Advances in Methods for Identifying Research Frontiers in Disciplinary Fields: A Postprint

Authors: Zhang Xue, Zhang Zhiqiang, Cao Lingjing, Weinan Ruan, Ren Xiaoya, Feng Zhigang

Date: 2023-04-01T15:51:27+00:00

Abstract

[Purpose/Significance] This study reviews relevant achievements on research fronts both domestically and internationally, synthesizes existing problems in current research, and provides references for identifying research fronts in disciplinary fields. [Method/Process] First, it summarizes the necessity of research front identification. Second, it clarifies relevant concepts. Then, based on investigating domestic and international related research, it organizes and summarizes from two dimensions: research on research front identification methods and new directions in research front identification. Finally, it points out existing research deficiencies and proposes prospects for future development. [Results/Conclusion] Regarding conceptual definition, by clarifying series of concepts related to research fronts from both temporal dimension and definitional scope, it ultimately elucidates the connotation of research fronts. Regarding identification methods, classical research methods include direct citation, co-citation analysis, bibliographic coupling, and word cluster-based research front identification methods; meanwhile, research front identification based on multi-source data, multi-dimensional indicators, and machine learning algorithms represents new directions for future research. Based on the above analysis, it summarizes deficiencies of different types of research front identification methods and universal issues, and proposes prospects for future research priorities.

Full Text

Abstract

[**Purpose/Significance**] This paper systematically reviews domestic and international research on research fronts, summarizes existing problems, and provides references for identifying research fronts in subject fields. [**Method/Process**] First, we summarize the necessity of research front identification and clarify relevant concepts. Then, based on investigations of related research at home

and abroad, we organize the literature from two perspectives: research front identification methods and new directions in research front identification. Finally, we point out existing research deficiencies and propose future development prospects. [**Results/Conclusions**] Regarding conceptual definitions, we clarify the connotation of research fronts by analyzing related concepts from both temporal and definitional perspectives. Regarding identification methods, classical approaches include direct citation analysis, co-citation analysis, bibliographic coupling, and word cluster-based methods. Meanwhile, research front identification based on multi-source data, multi-dimensional indicators, and machine learning algorithms represents new directions for future research. Based on the above analysis, we summarize the shortcomings of different types of research front identification methods and their universal problems, and provide prospects for future research priorities.

Keywords: research front; expert interpretation; citation analysis; word cluster analysis; multi-source data; multi-dimensional indicators

2. Discrimination of Research Front Related Concepts

D. J. Price is considered the pioneer of research front studies. In 1965, he first proposed the concept of research front, defining the collection of recent, highly cited literature in citation networks as active research fronts, which he metaphorically described as "growing tips" or "epidermal layers" [6]. In 1973, H. Small [7] proposed that research fronts result from clustering highly cited papers and introduced co-citation clustering to identify research fronts. Together, these two scholars laid the theoretical and methodological foundations for the research front field. Subsequently, scholars have supplemented and refined the concept from various perspectives, with representative definitions as follows: In 1991, R. R. Braam et al. [8] viewed research fronts as a series of topics with high scholarly attention, using similarity between co-citation clusters to detect the stability of research fronts. In 1994, O. Persson [9] considered cited literature as the "research base" and defined clusters of citing papers that reference the same papers as "research fronts." In the same year, E. Garfield [10] referred to both core documents in co-citation clusters and their citing literature collectively as research fronts. In 1998, S. Bhattacharya et al. [11] employed co-word analysis, extracting terms directly from paper titles for co-word clustering, treating the resulting research topics as research fronts. In 2003, S. A. Morris et al. [12] combined research fronts with paradigm theory, viewing research fronts as paper sets that cite a persistent, relatively fixed group of documents within a specific paradigm, emphasizing the stability of cited literature. In 2006, Chen Chaomei [13] argued that research fronts might be discontinuous and temporary issues in a field, defining them as a set of emergent dynamic concepts and potential research questions. In 2008, N. Shibata [14] defined the clustering of recent direct citations as research fronts. In 2010, S. P. Upham et al. [15] considered research fronts as the most dynamic and attention-grabbing research topics in scientific fields. In 2014, Xu Xiaoyang et al. [16] defined research fronts as

recently emerged, emerging research topics or fields. In 2016, Zheng Yanning et al. [17] viewed research fronts as the most active parts of specific research fields during particular time periods. Table 1 summarizes these representative definitions of research front.

