the United States during the late 1970s and early 1980s [1], and it gradually spread worldwide. In China, IRM has evolved from a research field into an independent discipline, exerting a profound influence on the theoretical construction, discipline development, professional education, and career advancement of library and information science[2]. Notably, in September 2022, after the primary discipline "Library Information and Archives Management" was renamed "Information Resources Management," the future trajectory of development has garnered significant attention. The advent of the data and intelligence era has endowed the research related to the IRM discipline with new connotations. Whether in terms of data acquisition or computational requirements, both have been realized with the advent of the big data era, and new disciplines related to IRM are gradually emerging, providing fresh impetus for the development of the discipline. This undoubtedly presents new opportunities for the field of IRM. The exploration and analysis of topics within this field, along with their novelty characteristics, can not only assist fellow scholars in understanding cutting-edge trends and gaining insights into state of research in the Chinese Information Resources Management field from a macro perspective, thus advancing the discipline [3].

Methods for measuring the novelty of academic articles are typically divided into two categories: those based on references and those based on content. References in a paper represent its knowledge sources and can be regarded as the input of knowledge into the paper. In reference-based methods, the novelty of a paper is measured by quantifying the number of new combinations

Joint Workshop of the 2th Innovation Measurement for Scientific Communication (IMSC) in the Era of Big Data (IMSC2024), Dec 20th, 2024, Hong Kong, China and Online

* Corresponding author.

EMAIL: tusthisrenyanqi@njust.edu.cn (Y. Ren); yizhao93@njust.edu.cn (Y. Zhao); zhancz@njust.edu.cn (C. Zhang) 0009-0009-3569-5352 (Y. Ren); 0009-0007-6751-7062 (Y. Zhao); 0000-0001-9522-2914 (C. Zhang)



© 2024 Copyright 2024 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

of its knowledge sources. Previous studies have indicated that, for “knowledge input-based” articles, this method overlooks the complexity of citation motivations [4], focusing instead on the paper’s exploration and integration across interdisciplinary fields, which highlights the diversity and innovative combinations of knowledge sources [5]. Content-based methods may focus more on measuring “knowledge output-based” articles. Traditional methods of measuring the novelty of an article often rely on the frequency of keywords, entities, and other elements within the text. However, calculating merely the combinational frequency of vocabulary, without considering semantic differences between combinations, may overlook important novelty features [6]. Knowledge entities are fundamental developmental trajectories but also demonstrate the current units of a discipline. Assessing academic paper novelty through fine-grained knowledge entities can not only address the shortcomings of traditional novelty measurement methods but also capture finer granularity of novelty differences between entities. As a result, many scholars have used knowledge entities as entry points for assessing paper novelty. For instance, Liu et al. quantified the scientific novelty of doctoral theses based on biological entities to explore the heterogeneity and gender differences in scientific novelty [7].

Current research mostly measures novelty from a single perspective within single or multiple disciplines, with few studies measuring and comparing novelty under two perspectives across different research topics within a discipline. Measuring the novelty of IRM articles from the dual perspectives of fine-grained knowledge entities and references enables the study of novelty characteristic differences among various topics. This approach aids researchers in better understanding the knowledge structure and development trends of the field, thus promoting interdisciplinary integration and innovation. Therefore, this study aims to utilize fine-grained knowledge entities and references in articles to explore, from two perspectives, the novelty score characteristics and differences of topics in the Chinese IRM field from 2000 to 2022, identifying traces of innovation in academic articles as they explore unknown areas, thereby providing references for researchers in choosing paper topics and designing research proposals to advance the depth and breadth of academic research. By analyzing the novelty characteristics of academic topics from two perspectives, it is hoped that this study will contribute beneficially to the theory and practice of academic innovation, providing strategies and insights for the academic community to maintain and enhance novelty within the research landscape.

2. Related Work

This article aims to measure the novelty of articles related to the IRM field from the perspectives of fine-grained knowledge entities and references, and to analyze the characteristics, similarities and differences in novelty scores across different topics within the field. To this end, this section will review previous related work.

2.1 The methods for measuring novelty in academic articles

The measurement of novelty in academic articles primarily relies on two perspectives: references and the content of the articles themselves. References provide an external perspective for gauging novelty, with the measurement based on the flow of knowledge from cited articles to citing articles. By contrast, measuring novelty from an internal perspective, focusing on the content of the paper, often involves keywords, entities, and sentences, which directly capture the innovation of knowledge within the paper.

2.1.1 Novelty measurement based on references

Uzzi et al. evaluated the novelty of articles by the rarity of pairwise journal combinations in the references. The study found that scientific innovation does not simply rely on novelty or conventionality, articles that combine highly novel with highly conventional elements are more likely to become highly cited [8]. Lee et al. build upon Uzzi’s method of measuring novelty by

addressing the issue of journal commonality, using the tenth percentile instead of the minimum value to reduce noise. The research decomposes scientific creativity into two aspects: novelty and impact, and explores how team size, domain, and task diversity influence these aspects. The results indicate an inverted U-shaped relationship between team size and novelty, with team size enhancing novelty by increasing knowledge diversity [9]. Foster used a community detection algorithm to cluster journals in the references. Journals categorized within the same community were considered conventional. Those in different communities were considered novel for innovation assessment. He discovered that novel strategies are present in the field of chemistry. Furthermore, he found that their importance increases over time [10]. Subsequently, Wang et al. defined the novelty of a paper based on whether the journals cited in it are being combined for the first time. They found that articles with high novelty, as defined, contribute significantly to science and are more likely to become the top 1% most cited articles in the long term, stimulating future highly-useful research. However, such articles exhibit higher variance in citations and are less likely to become top-cited articles in the short term, indicating high research risk[11]. Veugelers applied Wang et al.'s method and concluded that scientific articles ranked in the top 1% of a field are more likely to have direct technological impact, be cited, and generate innovative patents compared to their non-novel counterparts [12].

