# A Postprint of an Analysis Method for Knowledge Evolution of ESI Research Fronts Based on Knowledge Element Mutation

**Authors:** Sun Zhen, Leng Fuhai

**Date:** 2023-04-01T15:51:22+00:00

## Abstract

[Purpose/Significance] As an exploratory study addressing the scientific and technological intelligence needs of disciplinary domains, focusing on full-text critical semantic quantitative analysis, and aiming to realize the practical application from intelligence automation to knowledge automation, this paper, based on technologies such as semantic annotation and machine learning and building upon previous research that detected the evolution mechanisms of research fronts from the perspective of knowledge element co-occurrence, further proposes a research front knowledge evolution analysis method based on knowledge element mutation. [Method/Process] Using the Word2vec word embedding model to represent knowledge elements as word vectors, identifying clusters of knowledge elements with similar semantic and pragmatic associations through calculating the Euclidean distance of knowledge element vectors and employing the K-means clustering method, calculating the TF-IDF values of each knowledge element within diachronic clusters, quantitatively measuring the abrupt changes in the importance degree of mutated knowledge elements, and thereby excavating the characteristics and patterns of knowledge element mutation in the evolution of ESI research fronts. [Result/Conclusion] Through comparative verification of detection results, it is found that the scientometric method based on knowledge element mutation not only complements and expands upon previous research methods, making the excavation of knowledge movement patterns within research fronts more concrete and detailed, but also serves as powerful evidence for detecting, as early and timely as possible, the future development trends and key intelligence signals of research fronts within the temporal sequence framework.

## Full Text

# An ESI Research Front Knowledge Evolution Analysis Method Based on Knowledge Element Variation

**Sun Zhen**[1], **Leng Fuhai**[2]
[1] Institute of Information Management Research, Shandong University of Technology, Zibo 255000
[2] Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190

### Abstract

[**Purpose/Significance**] As an exploratory study oriented toward disciplinary domain scientific and technological information needs, focusing on full-text key semantic quantitative analysis, and aiming to realize the practical application shift from information automation to knowledge automation, this paper proposes a research front knowledge evolution analysis method based on knowledge element variation, building upon previous research that examined the evolution mechanism of research fronts from the perspective of knowledge element co-occurrence. [**Method/Process**] The Word2vec word embedding model is employed to represent knowledge elements as word vectors. By calculating the Euclidean distance between knowledge element vectors, K-means clustering is used to identify clusters of knowledge elements with similar semantic and pragmatic associations. The TF-IDF values of each knowledge element within diachronic clusters are calculated to quantitatively measure sudden changes in the importance of knowledge elements after variation, thereby mining the characteristics and patterns of knowledge element variation in ESI research front evolution. [**Result/Conclusion**] Comparative testing of detection results reveals that the scientometric method based on knowledge element variation not only supplements and expands previous research methods, making the excavation of internal knowledge movement laws within research fronts more concrete and detailed, but also serves as powerful evidence for detecting future development trends and key intelligence signals of research fronts as early and timely as possible within the time series framework.

**Keywords:** knowledge element; research fronts; machine learning; full-text semantic analysis; perovskite solar cells
**Classification Number:** G250
**DOI:** 10.13266/j.issn.0252-3116.2022.02.015

## 1. Introduction

With the massive accumulation and growth of scientific data resources, scientific literature—as a knowledge carrier documenting systematic processes of scientific discovery and technological innovation—constitutes more efficient and high-quality textual big data than raw scientific data, better facilitating sci-

entific knowledge discovery and innovation dissemination. However, the unstructured nature, weak semantic representation, and difficulty of knowledge computation in textual big data constrain the deep-level and high-efficiency utilization of scientific literature [1]. Therefore, how to semantically represent key knowledge content in scientific literature, develop fine-grained knowledge unit association distribution patterns, discover domain knowledge flow laws, and realize potential scientific discoveries has become a critical issue.