Regarding English terminology, "research front" and "research frontier" are the main expressions. Chinese scholar Zhong Zhen [18] elaborated on these terms from both theoretical and empirical perspectives: "research front" is an a priori evaluation representing expected results without practical verification, more commonly used in informetrics; whereas "research frontier" is a posteriori statistics representing analyzed results confirmed by peer experts, more applied in natural sciences. Some scholars argue that research fronts identified through scientific literature tend to be more like "research hotspots," while true research fronts should be a small number of cutting-edge fields such as the origin of the universe, biological evolution, and material structure, corresponding to the English term "research frontier."

Table 1. Representative Definitions of Research Front

Author	Definition
D. J. Price	Collection of recently highly cited literature
H. Small	Co-citation literature clustering
R. R. Braam et al. (1991)	Citing literature clustering
O. Persson (1994)	Bibliographic coupling clustering based on co-citation
E. Garfield (1994)	Collection of cited and citing literature
S. Bhattacharya (1998)	Co-word clustering
S. A. Morris et al. (2003)	Citing literature coupling clustering
Chen Chaomei (2006)	Emergent dynamic concepts and potential research questions
N. Shibata (2008)	Direct citation clustering
S. P. Upham et al. (2010)	Small highly cited clusters
Xu Xiaoyang et al. (2014)	Recently emerged, emerging research topics or fields
Zheng Yanning et al. (2016)	Latest and highly-attended research themes emerging in a field during a certain period

In the field of information science, many concepts are similar to research front, such as emerging research, research hotspots, scientific frontiers, and disciplinary frontiers. However, research front differs from these concepts in connotation, showing temporal differences with emerging research and research hotspots, and definitional scope differences with scientific frontiers and disciplinary frontiers.

(1) Temporal Dimension Differences. "Emerging research" refers to newly emerged research topics, characterized primarily by "newness" and temporal

novelty. Guo Hanning [19] interpreted emerging research as research that appears for the first time and develops vigorously, emphasizing the present moment, while research fronts are emerging research fields that have attracted widespread attention within a specific time period. Luo Rui et al. [20] argued that although emerging research shows trends of being "young" and "fast-growing," it does not necessarily represent research fronts with future value and prospects. In other words, research fronts should be valuable and stable emerging research. "Research hotspots" refer to topics with high attention, characterized primarily by "hotness" and extensive discussion. Zhong Zhen [18] pointed out that on the timeline, academically valuable research fronts from previous periods are likely to become research hotspots in new periods, indicating that research hotspots have a certain time lag relative to research fronts. In summary, we argue that research fronts refer to emerging research themes that are highly concerned by the scientific community, with innovative content and development potential, emphasizing thematic potential based on novelty and attention, characterized by high innovation and high impact. Figure 1 [Figure 1: see original paper] illustrates the relationship among these three concepts: under certain conditions, emerging research develops into research fronts, and some research fronts subsequently become research hotspots. If emerging research or research fronts fail to develop successfully, they will "disappear" and cannot develop into research fronts and research hotspots accordingly. Thus, research hotspots have obvious characteristics of temporal accumulation and sequence.

(2) Definitional Scope Differences. The difference between research fronts and scientific frontiers or disciplinary frontiers mainly lies in the distinction between research fields and science/disciplines. "Scientific frontier" refers to research that is forward-looking, pioneering, theoretical, and exploratory, with significant impact and leading role in future scientific development, also known as "science and technology frontier." Scientific frontier is a broad concept covering all disciplines and fields related to science and technology, while research fronts are usually limited to a specific research field. "Disciplinary frontier" refers to the most valuable development trends in a particular discipline, generally addressing major critical issues constraining current disciplinary development. Liu Haifeng [21] argued that the important difference between "discipline" and "research field" lies in permeability: discipline boundaries are usually impermeable with stable and integrated knowledge, while research field boundaries are permeable with relatively open and loose knowledge. Therefore, with increasingly refined disciplinary divisions, multiple disciplines usually correspond to one research field. Figure 2 [Figure 2: see original paper] shows the relationship among the three in terms of definitional scope: scientific frontier is the most macro-level research front, whose resolution may bring scientific research into new development stages, holding significant importance for national economy and social development. In the era of big science, solving many scientific problems is no longer limited to a single discipline. Therefore, research fronts refer to the frontiers of a certain research field, possibly involving multiple disciplines

or even being interdisciplinary.