2.1.2 Novelty measurement based on paper content

To measure the novelty of academic articles, some scholars have focused on the content itself, including keywords, entities, and sentences. An international expert panel, consisting of 57 leading experts from 16 countries, was mentioned by Zins in relation to the “informatics knowledge graph”, which explores systemic and comprehensive innovations in this field through group discussions [13]. In subsequent research, knowledge transfer and innovation models were categorized by Meng into three types: research-oriented innovation, front-led innovation, and disruptive innovation, based on the similarity of keywords in academic research topics [14]. Additionally, novelty was assessed through keywords by Mishra S et al. A single document’s thematic novelty was measured, based on a medical subject heading index, by proposing a series of methods that include improved word frequency statistics. It was found that the average conceptual novelty among most authors declines with age, yet the most innovative works may be published at any stage of their careers [15]. Regarding entity-based novelty measurement, scientific novelty in doctoral dissertations was quantified using biological entities by Liu et al., and it was found that the novelty declines over time, with gender differences also noted [7]. Moreover, the novelty was gauged by Liu using combinations of biological entities from COVID-19-related articles, highlighting the importance of international collaboration during pandemics [16]. The novelty of articles was measured by Chen et al. using fine-grained knowledge entity combinations, exploring the relationship between the composition of author teams and the novelty of academic articles [17]. Sentence-level novelty was also assessed by Zhang et al., quantifying the novelty of sentences by comparing the cosine similarity between current and historical sentences in the bag-of-words space, with the novelty score calculated as one minus the maximum similarity. The results were compared to those of English sentence novelty assessments, revealing that the performance of novelty detection at the Chinese sentence level can be comparable to that of English [18].

In summary, the novelty assessment based on references primarily includes evaluating the rarity of reference pair combinations and considering the commonality issues of journals. Additionally, the measurement of novelty based on paper content involves constructing evaluation metrics using factors such as the frequency and temporal aspects of keywords, entities, and sentences. Methods such as measuring novelty through entities and sentences are also utilized.

2.2 Research on topics related to novelty

In academic research, topic identification and novelty measurement are two important and closely related fields. Topic identification aims to determine hidden themes within a text, thereby helping researchers better understand the content. Novelty measurement, on the other hand, involves assessing the originality and innovativeness of an article or study. By comprehensively applying topic identification techniques and novelty measurement methods, the core themes of a paper can be thoroughly explored, and its innovative contribution within academia can be evaluated.

He et al. explored the predictive effect of paper innovation on literature growth in a certain field. They sorted word embeddings by time series to form time-ordered embeddings, calculated topic word similarity in vector space, and obtained a topic innovation index [19]. The frequency of topic changes over time was tracked by Mörchen et al., with frequency scores used to represent topic novelty. The results indicate that emerging trends can be predicted, and a trend-ranking function was provided to support interactive searches for the latest popular trends related to diseases [20]. Some studies utilize temporal relationships between topics to assess their novelty, but methods for constructing these relationships vary. Topic modeling was applied by He et al. to citation networks to determine pairwise relationships [21], whereas Yan employed similarity measures to establish these relationships [22]. These methods automatically detect research topics and assess their novelty based on textual information. Small et al. identified, classified, and analyzed the top 25 emerging topics from 2007 to 2010 each year, to understand the drivers of their novelty, including scientific discoveries, technological innovations, or external events. The novelty and value of these topics were evaluated by searching for these topics or significant awards recently received by key researchers. The findings indicate that this method provides a list of potentially important topics with novelty for review by decision-makers [23]. Additionally, Choi et al. used STM topic models to identify topics within patent data to mine potential novel topics [24]. Other scholars focused on the relationship between paper topics and time to explore the novelty of the articles. For instance, Tu et al. proposed a predictive index based on time, the number of topics, and frequency to identify the novelty of emerging topics in specific fields [25].

Current research in the IRM field mostly utilizes research topics for frontier analysis or for measuring novelty based on the topics themselves. However, studies that measure novelty from both the perspectives of knowledge entities and references, and combine them with paper topics to analyze characteristics and similarities and differences, are limited. This study starts from the content of articles, using fine-grained research method entities and references to compute the novelty of IRM field articles, and explores the distribution of novelty scores and their similarities and differences across different topics, with the aim to provide a reference for scholars in related fields to understand the IRM domain and select research topics.

3. Data and Methodology

This article aims to evaluate the novelty of IRM articles from 2000 to 2022. It uses fine-grained knowledge entities as the internal perspective and references as the external perspective. The article analyzes the novelty scores of different topics within the IRM field. It also examines the novelty differences across various topics under these perspectives. First, all academic articles related to the IRM field are collected. Then, the collected corpus is subjected to novelty calculation in two dimensions, and the BERTopic model is used to identify paper topics. Finally, the distribution of novelty across different topics in the IRM field and the similarities and differences under the two perspectives are analyzed. The specific research framework is shown in Figure 1.

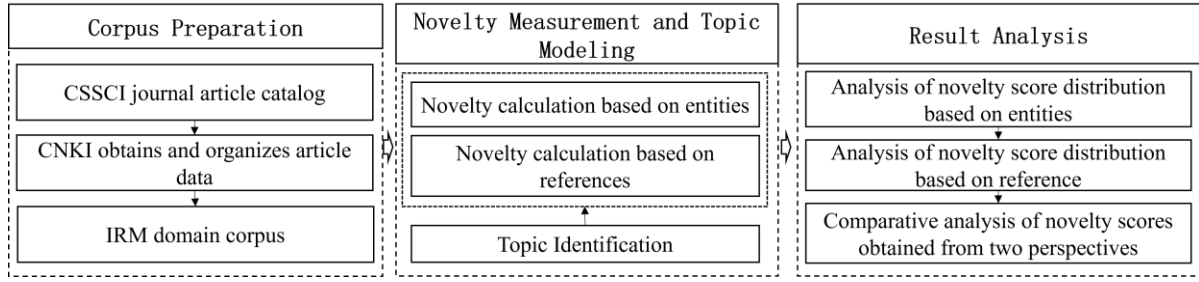


Figure 1: Research Framework Diagram

3.1 Corpus collection and organization

This study focuses on Chinese IRM articles, which are defined as those indexed by the Chinese Social Sciences Citation Index (CSSCI, <http://cssci.nju.edu.cn/>) and pertain to the field of information resource management. Currently, CSSCI is widely recognized in the Chinese academic and publishing communities and has become one of the most influential journal evaluation standards in social sciences in China [26]. Therefore, the CSSCI journal list (2021-2022) was used as the source, and the CNKI database (<https://www.cnki.net/>) was selected as the data source to obtain a total of 100,142 IRM articles. Due to varying initial publication years of journals, the time range was set from 2000 to 2022 for consistent analysis. Considering the needs of subsequent research, articles lacking publication dates, abstracts, full text (approximately 11.3% of the total data), references (approximately 20.6% of the total data), and those classified as cover articles were excluded, resulting in a valid dataset of 59,084 articles. The distribution of journals in the dataset is shown in Table 1. Notably, the journal *New Technology of Library and Information Service* was renamed *Data Analysis and Knowledge Discovery* in 2017. Thus, the paper counts from these journals have been combined under the latter title. The unequal distribution of source journals could affect topic identification. However, the development of the IRM field has led to significant trends of thematic intersection, with integration observed among its sub-disciplines, which mitigate potential biases in topic identification results due to unequal numbers. Additionally, the temporal distribution of journal articles was also analyzed, as shown in Figure 2.