The ESI (Essential Science Indicators) Research Fronts database represents a dataset that captures the distribution patterns of world science and technology frontier knowledge within a certain time period [5]. Since 2001, the Institute for Scientific Information (ISI) in the United States has launched the Essential Science Indicators (ESI) database and employed co-citation analysis methods for research front analysis. "Research Fronts," as a professional domain methodology, originates from certain commonalities among scientific research, which may derive from experimental data, scientific hypotheses, research methods, or computational approaches, and has become an important task for current scientific and technological intelligence services oriented toward disciplinary domains [2-3].

Research fronts document the continuous processes of emergence, convergence, and development in dispersed research fields. Through analysis of citing literature, the latest development directions of a field can be discovered. While keywords and subject terms can reflect research topics and knowledge characteristics to some extent, for STEM (Science, Technology, Engineering & Mathematics) literature, domain knowledge is often sealed within documents in the form of associated knowledge units, particularly in the "Methods/Experimental Section" of papers [4]. If the scope of literature data representing knowledge distribution patterns in a scientific field within a certain period can be defined, and specific knowledge units in such literature can be targeted with designed computational methods for mining key semantics, then domain knowledge maps and other visualization methods can be employed to reveal the implicit knowledge structure of that field.

The "Recommendations of the CPC Central Committee on Formulating the 14th Five-Year Plan for National Economic and Social Development and the Long-Range Objectives Through the Year 2035" emphasizes: "Facing the frontiers of world science and technology… strengthening basic research, emphasizing original innovation, and optimizing discipline layout and research deployment." In this context, constructing semantic processing methods for disciplinary domain knowledge structure characteristics based on ESI research front data for basic research, and conducting forward-looking strategic intelligence analysis for frontiers, can provide researchers with early insights into R&D trends, frontier development trajectories, and opportunities to seize commanding heights, while also providing decision-making support and intellectual backing for national and local governments to map out disciplinary strategic layout priorities and deploy main directions for scientific and technological innovation.

However, the current status of intelligence research on ESI "Research Fronts" primarily involves its use as foundational data for "research front" series detection [6] and exploring national performance through scientific mapping [7], with few attempts to utilize full-text scientific literature data to explore internal knowledge structure changes in research fronts from the perspectives of semantic analysis and knowledge computation.

Given this background, and since our previous research has already defined the conceptual connotation of knowledge elements in disciplinary domains [8], preliminarily constructed the theoretical foundation of knowledge element measurement methods [9], and empirically demonstrated the feasibility and advancement of knowledge element methods for ESI research front knowledge evolution [4], this paper further proposes a scientometric method based on knowledge element variation. This approach focuses the evolution analysis of ESI research fronts on the level of clustered knowledge elements, emphasizing changes in the semantic and pragmatic functions of knowledge elements during research front evolution. This is indeed key to further understanding the evolution mechanism of research fronts from the perspective of domain knowledge content itself.

Previous research [4] has demonstrated that from an operational concept of intelligence practice, research fronts can be viewed as collections of knowledge elements with semantic and pragmatic functions. The foundation of research front evolution lies in the innovative recombination and application changes related to knowledge elements. While the interpretation of knowledge element co-occurrence linkages is based on the analysis of domain knowledge entities extracted through text mining, it essentially still relies on manual semantic and pragmatic annotation of knowledge element measurement results, with machines merely completing the automatic collection and classification of domain knowledge during this process.

Against this backdrop, this paper continues to use the "high-efficiency perovskite solar cells" hotspot front in the chemistry and materials science domain from the *2016 Research Fronts* report as an example. It attempts to utilize machine learning technology to automatically identify changes in the semantic and pragmatic functions of knowledge elements during research front evolution, detecting sudden variation phenomena in knowledge element clusters within the same pragmatic environment during research front evolution. The goal is to understand the change mechanisms of disciplinary knowledge structure during research front evolution from different perspectives and in greater depth, while also providing references for intelligent scientific and technological intelligence work oriented toward disciplinary domains.