3. Main Research Content of Research Front Identification

This paper focuses on how to identify research fronts in subject fields, specifically discussing methodologies and approaches for extracting themes with high novelty, development potential, and impact from subject domain topics to advance forward-looking and valuable project development. Based on investigations of domestic and international research, we summarize existing studies from two aspects: research front identification methods and new directions in research front identification. Identification methods represent the cornerstone and foundation for successful subsequent research, while new directions represent novel approaches for future research under the backdrop of rapid development in machine learning and large-scale text processing technologies.

3.1 Research on Research Front Identification Methods

Since D. J. Price introduced the concept of research front into the scientific field, scholars have explored its conceptual connotation and identification methods from various aspects, including qualitative interpretation and quantitative calculation, with quantitative methods primarily focusing on citation network analysis. This section summarizes relevant qualitative research and focuses on quantitative analysis. First, we outline the proposal, methodological processes, and classic studies of three major perspectives: direct citation, co-citation analysis, and bibliographic coupling. Second, we summarize research on performance comparison and method improvement. Since the above analyses are based on literature-level data with inherent limitations, word cluster-based analysis methods can serve as supplements or alternatives to citation analysis methods. Therefore, we finally summarize the progress of word cluster-based identification methods to provide a more comprehensive and systematic overview of the research front identification method system.

3.1.1 Expert Interpretation Based on Subjective Data The Delphi method is widely applicable for long-term trend prediction in specialized and comprehensive scientific and technological research, making it one of the most commonly used methods in research front identification. Similar foresight methods include brainstorming and expert consultation, which do not require repeated questionnaires and can directly solicit expert opinions, offering convenience and time savings. However, these methods may lead to identification results lacking democratization and socialization, with accuracy not maximized [22]. Additionally, numerous studies combine interviews with citation analysis results, noting that interviews are crucial sense-making tools [23-24]. For example, S. Upham et al. [15] interviewed 30 researchers for durations ranging from 30 minutes to 2 hours. Currently, in research front identification studies, expert participation is required both for setting thresholds for core literature collections in the early stage and for interpreting and revising quantitative analysis results

in the later stage. However, due to defects in information sources and analytical methods for selecting experts, these methods suffer from low accuracy and reliability, as well as poor objectivity. With data intensity becoming a notable feature across disciplines, how to make quantitative analysis results better support expert decision-making and how experts can play a role in research front identification are directions for further research.

3.1.2 Research Front Identification Based on Target Literature and Its Forward/Backward Citations With the rapid accumulation of scientific literature, scientometric methods have become important means for quantitatively identifying research fronts. Literature surveys reveal that citation analysis is one of the earliest, most theoretically grounded, and most widely used methods in research front identification. Citation analysis can be divided into direct citation, co-citation, and bibliographic coupling based on different types of citation relationships, with existing research primarily developing from these three perspectives. Table 2 provides a comparative analysis of the meanings of these three citation analysis methods and their analytical processes for research front identification.

Table 2. Comparative Analysis of Three Citation Analysis Methods and Their Processes for Research Front Identification

Method	Process
Direct Citation	Data download: Identify research field and time window, download target literature and their references from database platforms Citation pair identification: Identify citation relationship pairs in the dataset, remove literature without citations or not cited Similarity calculation: Calculate similarity based on citation frequency between documents Clustering: Cluster target literature based on similarity to form document clusters Naming: Assign thematic names to different document clusters

Method	Process
Co-citation Analysis	Data download: Identify research field and time window, download target literature and their citing literature from database platforms Data screening: Remove references with excessively high citation frequency to avoid over-aggregation in coupling Similarity calculation: Calculate similarity between two target documents based on frequency of citing the same references Clustering: Cluster target literature based on similarity to form bibliographic coupling clusters Naming: Assign thematic names to different document clusters
Bibliographic Coupling	Data download: Identify research field and time window, download target literature and their references from database platforms Highly cited target literature screening: Since citation frequency is time-dependent, group target literature and their citing literature by year. For each annual set, screen target literature according to specific rules, such as selecting literature with citation lag = 0 and citation frequency ≥ 3, or citation lag < 3 and citation frequency $\geq$ citation lag + 1, or citation frequency ≥ 5 as highly cited literature sets [25] Similarity calculation: Calculate similarity between two target documents based on frequency of being co-cited by citing literature Clustering: Cluster target literature based on similarity to form co-citation clusters. If network average degree centrality is too high, the network can be pruned by setting similarity thresholds between different target literature Naming: Assign thematic names to different document clusters