Table 1
Statistical Results of The Number of Journal Articles

Journal name	Frequency	Journal name	Frequency
<i>Journal of Academic Libraries</i> (大学图书馆学报)	1164	<i>Data Analysis and Knowledge Discovery</i> (数据分析与知识发现)	3052
<i>Archives Science Bulletin</i> (档案学通讯)	1010	<i>Library Development</i> (图书馆建设)	3130
<i>Archives Science Study</i> (档案学研究)	1109	<i>Library Tribune</i> (图书馆论坛)	4201
<i>Journal of the National Library of China</i> (国家图书馆学刊)	869	<i>Researches on Library Science</i> (图书馆学研究)	1788
<i>Information Science</i> (情报科学)	5626	<i>Library Journal</i> (图书馆杂志)	2245
<i>Information studies: Theory& Application</i> (情报理论与实践)	4137	<i>Library and Information Service</i> (图书馆情报工作)	7517
<i>Journal of the China Society for Scientific and Technical Information</i> (情报学报)	1102	<i>Document Information & Knowledge</i> (图书情报知识)	1751
<i>Journal of Intelligence</i> (情报杂志)	7971	<i>Library and Information</i> (图书与情报)	1675
<i>Information and Documentation Services</i> (情报资料工作)	1659	<i>Journal of Modern Information</i> (现代情报)	7922
<i>Journal of Library Science in China</i> (中国图书馆学报)	716	<i>Journal of Information Resources Management</i> (信息资源管理学报)	439

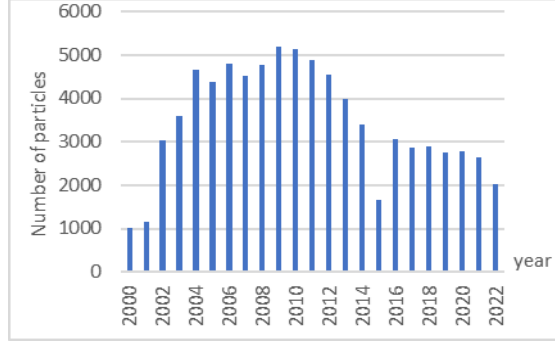


Figure 2: Age Distribution of Journal Articles

3.2 Novelty calculation of IRM articles

In this study, five research method entities were defined and extracted, and these fine-grained entities, along with the references of the articles, were used to measure the novelty of IRM articles from 2000 to 2022. The novelty results from the two perspectives were compared. This section primarily introduces the methods for measuring novelty from both perspectives.

3.2.1 Novelty calculation of entities based on fine-grained research methods

(1) Annotation of fine-grained research method entities in the IRM field corpus

Table 2

Five Types of Fine-Grained Knowledge Entities & Definitions in the IRM Field

Type	Definition	Samples
Theory	Theoretical frameworks, laws, regulations, or academic theories, etc.	文件运动理论(Record movement Theory)、文件生命周期理论(Theory of Records'life Cycle)、赖普斯定律(Price Law)
Method	Algorithms, models, and methods, etc.	LDA、SVM、CNN、主成分分析法(Method of Principal Component Analysis)
Data	Datasets, lexicons, dictionaries, literature, catalogs, etc.	NTU、WordNet、Hot Net、DBpedia、
Tool	Open-source tools, software, programming languages, or platforms, etc.	SPSS、stata、JAVA
Metrics	Metrics, evaluation criteria, etc.	召回率(Recall), F ₁ 、精确率(Precision)

For subsequent work on novelty measurement based on fine-grained research method entities, machine learning was employed to automatically identify research method entities in the full text. Based on Chu et al.'s classification standards [27] and related studies [28], theory, method, data, tool, and metric entities used to solve problems were extracted from research articles for novelty evaluation. Considering the broad range of disciplines in the information resource management field, where research methods integrate multiple disciplines and theory acts as a cornerstone for guiding practice and driving innovation, a theoretical entity was added to the original four fine-grained knowledge entities in this study. The specific definitions are shown in Table 2.

To minimize subjective bias in manual annotation, the process was divided into three parts: pre-annotation, consistency calculation, and formal annotation. Specific annotation guidelines are provided in the appendix. The pre-annotation phase was used to develop annotation rules, and during the consistency calculation phase, two sets of annotation consistency results were obtained and measured using Kappa coefficients [29], which were 0.69 and 0.73, meeting consistency requirements. Finally, all data underwent formal annotation. Ultimately, 2,716 sentences containing

method entities were obtained from the 249 sampled articles. The numbers of entities and sentences corresponding to the five types of method entities are shown in Table 3.

Table 3

Statistical Information of Method Entity Annotation Results

Type	# Entities	# Sentences
Theory	788	580
Method	1252	798
Data	819	525
Tool	529	296
Metrics	1006	518

(2) Fine-grained research method entity extraction in the IRM domain

Table 4

Statistical Information of Method Entity Extraction Results

Type	# Sentences
Theory	770143
Method	698713
Data	581840
Tool	719252
Metrics	547925

Chinese-BERT-WWM-Ext [30] is a Chinese pre-trained model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture. It has been trained using Chinese vocabulary and language characteristics, enhancing its ability to understand Chinese semantics. “WWM” stands for “Whole Word Masking,” meaning that entire words are masked during training as opposed to individual characters as in the original BERT model, which helps improve the model’s comprehension of complete words. “Ext” stands for “Extended,” indicating that this Chinese BERT model has been adjusted in terms of training size and steps to enhance model performance and effectiveness. Based on these characteristics, the Chinese-BERT-WWM-Ext model was fine-tuned on a training set, with optimal model parameters determined using a validation set. To ensure a balanced number of entities across all categories, all sentences are randomly shuffled and then divided into training, validation, and test sets in an 8:1:1 ratio. Its performance was tested on a test set using accuracy, recall, and F_1 score, achieving scores of 0.79, 0.75, and 0.77 respectively. Finally, the trained model was used to extract method entities from unannotated articles, with the extraction results for each entity type presented in Table 4.