## 2. Related Research

Scientific literature knowledge features can be divided into knowledge representation features and knowledge entity features. Knowledge representation features are primarily used to evaluate the academic influence of literature, while knowl-

edge entity features can be further divided into external and internal knowledge entity features. External knowledge entities mainly refer to surface-level keywords and subject terms, commonly used in research on knowledge utilization and transfer aimed at promoting knowledge discovery. However, measurement analysis based on external knowledge entity features has numerous limitations in constructing disciplinary domain knowledge maps from scientific literature, capturing disciplinary domain ideas, and discovering potential content associations [10].

In recent years, many scholars have focused on scientometric analysis based on internal knowledge entity features. Y. Ding et al. first proposed the concept of "entitymetrics" in 2013, extracting gene, disease, and other biological knowledge entities from each document to construct entity citation networks and analyze knowledge utilization and transfer patterns related to Metformin [10]. Subsequently, M. Song et al. used PubMed seed literature and their references to establish a "gene-citation-gene" network to detect implicit interactions between genes [11]. Q. Yu et al. extracted biological database knowledge entities from PubMed Central full texts and their references to construct database linkage networks, tracking usage link relationships of biological databases [12]. D. Lee et al. used "Alzheimer's disease" as a search term to construct four types of networks, capturing cognitive state maps of the field from different perspectives of indexers, authors, or citers [13]. K. Lee et al. used tomographic content analysis with proteins, genes, and MeSH terms as entities to explore the impact of queried knowledge entities on heterogeneous knowledge entity networks within literature [14]. K. Li et al. extracted R package entities from each document based on dictionaries to construct paper-level co-mention networks, investigating the role and usage status of R language packages in biomedical literature [15].

Entitymetrics takes knowledge entities as the basic operational unit, truly achieving the deepening of measurement objects toward semantic knowledge ontologies within literature, and can be better applied to domain knowledge discovery. Unfortunately, compared with the biomedical field, other disciplinary fields have few reports on measurement analysis based on domain knowledge entities within literature. The reasons are threefold: First, it is difficult to obtain full-text data from publishers other than PubMed, and both acquisition and reuse of full-text data face challenges. Second, knowledge entity relationship construction in the biomedical field is often limited to full-text XML documents and articles with PMIDs, which are relatively easy for machines to read and process, whereas full texts in other scientific and technological fields are often PDFs requiring paid download, and converting them to machine-readable text format is not only time-consuming but also compromises accuracy. Third, compared with the rich and complete dictionaries in the biomedical field, other scientific and technological fields evolve rapidly at the frontiers with diverse change directions, making it difficult to form comprehensive knowledge entity dictionaries covering all directions of a field, thus hindering high-accuracy and high-recall extraction processing.

## 3. Basic Theory Explanation

### 3.1 Knowledge Element Variation Theory

The renowned Chinese scientist of science Zhao Hongzhou once proposed [16]: "Any scientific creation process first dissociates crystallized knowledge units and then recrystallizes them in a completely new thinking potential field. This process is not simple repetition but produces entirely new knowledge systems and knowledge units through recombination." Similarly, viewing a research front as a complex scientific knowledge ecosystem and its evolution as a process of scientific knowledge thought recrystallization, then within the specific knowledge scope surrounding the frontier theme, the internal knowledge of the frontier also undergoes knowledge element dissociation and recombination, evolution and sublimation, derivation and transformation, forming an ascending process from simple to complex and from low-level to high-level for the research front's "knowledge crystallization." During this period, certain key knowledge elements may play the role of "knowledge genes," determining the promotion and mutation of knowledge in specific domains.