(1) Direct Citation. E. Garfield [26] in 1963 noted that direct citation analysis could serve as a key method for evaluating the impact of scientific discoveries. D. J. Price [1] used collections of recently published and frequently cited literature to identify research front themes. R. Klavans et al. [27] pointed out that direct citation can reveal the current status and future development trends of a field. However, direct citation requires a relatively long time window to obtain sufficient citations to ensure clustering effectiveness, thus limiting its widespread use.

(2) Co-citation Analysis. If papers E and F (regardless of their publication dates) are both cited by paper A, then papers E and F have a co-citation relationship. Co-citation strength refers to the number of papers that simultaneously cite papers E and F; the more papers, the higher the relevance between the two co-cited documents. Co-citation is forward-looking, and co-citation strength may change over time. ESI has consistently used co-citation analysis for field front prediction, specifically targeting clustered, frequently co-cited highly cited papers. H. Small [7] in 1973 proposed detecting research fronts based on document co-citation relationships, arguing that co-citation relationships more objectively represent the intellectual and social cognitive structure of science than direct citation. In 1974, H. Small and B. C. Griffith et al. [28-29] used this method for empirical analysis of document similarity and visualized clustering results. In 1985, he further revised the method, proposing fractional co-citation clustering to eliminate the impact of different reference counts per paper on clustering results [30]. I. V. Marshakova [31] also in 1973 noted that compared with bibliographic coupling (retrospective research), co-citation analysis (prospective research) is more complex and better reveals the evolutionary characteristics of research fronts. However, co-citation-based research front identification methods have limitations: research fronts can only be monitored when citing literature reaches a certain scale, creating temporal lag. Co-citation methods cannot immediately identify a research front when it emerges but only at later stages of field development. Co-citation clustering is an a priori method that clusters based on co-citation patterns among highly cited papers in a field, but if no papers in a field are highly cited, effective identification of that field's research fronts becomes impossible.

(3) Bibliographic Coupling. This term was first proposed by Fano, with M. M. Kessler [32] defining it in 1963. The basic premise of bibliographic coupling theory is that two documents citing one or more common documents must be related. If papers A and B both cite the same article(s), then papers A and B have a bibliographic coupling relationship. Coupling strength refers to the number of common cited documents; more common citations indicate higher relevance between the two coupled documents. W. Glänzel et al. [33] argued that clusters of highly coupled, recently published papers can identify early development trends in a field, offering advantages over co-citation analysis. They identified research fronts in fields such as Alzheimer's disease and fullerenes using bibliographic coupling, with expert consultation confirming the method's value. M. H. Huang et al. [34] used bibliographic coupling with a sliding window

to detect research fronts in organic light-emitting diodes from 2000-2009, finding results aligned with expert opinions. S. A. Morris et al. [12] used bibliographic coupling to identify and map research front evolution in anthrax data, with technology foresight expert panels deeming the results valuable. Since bibliographic coupling is retrospective and coupling strength is fixed, it is less dynamic than co-citation analysis. Additionally, two documents may cite the same literature for different content, potentially creating false high coupling strength.

After the three methods were proposed, scholars conducted deeper discussions on performance comparison, method fusion, and improvement:

(1) Comparative Analysis of the Three Methods' Performance. Some studies show direct citation yields more meaningful results. For example, N. Shibata et al. [35] compared direct citation, co-citation, and bibliographic coupling in identifying research fronts in gallium nitride, complex networks, and carbon nanotubes, finding direct citation produced more accurate research fronts and that bibliographic coupling better monitored research fronts than co-citation. Some studies show co-citation analysis performs better. For instance, J. Sharabchev [36] compared co-citation and bibliographic coupling in mapping immunology research fronts, finding co-citation performed better in drawing immunology science maps. Some studies show bibliographic coupling performs better, such as B. Järneving [37-38] who argued bibliographic coupling produces more micro-level and interpretable themes than co-citation. M. H. Huang et al. [39] used both methods to analyze research front evolution in organic light-emitting diodes, finding both effective but bibliographic coupling could identify more research fronts earlier and better. Some studies show no significant differences among the three methods. For example, K. W. Boyack et al. [25] used all three methods plus citation-text hybrid methods to identify biomedical research fronts, concluding each could be considered a valid method. Overall, due to different analysis units and research objects, different citation-based front identification methods may yield different results, but the three methods can complement each other to provide more comprehensive knowledge of field research fronts.