(3) Novelty computation based on fine-grained entities

This study employs the method developed by Liu et al. [16], which measures the novelty of articles based on entity combinations and distance calculations, to calculate the novelty of IRM articles based on fine-grained knowledge entities. Liu et al., extracted entities from the titles and abstracts of COVID-19 articles. Then, these entities were paired, and the distances between each pair were captured. A distribution of distances between entity pairs was obtained, and those pairs whose distances fell within the top 10% of this distribution were considered novel entity combinations. Equation (1) shows the specific calculation method of distance for entities. The novelty score of each article was measured by the ratio of novel entity pairs to the total number of possible entity pairs in the paper, as shown in Equation (2). After obtaining all five types of fine-grained knowledge entities for each article, the novelty scores for all IRM domain articles were calculated using the aforementioned method and equations.

$$Distance_{E_a, E_b} = 1 - \frac{(E_a \cdot E_b)}{(\|E_a\| \cdot \|E_b\|)} \quad (1)$$

Where E_a and E_b represent the two entities in an entity pair, specifically, all five types of entities extracted from the aforementioned articles, $E_a \cdot E_b$ is the dot product of E_a and E_b , and $\|E_a\| \cdot \|E_b\|$ represents the product of the Euclidean norms of E_a and E_b .

$$Novelty_i = \frac{m}{C_n^2} \quad (2)$$

Where i represents the academic article, n is the number of entities extracted from i , C_n^2 is the total number of combinations of two entities that can be extracted from the set of n entities extracted from i (i.e. the number of entity pairs generated by n entities), and m denotes the count of entity pairs in i compared to all entity pairs generated in IRM domain articles, where the distance between the two entities in the pairs from i falls within the top 10% of the distance distribution of entity pairs.

3.2.2 Novelty computation based on references

This study employs a method for calculating novelty based on references, adopting the concept of Lee et al. [9], which uses the rarity of journal combinations in references as a measure of novelty, implemented by ranking and recording the percentiles of commonality. This method of measuring novelty was inspired by the research of Uzzi et al. [8]. Specifically, researchers first calculated the number of co-cited journal pairs in the database and recorded the cited journal pairs for each article. Then, for articles published in the same year, these journal pairs were aggregated into an annual set of journal pairs. Subsequently, for articles published in year t , the commonality of each cited journal pair was recorded. These commonality values were ranked, and the 10th percentile was recorded as an indicator of commonality at the paper level, as shown in Equation (3). In this manner, the novelty of articles can be objectively assessed without relying on other factors such as impact and citation counts.

$$Commonness_{ijt} = \frac{N_{ijt}}{\frac{N_{it}}{N_t} \cdot \frac{N_{jt}}{N_t}} \quad (3)$$

Where U_t represents the entire dataset, N_{ijt} indicates the number of journal pairs containing journals i and j in t , N_t represents the total number of journal pairs in U_t . $\frac{N_{it}}{N_t}$ is the probability of journal i appearing in U_t . $\frac{N_{it}}{N_t} \cdot \frac{N_{jt}}{N_t}$ is the joint probability of journal i and j appearing together.

3.3 Identify research topics

After evaluating the novelty of IRM field articles from both internal and external perspectives, BERTopic is employed for topic modeling to further explore the thematic characteristics of novelty from different perspectives, facilitating the subsequent analysis of thematic features of paper novelty scores.

3.3.1 Topic modeling based on BERTopic

This study primarily utilizes the BERTopic model for topic extraction from abstracts and titles of articles within the IRM field. The BERTopic model is based on BERT (Bidirectional Encoder Representations from Transformers) and topic modeling. It is applied to generate and interpret topics within documents. Compared to traditional topic models, BERTopic has the capability to handle multilingual text data and demonstrates superior performance. It more accurately captures the semantic and thematic information of text data, ensuring higher levels of topic coherence and diversity while retaining keywords within topics. Its dynamic topic modeling results can provide a clearer explanation of trend analysis [31]. Considering the characteristics of the data and research objectives, BERTopic is utilized for topic identification. The model employs BERT to generate document embeddings, uses UMAP for dimensionality reduction while preserving positional information, and clusters using the HDBSCAN algorithm. Finally, c-TF-IDF and maximal marginal relevance are used to optimize topic generation and obtain topic representation [32].

The detailed process of topic modeling in this study is as follows: Initially, since the study involves processing Chinese text, the “paraphrase-multilingual-mpnet-base-v2” was selected as the word embedding model. Subsequently, after the initialization of the UMAP model, “cosine” was used to measure the distance between points, informed by relevant literature [24] and multiple experimental results. To ensure tighter embeddings, the parameter “min_dist” was set to 0.01. Next, HDBSCAN was initialized. Considering the data volume and the handling of outliers in the study, the parameter “min_cluster_size” was set to 700 while “min_samples” was set to 1 to ensure the identification of topics while minimizing outliers as much as possible. Finally, the parameter “nr_topics” was set to “auto,” allowing BERTopic to iteratively generate topics. This approach avoids the inconvenience of subjectively setting too many or too few topics.

3.3.2 Topic identification results in the IRM domain

Topic modeling was conducted on all IRM field data for each paper, categorizing them into corresponding topics. The number of articles in each topic was counted, resulting in a total of 17 topics. Additionally, 12,908 articles were considered as noise due to their unclear topics, and they were labeled as -1. Combining the thematic names from previous studies with the characteristic words identified in this study, the research topics in the field of information resources management include but are not limited to University Library (topic0), Bibliometrics and Evaluation (topic1), Enterprise Knowledge Management and Organization (topic2), Online Public Sentiment (topic3), Resource Service Construction of Digital Libraries (topic4), National Security Intelligence Analysis (topic5), Electronic Data and Information Management in Government (topic6), Text Semantic Analysis (topic7), Information Literacy Education (topic8), Book Preservation and Classification (topic9), Copyright Protection (topic10), Reading Promotion (topic11), Patent Technology Protection (topic12), Virtual Consulting Services (topic13), Network Information Retrieval (topic14), Enterprise Competitive Intelligence (topic15), and User Information Behavior (topic16), aligning well with existing research findings.

4. Results Analysis

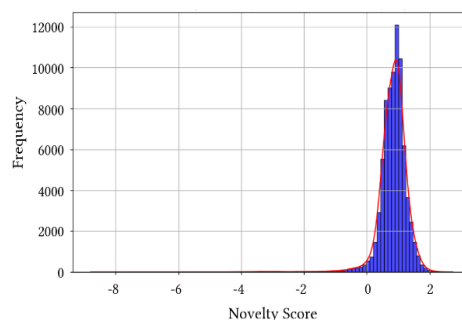
This section primarily investigates the distribution of novelty scores calculated based on references and fine-grained entities, comparing the results obtained by both methods to analyze the thematic characteristics of highly novel articles.