(2) Method Improvement and Cross-fusion. These discussions aim to improve readability and accuracy of identification results, primarily by adding relevant field information, dividing different time windows, and fusing multiple methods. Specific studies include: D. Zhao et al. [40] proposed author bibliographic coupling analysis (ABCA) in information science, comparing it with author co-citation analysis (ACA) [41], showing both have strengths and combined analysis could provide a fuller picture of field knowledge structure. C. Chen et al. [42] proposed a method combining author co-citation and reference co-citation analysis for more flexible and efficient naming of co-citation clusters. K. W. Boyack et al. [43-44] created highly detailed, dynamic global science maps based on co-citation analysis and attempted to improve result accuracy by considering different citation frequency thresholds, time slices, layout algorithms, and whether to include bibliographic coupling methods.

The above analyses reveal that co-citation analysis tends to cluster older published articles and cannot effectively include recently published, uncited articles, while bibliographic coupling tends to cluster recently published articles and cannot effectively include older, cited articles. Direct citation can more evenly cluster all documents across the entire time window. However, due to resource constraints, all three methods focus on literature reaching certain citation or coupling strength thresholds, preventing comprehensive analysis of all literature potentially related to research fronts. For example, M. H. Huang et al. [39] set the coupling strength threshold at 5, while ESI extracts the top 10% most citation-influential papers in each ESI discipline as research objects. This may cause loss of some research front-related literature and subsequent omission of some front topics. Second, due to different citation motivations and positions, literature in the same cluster may have low similarity, causing misidentification of fronts. Finally, none of the three methods can directly name clusters, with thematic cluster naming mostly based on manual interpretation of selected literature titles, keywords, and abstracts, involving significant subjectivity and requiring further expert validation.

3.1.3 Word Cluster-Based Research Front Identification Citation-based research front identification analyzes highly cited literature, making it difficult to include low-cited or zero-cited literature. To overcome this limitation, some scholars have shifted focus to finer-grained word clusters. When scholars pay attention to a research front, numerous related publications and keywords emerge in that field. Researchers have attempted to use word frequency statistics to discover more valuable research fronts more directly. Current identification methods mainly include burst word-based front identification and co-word-based front identification.

(1) Burst Word-Based Front Identification. With extensive application of burst word detection (BTD) in text mining, traditional scientometrics began using burst word monitoring technology to explore research fronts in related fields. Burst words refer to words whose frequency suddenly surges in text streams, characterized by frequency changes and burst time intervals [45]. J. Kleinberg [46] noted that the emergence of a research theme is accompanied by sharp increases in certain characteristic frequencies, indicating "activity bursts" marking the appearance of field research topics, and developed algorithms to identify such words. Chen Chaomei integrated Kleinberg's algorithm into CiteSpace, noting that burst words in literature collections can partially reveal potential fronts of research themes [13]. M. N. Li et al. [47] introduced burst word monitoring to enhance traditional co-word analysis, constructing an association rule mining model between keywords and burst words, with results confirmed as effective supplements to traditional research front identification. Unlike traditional high-frequency words, burst words emphasize sudden frequency increases. Combined with research front definitions, traditional high-frequency word analysis identifies research hotspots, while burst word analysis is more likely to identify research fronts. However, burst word monitoring's effectiveness depends on accu-

rately identifying burst words, with time slice and frequency threshold selection significantly impacting results. The same algorithm may achieve different accuracy rates across research fields, and despite ongoing efforts to develop different algorithms to improve monitoring, no universal algorithm has yet emerged.

(2) Co-word-Based Front Identification. Since papers require time to be cited by other papers, citation-based research front identification methods struggle to capture the latest trends. Co-word networks can be constructed immediately upon paper publication, enabling timely discovery of research fronts. As single keywords may weaken semantic expression of research topics, research themes are typically composed of co-occurring word clusters. M. Callon et al. [48] published the first academic monograph on co-word analysis in 1986, considered a milestone in the field. Co-word analysis first divides datasets into time-period subsets, using keywords from paper titles and abstracts as research objects, then calculates co-occurrence frequencies of different keyword pairs, and finally draws co-occurrence network maps for each period. Nodes in co-word networks correspond to keywords, and edges correspond to co-occurrence relationships. Co-word analysis typically uses common, frequently occurring keyword pairs to represent research themes, such as J. Joung et al. [49] using hierarchical clustering algorithms on keyword correlation matrices to identify research front technologies. However, these stable keyword pairs may interfere with discovering specific, suddenly attention-grabbing, period-specific research themes, which often better represent field research fronts. To overcome this, M. Katsurai [50] developed the TrendNets algorithm, calculating differences in keyword co-occurrence frequencies between two consecutive periods to quickly monitor changes in edge weights in dynamic co-word networks, thereby identifying themes widely discussed in one period but not previously. However, most current research still uses traditional co-word analysis, i.e., using keyword frequency statistics and high-frequency word clustering to identify research fronts, which tends to identify research hotspots rather than fronts. Moreover, co-word analysis also relies on keywords and ignores semantic information, prompting researchers to attempt using machine learning methods like topic models to improve semantic information in research front identification results.

Word cluster-based research front identification cannot directly define identified themes as research fronts; they require further expert interpretation or combination with other methods. Related research combines word cluster analysis with citation analysis: first identifying highly cited literature sets through citation analysis, then applying word cluster analysis to mine front topics. For example, R. R. Braam et al. [8] combined word frequency analysis with co-citation analysis to identify research fronts through word frequency analysis of highly cited literature titles and abstracts. Hou Haiyan et al. [51] combined co-word and co-citation analysis. Zhou Liying et al. [52] combined co-word analysis with bibliographic coupling. P. Van den Besselaar et al. [53] combined word frequency analysis with co-citation analysis. Combining different methods can effectively compensate for single-method limitations and has become a commonly used approach in current research front identification.

3.2 New Directions in Research Front Identification

Beyond these mainstream methods, scholars have explored various aspects of research front identification with the improvement of various databases and the rise of machine learning algorithms, attempting to enrich research objects, improve topic validity and readability, and make identification results more realistic and better serve decision-making.

3.2.1 Multi-Source Data as Research Objects Some scholars argue that citation and word cluster-based research front identification focuses on "how to measure" rather than "how to identify" [3]. Additionally, scholars have different views on the essential characteristics of research fronts: some believe research fronts should create new methods based on existing research [60-61], while others emphasize their disruptive nature to existing research [62-63]. Therefore, to comprehensively identify research fronts from multiple dimensions, scholars have attempted to design different scientometric indicators based on research front concepts to analyze performance of different front topics across dimensions, then classify front topics into different types to provide more targeted policy recommendations. For example, S. Cozzens et al. [64] summarized four main characteristics of research fronts: recent rapid growth, change in new things, high market or economic potential, and growing scientific nature. This pioneering work provided new ideas for research front identification from a multi-indicator perspective. Subsequent scholars supplemented and modified this indicator system from different dimensions. For instance, H. Guo et al. [65] proposed that sudden increases in specific word frequency, accelerated attraction of new authors, and enhanced interdisciplinary citation could serve as markers for research fronts. H. Small et al. [61] noted that scholars have reached consensus on the novelty and rapid growth characteristics of research fronts. D. Rotolo et al. [66] clearly defined five characteristics of research fronts: high novelty, relatively rapid growth, coherence, significant impact, and uncertainty/ambiguity. A. L. Porter's team [67-70] used novelty, persistence, author publication networks, and growth as four characteristics to define research fronts. Domestic scholars have also conducted systematic research, mostly building upon the above studies. Table 3 summarizes existing research front characterization dimensions and their measurement indicators.

Table 3. Multi-Dimensional Measurement Indicators for Research Front Identification

Dimension	Measurement Indicators
Novelty Measurement	- Publication year of references and their knowledge base- Average publication year of literature clusters (newer = more novel)
Growth Measurement	- Annual growth rate of publications in a theme- Growth rate of citations received- Growth rate of authors

Dimension	Measurement Indicators
Impact Measurement	- Citation frequency of publications in a theme- Knowledge redundancy between new and existing literature- Number of annual publishing authors
Public Recognition Measurement	- Social media attention, likes, shares- Mentions/citations in academic papers/reports- Influence of mentioning media/users and publishing journals/researchers
Interdisciplinarity Measurement	- Disciplinary richness, balance, and diversity- Network density, centrality, core-periphery degree
Other Secondary Indicators	- Life cycle characteristics- Nature Index (NI)- Productivity Index (π-index)