4.1 Distribution analysis of novelty calculation results of academic articles in the field of IRM

The distribution of novelty obtained by two methods will be explored separately in this section, in order to reveal the state of novelty in current academic research and its characteristics from different perspectives.

4.1.1 Novelty distribution analysis based on fine-grained research method entities

Based on the idea of entity combination, this study calculated the novelty of IRM articles from 2000 to 2022, with an overall novelty score range of 0-1. To more intuitively observe the general characteristics of the novelty score distribution derived from fine-grained entities, a histogram of



the score distribution was also created, as shown in Figure 3.

Figure 3: Histogram of Novelty Score Distribution Based on Fine-Grained Knowledge Entities

Overall, the novelty scores of the majority of articles are distributed in the lower range, showing a significant right-skewed distribution, indicating that novelty in academic research is still an issue requiring further attention and enhancement. Specifically, the most notable part of the graph shows that over half of the novelty scores are concentrated in the 0.0-0.2 range, with 55.8% and 34.8% of scores falling into the 0.0-0.1 and 0.1-0.2 intervals, respectively. This implies that most articles may primarily extend or slightly develop existing research, with relatively limited novelty. Such a distribution may raise concerns about the conservative or highly repetitive nature of academic research, suggesting that academic institutions and researchers need to more actively promote highly novel research. Articles with novelty scores exceeding 0.2 are significantly fewer. Articles falling into the 0.2-0.25 range account for 5.3%, those in the 0.25-0.3 range only make up 2.4%, and those scoring between 0.3-0.35 are as few as 1.1%. These articles may propose new theories or significant breakthroughs, yet they are relatively scarce. Although low in proportion, these articles with high novelty may have important implications for the development of the discipline.

In summary of the above analysis, the histogram indicates that although some articles exhibit high novelty, the majority demonstrate low novelty. To enhance overall research novelty in the IRM field, various measures may need to be adopted, such as strengthening interdisciplinary collaboration[33], encouraging high-risk, high-reward research projects[11], and increasing support and funding for original research. Additionally, academic review mechanisms could be oriented towards high-novelty research, encouraging more researchers to dare to challenge and explore new areas.

4.1.2 Novelty distribution analysis based on references

This study also utilized Lee's improved method for calculating novelty, which relies on the theory of the combinatorial rarity of reference journals to assess the novelty of IRM articles. This approach provides a basis for the reliability of subsequent novelty calculations. The distribution of novelty scores obtained from references is shown in Figure 4, where the horizontal axis represents the novelty scores in the range of -8 to 3. The vertical axis indicates the frequency of articles per score interval. The red line depicts the smoothed Kernel Density Estimation (KDE) curve [34]. It can be observed from the histogram and its KDE curve that the distribution of novelty scores exhibits a notable skewness, particularly showing a left-skewed distribution. Specifically, most research articles have novelty scores concentrated between 0 and 1, with this interval accounting for 63.25% of the total paper proportion, indicating that only a small number of articles in the academic dataset exhibit high novelty. Furthermore, it can be directly observed from the distribution graph that most scores are concentrated, with no significant dispersion seen in the scores obtained using references, a finding consistent with previous research conclusions and expectations [3].

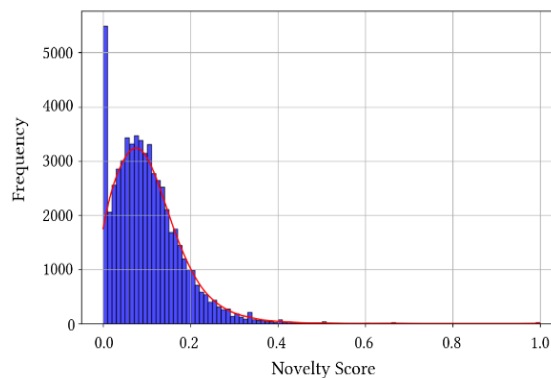


Figure 4: Histogram of Novelty Score Distribution Based on Reference

4.2 Comparative analysis of novelty calculation results

This section will provide a comparative analysis of the novelty results of academic articles obtained from two different perspectives. This comparison not only aids in revealing the application differences of citation-based and content entity-based novelty measurement methods across various research topics but also facilitates a deeper understanding of their characteristics and strengths in capturing academic innovation. Through comparative analysis, it is expected that the applicable scenarios for each method will be identified, along with their specific contributions to novelty assessment.

4.2.1 Comparative analysis of novelty calculation results based on fine-grained entities and references — global perspective

In this section, a global perspective is adopted to analyze the novelty characteristics based on fine-grained entities and references. Building on the existing classification system of research methods, this study introduces “theory” entities and uses five extracted fine-grained research method entities to calculate the novelty of IRM articles. Previous related studies have calculated novelty using references from a small dataset of academic articles. In contrast, this study has collected IRM academic articles from 2000 to 2022 and calculated their novelty using references, yielding a distribution of scores consistent with existing research. This study utilizes both internal and external texts of academic articles. After normalizing and aligning the novelty score data obtained from fine-grained entities and references, the results are plotted on a quadrant chart, as shown in Figure 5. The specific normalization method is shown in formula (4). The analysis combines the consistency and differences of novelty scores under the two methods to explore the novelty characteristics of IRM articles in depth.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (4)$$

Among them, X_{max} and X_{min} are the maximum and minimum values in the original data, respectively, X represents each feature value or observation value in the original data, X_{norm} represents normalized eigenvalues or observations.

Figure 5 illustrates the scatter distribution of two novelty calculation methods, Score_reference and Score_entity. The horizontal axis represents the Score_reference, while the vertical axis denotes the Score_entity, with each point representing an individual article. The obtained novelty score distribution is divided into four quadrants. The red and green lines indicate the demarcation lines for high and low novelty scores based on fine-grained entities and references, respectively, representing the average scores from the two perspectives. Overall, more score points are distributed in the lower quadrants of the chart. Relative to the demarcation lines, the novelty scores of articles based on references are generally higher than those based on fine-grained method entities. Articles located in the Q1 quadrant have novelty scores that exceed the average in both perspectives, indicating that these articles demonstrate significant innovation in content entities as well as in their reference citations compared to other articles. These articles may propose new theories, methods, or applications, combined with the citation of cutting-edge and diverse literature, to exhibit high overall novelty. Therefore, such articles are more likely to exert a positive impact on their respective fields and advance academic research.

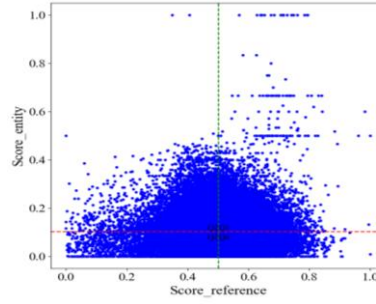


Figure 5: Quadrant diagram of two novelty scores

Articles in the Q3 quadrant score below average in both novelty measurements, indicating limited innovation in both content entities and reference citations. These articles predominantly follow established research paths, lacking novel insights or cutting-edge literature citations. If fine-grained research method entities are considered as knowledge output and reference citations are seen as knowledge input, then articles located in the Q2 and Q4 quadrants indicate that some articles exhibit outstanding innovation in knowledge output or demonstrate a certain level of novelty through the integration of pioneering, cross-disciplinary literature as input. Furthermore, by examining a few extremely high novelty articles ($\text{Score_entity} > 0.5$ and $\text{Score_reference} > 0.5$) in relation to internal and external features, it is found that they were mostly published between 2002 and 2013. This suggests that there is no direct correlation between novelty and publication time, and newly published articles do not necessarily exhibit high novelty.

4.2.2 Comparative analysis of novelty calculation results based on fine-grained entities and references - thematic perspective

This section analyzes the novelty calculation results based on fine-grained entities and references from a thematic perspective, further exploring the thematic characteristics of article novelty using both methods. This study selects the top 1000 articles with the highest novelty calculated based on fine-grained entities and the top 1000 based on references for thematic analysis. This allows for an understanding of the similarities and differences in how the two methods identify novel themes, providing a more comprehensive perspective on studying paper novelty.

- (1) Topic distribution characteristics of the top 1000 articles with novelty scores from two perspectives

The thematic overlap of the top 1000 articles in terms of novelty, based on fine-grained research method entities and references, was assessed using the Jaccard similarity coefficient, yielding a score of 0.5661. This indicates that the thematic overlap for the top 1000 articles in novelty scoring, under the fine-grained entities dimension and the references dimension, is approximately 56.61%. The thematic distributions of the two dimensions are similar in more than half of the cases, yet they are not completely identical. This suggests that the internal and external dimensions of an article may each emphasize different aspects of novelty assessment: entity innovation and reference innovation represent distinct contributions of content and citation networks, respectively.

The thematic distributions of the top 1000 articles in novelty, obtained from two dimensions, are tabulated in Tables 5, where topics labeled as -1 are considered noise. Specifically, the primary themes of high novelty articles based on fine-grained entities include University Library(topic0), Bibliometrics and Evaluation (topic1), Enterprise Knowledge Management and Organization (topic2), while those based on references include University Library(topic0), Bibliometrics and Evaluation (topic1), and Resource Service Construction of Digital Libraries(topic4).

Among these, the topic of University Library(topic0) has a high proportion in the top 1000 articles across both internal and external characteristics. On one hand, university libraries are a core area of library science with relatively rich research accumulation. On the other hand, aspects

such as information resource development, digital transformation, and information literacy education in university libraries easily intersect with other topics, resulting in high relevance in both entity and reference dimensions. Consequently, this leads to a high novelty score for this topic in both dimensions. Furthermore, articles under the theme of Bibliometrics and Evaluation (topic1) also exhibit a high proportion in the top 1000 articles by novelty score in both dimensions. As a significant research direction in the IRM field, bibliometrics and evaluation involve issues such as how to assess the quality, impact, and academic productivity of articles. Scholars in this research area are quite active, often proposing novel research methods and perspectives. Research in bibliometrics spans multiple disciplines and is particularly widely applied in various subfields of the IRM domain. Additionally, bibliometrics is inherently a field that necessitates continuous innovation. Researchers often draw from methods in other areas to develop new research techniques, resulting in novel findings that elevate the novelty score of this theme. Moreover, the integration of vast bibliographic and citation data, along with the trend of “cross-disciplinary integration” in the information resource management field, may contribute to the high proportion of this research theme among articles with high novelty scores.

Table 5
Proportion of Top 1000 Topics Based on Novelty Score

Topic No.	Topic	Percentage (entity)	Percentage (reference)
-	-1	25.10%	20.90%
topic0	University Library	27.60%	38.20%
topic1	Bibliometrics and Evaluation	6.20%	10.10%
topic2	Enterprise Knowledge Management and Organization	6.00%	2.70%
topic3	Online Public Sentiment	4.90%	1.20%
topic4	Resource Service Construction of Digital Libraries	3.60%	4.50%
topic5	National Security Intelligence Analysis	2.40%	2.10%
topic6	Electronic Data and Information Management in Government	4.30%	2.40%
topic7	Text Semantic Analysis	2.20%	1.20%
topic8	Information Literacy Education	2.10%	2.30%
topic9	Book Preservation and Classification	2.50%	5.00%
topic10	Copyright Protection	2.50%	1.50%
topic11	Reading Promotion	1.20%	3.10%
topic12	Patent Technology Protection	2.30%	1.20%
topic13	Virtual Consulting Services	2.30%	1.40%
topic14	Network Information Retrieval	2.00%	0.50%
topic15	Enterprise Competitive Intelligence	1.80%	1.50%
topic16	User Information Behavior	1.00%	0.20%

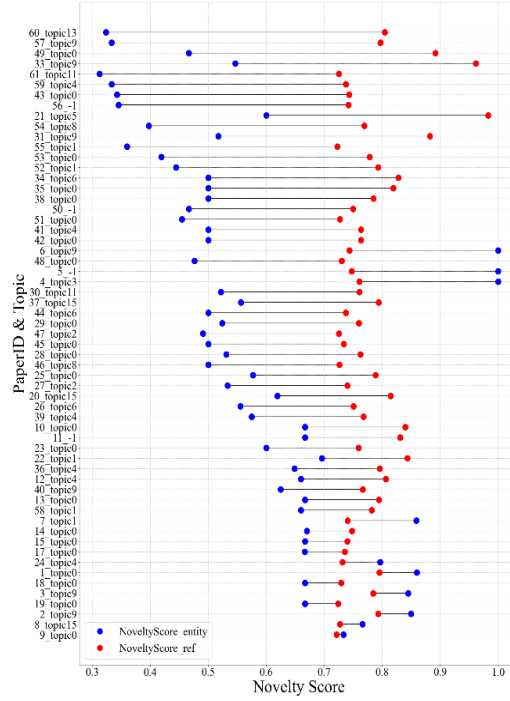
Note: “-1” represents articles with themes that are ambiguous and cannot be classified. This study focuses on the mainstream themes of high novelty articles, and those labeled as -1 further represent thematic diversity. The small number of unclassifiable articles does not affect the conclusions of this study.