Multi-dimensional indicator-based research front measurement typically follows this process: first, identify research fronts using the methods mentioned above (citation, word cluster, or text clustering algorithms); second, construct measurement frameworks based on research front concepts to extract indicators that can express or comprehensively measure original data information; finally, calculate scores for each theme across indicators, classify themes into different front types based on indicator importance and threshold division. For example, Bai Rujiang et al. [54] combined topic strength and novelty to classify research themes into hot, emerging, declining, and potential fronts. Liu Ziqiang [71] selected themes ranking in the top 10% for both emergence and attention as research fronts. Fan Yunman et al. [72] used intersection points of publication volume, citation volume, and novelty curves to characterize topic development stages, identifying research fronts by comparing LDA topic model results with hybrid baselines.

Existing studies mostly cover one or several dimensions, typically using 3-5 indicators. Composite indicators are complex and difficult to generalize, and no unified, systematic quantitative system has been formed. Different indicator selections for the same dimension make horizontal comparison of front identification results across fields nearly impossible. Moreover, since thresholds for each dimension require manual specification and their accuracy needs verification, classification of different front topic types involves considerable subjectivity.

3.2.3 Machine Learning Algorithm-Based Research Front Identification Methods The accuracy and decision-making efficiency of machine learning have improved with large-scale text processing technologies, with its usage growing exponentially in recent years. Machine learning applications in research front identification mainly include text clustering, text classification, and time series analysis. Text clustering uses unsupervised learning algorithms for topic identification, reducing keyword dimensions and improving semantic connotation and interpretability compared with traditional co-word analysis. Text classification uses supervised learning algorithms to predict highly cited literature

sets in advance, effectively solving citation lag problems. Time series analysis uses measurement indicator trends over time to predict future development trends through time series models.

Specific research includes: Text clustering-based front topic identification. With increasing data volume and unstructured data, text topic mining technology has become increasingly important. Text topic mining extracts valuable information and knowledge from structured, semi-structured, or unstructured text data, including text collection, data cleaning, feature extraction, feature pruning, and text clustering. To consider contextual relationships between words, scholars have developed different algorithms for feature extraction and text clustering. For example, J. Yoon et al. [73] proposed SAO structure analysis, where sentences organized by grammatical structure clearly describe relationships between components [74], making identified topics more understandable [75-76]. W. M. Pottenger et al. [77] used neural network models to identify emerging concepts or topics in datasets. A. Kontostathis et al. [78] proposed the Emerging Trend Detection method, first showing topics in different periods through co-word analysis, then using text mining for topic extraction and classification based on co-occurrence features, and finally validating topics and judging development trends through evaluation criteria. To enrich topic semantic structures, D. M. Blei et al. [79] proposed using LDA topic models to mine implicit semantic structures in datasets unsupervised. T. Mikolov et al. [80] proposed the Word2vec word embedding model to learn implicit vector representations of words. S. Xu et al. [81] used topic n-grams models to extract term-based topics. These various algorithms enrich semantic information in research front topic identification results, facilitating expert understanding and interpretation.

Text classification-based front dataset prediction. As previously noted, most research is based on highly cited papers, but citation accumulation requires time, making high-citation paper collections unable to effectively cover recently published but potentially highly cited literature, thus compromising the novelty of identified research fronts. Therefore, using machine learning classification models to early predict valuable and potentially strong literature provides new research methods for current and future research front identification. For example, C. Lee et al. [82] built 18 high-value patent discrimination indicators based on patent data, using feedforward multilayer neural network models to capture relationships between input and output indicators, thereby predicting patents with research front characteristics in early application stages. Li Xin et al. [83] first built machine learning models to identify potentially highly cited papers, then used these as data sources for cluster analysis to identify research front themes.

Time series analysis-based indicator evolution prediction. Current research in this area is preliminary. For example, S. Xu et al. [84] defined research front indicators as novelty and consistency, calculated evolution trends of various theme indicators from 2001-2017, and used time series analysis models and

multi-task least squares support vector machines to predict two-year indicator changes, thereby forecasting potential research front themes. Similarly, Li Jing et al. [85] and Yue Lixin et al. [86] used support vector machines and ARIMA models respectively for topic trend prediction.