(2) Characteristics of topic distribution in high novelty articles from two perspectives

This study considers the top 1000 articles, calculated under both dimensions for novelty, as high novelty articles in their respective dimensions. Extract the articles that appear in both the Top 1000 based on novelty calculated from fine-grained entities and the Top 1000 based on novelty calculated from references, totaling 61 articles. A PaperID was assigned to each paper along with its associated theme, and a Lollipop chart was drawn, as shown in Figure 6. Red dots represent novelty scores based on references, while blue dots represent novelty scores based on entities. The horizontal axis represents the novelty scores, and the line connecting the two dots indicates the score difference. The vertical axis represents 61 high-novelty articles assigned with PaperIDs and their corresponding topics. For instance, on the vertical axis, ‘60_topic 13’ represents the paper with ID 60 from the 61 jointly highly novel articles, which belongs to Virtual Consulting Services

(topic13).The chart shows that the novelty scores based on fine-grained research method entities are generally lower than those based on references among these 61 high novelty articles, consistent with the distribution shown in Figure 5. The main themes involved include University Library (topic0), Bibliometrics and Evaluation (topic1), Enterprise Knowledge Management and Organization (topic2), Online Public Sentiment (topic3),Resource Service Construction of Digital Libraries (topic4),National Security Intelligence Analysis (topic5), Electronic Data and Information Management in Government (topic6), Information Literacy Education (topic8), Book Preservation and Classification (topic9), Reading Promotion (topic11), Virtual Consulting Services (topic13), Network Information Retrieval (topic14), and Enterprise Competitive Intelligence (topic15). Notably, shared high novelty paper themes do not include Text Semantic Analysis (topic7), Copyright Protection (topic10), Patent Technology Protection (topic12), and User Information Behavior (topic16). First, themes like Copyright Protection (topic10) and Patent Technology Protection (topic12) have established research foundations with fixed citation networks and research methods, leading to weaker performance in external novelty calculations. Moreover, research on these topics is strictly limited by national laws and regulations, posing challenges for innovation within this framework. Secondly, themes like Text Semantic Analysis (topic7) and User Information Behavior (topic16) often involve interdisciplinary approaches, with complex research methods and application scenarios, resulting in varying novelty scores across dimensions, and thus less prominent performance in a composite dimension. Finally, in practical academic dissemination, the impact of these themes might not be promptly reflected in novelty calculations due to delays in dissemination and citation.

The difference in novelty scores for the theme of University Library (topic0) between fine-grained research method entities and reference-based scores is significant, indicating that the novelty of articles in this theme is more readily reflected through references, while it remains relatively conservative in terms of method entities. Several articles demonstrate relatively consistent novelty in the theme of Book Preservation and Classification (topic9). This is closely related to technological advancements in the digital age, including artificial intelligence and big data analytics. Research on book preservation and classification has benefited from the application of emerging technologies, continuously yielding new methods and tools. For instance, the emergence of large language models has spurred the intellectualization and automation of librarianship [35], and the use of electronic books in libraries has been extensively promoted. These technological advancements have to some extent increased the interdisciplinarity of this theme, facilitating new integrations and enhancements in both research methods and reference utilization.



Note: Red dots indicate reference-based novelty scores, blue dots indicate entity-based novelty scores, with the horizontal axis showing novelty scores and lines between dots showing score differences. The vertical axis represents 61 high-novelty articles with PaperIDs and their topics.

Figure 6: A Lollipop Chart of Novelty Scores and Topics for the Intersection of Top 1000 Articles by Novelty Score from Two Angles

In summary, this study analyzes the novelty scores and thematic characteristics calculated from two dimensions, from macro to micro perspectives, providing references for scholars assessing the novelty of articles across various themes. Meanwhile, the two perspectives emphasize different aspects when evaluating the novelty of academic articles: the fine-grained entity-based approach reveals the uniqueness and novelty of the content itself, whereas the reference-based approach highlights its position and role in academic dissemination and citation networks. Combining these two approaches allows for a more accurate assessment of the overall novelty of an article, thereby providing a more comprehensive and objective perspective for academic research and evaluation.

1. Conclusion and Future Works

This study employs the BERTopic model to identify research themes in Chinese IRM articles from 2000 to 2022, defining and extracting five types of fine-grained knowledge entities, and calculating novelty based on these entities. The novelty scores calculated from references are combined with those derived from fine-grained entities to analyze the differences and characteristics of paper novelty scores and themes under these two perspectives. The results indicate that articles with high novelty scores evaluated through fine-grained methods of entity and reference perspectives exhibit high novelty in topics such as university libraries, bibliometrics, and evaluation. The novelty score from fine-grained method entities shows a right skewed distribution, with a novelty range of 0 to 1. The novelty score based on references shows a left skewed distribution, with a novelty score range of -8 to 3. Both dimensions indicate a high degree of consistency in book preservation and classification themes. This study also normalized the novelty scores from both perspectives, analyzing them from a global and thematic perspective. From a global perspective, the novelty score of academic papers in the field of Chinese IRM based on references is generally higher than that based on fine-grained research methods and entities. From a thematic perspective, the overlap between the top 1000 topics in terms of novelty scores in the fine-grained entity dimension and the reference dimension is approximately 56.61%, indicating that the internal and

external dimensions of the paper may have their own focus in novelty assessment. Meanwhile, there is no direct correlation between the novelty of the paper and its publication time, and the novelty of newly published papers may not necessarily be high. By analyzing the novelty characteristics of IRM themes over the past 22 years under both perspectives, the study provides guidance for researchers in theme selection and reveals the importance of interdisciplinary integration in the digital age. Through this comparative analysis, it is expected that applicable scenarios for each method and their specific contributions to novelty evaluation can be identified, providing valuable insights for curriculum development and interdisciplinary collaboration. Additionally, in evaluating novelty across different themes, multiple factors should be considered to promote the development of the IRM field.

Future research can focus on enhancing entity extraction performance and optimizing paper novelty calculations by exploring more efficient algorithms. Integrating the BERTopic model with other models or technologies may improve theme recognition accuracy. Additionally, examining relationships between research method entities can uncover patterns in novel articles, offering a comprehensive understanding. Meanwhile, a deeper exploration of the proportion of different types of entities in papers with varying degrees of novelty may provide better assistance in understanding novelty. Finally, considering the document structure of fine-grained entities and references in novelty evaluation can lead to a more thorough assessment.