From word frequency statistics to text clustering algorithms, from subjectively thresholded highly cited literature sets to early prediction of potential highly cited literature sets, and from indicator calculation in specific periods to future trend prediction, new research front identification methods enrich semantic information, provide more flexible identification granularity, overcome citation analysis time lag, and produce more forward-looking topics, offering new research ideas and measurement methods for research front identification.

4. Summary and Outlook

This paper begins with the background and concept discrimination of research fronts, summarizes relevant content from qualitative and quantitative perspectives on research front identification methods and key technologies, and proposes deficiencies and future development directions based on existing measurement indicators and research objects.

4.1 Further Clarification of Research Front Concepts Needed

Current research front concepts remain unclear, mostly defining and discovering research fronts from literature perspectives. Whether these identified "research fronts" are truly research fronts requires further discussion, with some scholars arguing that literature-based measurement identifies more like "research hotspots." However, research front identification is undoubtedly important for predicting scientific development trends and scholars' research direction selection. This paper discriminates related concepts, attempting to clarify relationships among research fronts, research hotspots, emerging research, scientific frontiers, and disciplinary frontiers to provide references for scholars.

4.2 Threshold Setting Rationality and Effectiveness Need Further Verification

Research front identification typically first selects representative highly cited literature by setting citation frequency thresholds, making it difficult to include low-cited or zero-cited literature. For example, ESI research front reports use papers ranking in the top 1% by publication year and discipline as highly cited core literature sets. Upham et al. [15] used top 1% papers by citation frequency, adjustable upward or downward. Glänzel [33] defined core literature as documents with more than 9 links and link strength of at least 0.25. Huang et al. [34] set the coupling strength threshold at 5. Due to time lag, reaching high citation rates requires years, with variations across disciplines. Different thresholds also produce different front topics. How to determine threshold size lacks clear scientific basis, generally set subjectively by researchers based on data volume and

desired front topic count, requiring further exploration of scientific rationality.

4.3 Method Applicability Scope Needs Further Definition

If we view research fronts as future-oriented exploration, they should be dynamic, pluralistic, and multi-dimensional concepts. However, current data sources are mostly papers and citations, with researchers typically setting citation windows and dividing time ranges into intervals. Without accurate, recognized citation windows, researchers usually choose windows based on research purposes, with 5-year fixed windows most common. On one hand, citation counts are affected by publication time, author motivation, and article availability, making lag issues difficult to overcome, leaving whether identified topics are truly research fronts open to question. On the other hand, whether fixed time windows can effectively capture all research fronts given burst words and "sleeping beauty" literature requires consideration. Future research could integrate multi-source data to capture front characteristics from different data sources and consider sliding windows for topic evolution trends.

4.4 Multi-Source Fusion Data and Multi-Dimensional Indicators Need Further Systematization

Current research primarily uses paper data, with some studies adding funding projects, technology plans, and patents as supplementary sources. Studies show paper research themes lag funding project themes by an average of 2 years, but existing research simply combines different data sources, separately mining themes and ignoring differences caused by data source characteristics. How to incorporate data source differences and features, effectively fusing different data sources for the same field based on integrated data matrices, is a breakthrough for addressing lag issues. Regarding multi-dimensional indicators, few comprehensively consider the essential multi-dimensional characteristics of research front concepts. Most indicators are complex and difficult to generalize, with no unified quantitative system formed. We should further clarify research front essence, combine different features, simplify calculation methods while ensuring indicators reflect all elements, and maximize objectivity.

4.5 Targeted Value of Existing Research Front Identification Results Needs Improvement

First, research fronts' most common characteristic is their potential to change and inject new understanding into our cognition of problems. To some extent, research fronts are not entirely quantifiable, meaning scientometric methods inevitably have limitations. Second, current methods involve expert participation and scoring in both early and late stages, but scientific communities already have their own understanding of research fronts. Therefore, the true value of scientometrically identified research fronts for researchers needs clarification. We should clarify service objects and purposes: scientometric methods mainly

provide relatively objective trend information at the macro level, requiring domain experts to further interpret implicit information behind quantitative data. Regarding identified research front topic scope, we should focus on identifying potentially valuable but overlooked research fronts, hoping our research can attract more attention to certain topics and provide references for research selection and policy formulation.

References: [The references section would follow here, preserved exactly as in the original]

Note: Figure translations are in progress. See original paper for figures.