Acknowledgements

The paper is presented at the second Workshop on “Innovation Measurement for Scientific Communication (IMSC) in the Era of Big Data” at 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL). This work was partially supported by the National Natural Science Foundation of China (No. 72074113).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to correct grammatical errors, typos, and other writing mistakes. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication’s content.

References

- [1] Horton F W. Information resources management: Concept and cases, Cleveland : Association for Systems Management,1979.
- [1] M. A. F. C., Building consensus and promoting the first-level discipline construction of information resource management, *Journal of Information Resources Management* 13 (2023) 4-8.
- [2] Yang C., Wang Y., et al.Unveiling novelty evolution in the field of library and informationscience in China,*The Electronic Library* , 42(6)(2024)854-878.
- [3] Wang S, Ma Y, Mao J, et al., Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities, *Journal of the Association for Information Science and Technology*, 74 (2023) 150-167.
- [4] Fontana M, Iori M, Montobbio F, et al., New and atypical combinations: An assessment of novelty and interdisciplinarity, *Research Policy* 49 (2020) 104063.
- [5] Azoulay P, Graff Zivin J S, Manso G., Incentives and creativity: evidence from the academic life sciences, *The RAND Journal of Economics* 42 (2011) 527-554.
- [6] Liu M., Xie Z., Yang A J et al., The prominent and heterogeneous gender disparities in scientific novelty: Evidence from biomedical doctoral theses, *Information Processing & Management* 61 (2024) 103743.
- [7] Uzzi B, Mukherjee S, Stringer M, et al ., Atypical combinations and scientific impact, *Science* 342 (2013) 468-472.

- [8] Lee Y-N, Walsh J P, Wang J., Creativity in scientific teams: Unpacking novelty and impact, *Research Policy* 44 (2015) 684–697.
- [9] Foster J G, Rzhetsky A, Evans J A. Evans, Tradition and innovation in scientists' research strategies, *American Sociological Review* 80 (2015) 875–908.
- [10] Wang J, Veugelers R, Stephan P. Stephan, Bias against novelty in science: A cautionary tale for users of bibliometric indicators, *Research Policy* 46 (2017) 1416–1436.
- [11] Veugelers R, Wang J , Scientific novelty and technological impact, *Research Policy* 48 (2019) 1362–1372.
- [12] Zins C., Knowledge map of information science, *Journal of the American Society for Information Science and Technology* 58 (2007) 526–535.
- [13] Meng J, Chen X., Transnational Knowledge Transfer and Innovation Based on Academic Subjects: The Patterns and Characteristics of Knowledge Transfer and Innovation by Chinese Scholars Returned from the United States, 2017 International Conference on Innovations in Economic Management and Social Science, Zhejiang Hangzhou, 2017, pp. 47–53.
- [14] Mishra S, Torvik V I, Quantifying conceptual novelty in the biomedical literature, *D-Lib Magazine* 22 (2016) 9-10.
- [15] Liu M., Bu Y, Chen C, et al., Pandemics are catalysts of scientific novelty: Evidence from COVID-19, *Journal of the Association for Information Science and Technology* 73 (2022) 1065-1078.
- [16] Chen Z., Zhang C., et al.Exploring the Relationship Between Team Institutional Composition and Novelty in Academic Articles Based on Fine-Grained Knowledge Entities, *The Electronic Library*,42(6)(2024)905-930.
- [17] Zhang Y, Tsai F S., Tsai, Chinese novelty mining, in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009, pp. 1561-1570.
- [18] He J, Chen C., Predictive effects of novelty measured by temporal embeddings on the growth of scientific literature, *Frontiers in Research Metrics and Analytics* 3 (2018) 9-24.
- [19] Mörchen F, Dejori M, Fradkin D, et al., Anticipating annotations and emerging trends in biomedical literature, in: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, 2008, pp. 954-962.
- [20] He Q, Chen B, Pei J, et al., Detecting topic evolution in scientific literature: how can citations help?, in: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, 2009, pp. 957-966.
- [21] Yan E., Research dynamics: Measuring the continuity and popularity of research topics, *Journal of Informetrics* 8(1) (2014) 98-110.
- [22] Small H, Boyack K W, Klavans R., Identifying emerging topics in science and technology, *Research Policy* 43(8) (2014) 1450-1467.
- [23] Choi H, Woo J., Investigating emerging hydrogen technology topics and comparing national level technological focus: Patent analysis using a structural topic model, *Applied Energy* 313 (2022) 118898.
- [24] Tu Y-N, Seng J-L. Seng, Indices of novelty for emerging topic detection, *Information Processing & Management* 48(2) (2012) 303–325.
- [25] Wang B., A bibliometrical analysis of interpreting studies in China: Based on a database of articles published in the CSSCI/CORE journals in recent years, *Babel: International Journal of Translation* 61(1) (2015) 62-77.
- [26] Chu H, Ke Q., Research methods: What's in the name?, *Library & Information Science Research* 39(4) (2017) 284–294.
- [27] Wang Y., Zhang C., Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing, *Journal of Informetrics* 14(4) (2020) 101091.
- [28] McHugh M L. Interrater reliability: the kappa statistic, *Biochemia medica* 22(3) (2012) 276-282.

- [29] Zhou X, Huang H, Chi Z, et al. RS-BERT: Pre-training radical enhanced sense embedding for Chinese word sense disambiguation. *Information Processing & Management*, 61(4)(2024) 103740.
- [30] Contreras K, Verbel G, Sanchez J, et al., Using topic modelling for analyzing Panamanian parliamentary proceedings with neural and statistical methods, in: *2022 IEEE 40th Central America and Panama Convention*, Panama, IEEE, 2022, pp. 1-6.
- [31] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (2022) arXiv preprint arXiv:2203.05794.
- [32] Fontana, M., Iori, M., Montobbio, F., & Sinatra, R. New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy*, 49(7) (2020), 104063.
- [33] Terrell G R, Scott D W., Variable Kernel Density Estimation, *The Annals of Statistics* 20(3) (1992) 1236-1265.
- [34] Zhao R., Huang Y., Ma W. et al., Insights and reflections of the impact of ChatGPT on intelligent knowledge services in libraries, *Journal of Library and Information Science in Agriculture* 35(1) (2023) 29-38.