J.D. Bernal, the founder of the science of science, believed [17]: "As a requirement of science itself, the formation and selection of topics are the most complex stages in research work. Generally speaking, proposing a topic is more difficult than solving it, and evaluating and selecting topics becomes the starting point of research strategy." Reflected in research fronts, the creation of research fronts often concentrates in newly emerging scientific directions, which arise from new scientific topic selections, which in turn originate from new scientific concepts or cognitions. Specifically, in scientific and technological literature on perovskite solar cells, what represents authors' initial scientific cognition and core scientific concepts are the "Methods" or "Experimental Section" descriptions in scientific texts. When scientists make a new scientific discovery or invention, whether technological innovation and upgrading or new material preparation and R&D, they elaborate on it in detail in this section to facilitate peer supervision and scientific experiment repetition. If a frontier theme forms during a certain period, the scientific phenomenon manifests as the emergence of a new scientific concept, discovery, method, technology, or material, which in literature maps to mutations in experimental material components, equipment technologies, and operational methods—this is the variation phenomenon of knowledge elements. In other words, compared with the previous period's knowledge ecosystem, the variation in knowledge element structure originates from changes in knowledge elements representing specific knowledge forms and features, where frequently appearing knowledge element components from the previous period are replaced by suddenly emerging knowledge elements.

### 3.2 Knowledge Element Variation Phenomenon in ESI Research Fronts

From the perspective of knowledge element transition and recombination, evolving ESI research fronts contain many knowledge element variation phenomena. Using perovskite solar cell fronts as an example, transparent conductive glass substrates, metal counter electrodes, perovskite light-absorbing layers, electron transport layers, and hole transport layers are the most important core components of perovskite solar cells [18]. If the material component of any one of these five core devices changes—where one material is replaced by another (i.e., the knowledge element representing the research material varies)—it triggers the recombination of material structural components in perovskite solar cell devices, thereby affecting the overall photoelectric conversion efficiency and stability characteristics of the solar cell. For instance, when all other device material components and preparation temperatures remain identical, simply replacing the electron transport layer material from single anatase (anatase $TiO_2$) [19] to a composite of anatase and nanofiber $TiO_2$ (anatase $TiO_2$ & nanofiber $TiO_2$) [20] results in different performance in short-circuit current density, open-circuit voltage, fill factor, and photoelectric conversion efficiency, with significant differences in most performance indicators. This variation is based only on homogeneous $TiO_2$ materials; if $TiO_2$ were replaced with other chemical materials from different groups, the resulting performance changes in perovskite solar cells would be even greater.

Thus, signals of variation in the internal components of knowledge element structures in ESI research fronts often signify scientific knowledge recrystallization movements triggered by knowledge element recombination. Under certain conditions, these catalyze changes in the scientific connotation represented by knowledge elements, further promoting the reconstruction of scientific elements and causing transformations in scientific research characteristics and properties.

In summary, this paper defines knowledge element variation in the perovskite solar cell domain within ESI research fronts as follows:

**Definition 1:** The N chemical materials constituting perovskite solar cell devices—transparent conductive glass substrate (Substrate), metal counter electrode (Pole), perovskite light-absorbing layer (Layer), electron transport layer (ETM), hole transport layer (HTM), etc.—are viewed as an N-element knowledge element tuple {M_S, M_P, M_L, M_E, M_H, …, M_N}. If at time $T_1$ the knowledge element composition in the text is {$M\_S_1$, $M\_P_1$, $M\_L_1$, $M\_E_1$, $M\_H_1$, …, $M\_N_1$}, and at time $T_2$ at least one knowledge element material component changes—for example, due to electron transport layer material variation generating a new N-element knowledge element tuple {$M\_S_1$, $M\_P_1$, $M\_L_1$, $M\_E_2$, $M\_H_1$, …, $M\_N_1$}—then a knowledge element variation phenomenon has occurred. In this phenomenon, the position of the varied knowledge element in the tuple remains unchanged (corresponding to scientific literature experimental text corpora, where the component composition and arrangement order

of knowledge element materials in the contextual position have not changed). However, precisely because it is in the same position representing the same perovskite solar cell component material, the change in semantic components triggers changes in overall battery performance and technical characteristics.

## 4. Research Method

The description of knowledge element semantics in this paper is primarily based on the distributional hypothesis in computational linguistics [21]. The distributional hypothesis posits that word semantics and semantic comparisons are determined by their contextual content. The "semantics" of knowledge elements studied in this paper are judged based on their context (i.e., surrounding knowledge elements) and the scope of the frontier theme they belong to, referring to the meaning of knowledge elements after entering communication (i.e., the arrangement and application of chemical materials represented by knowledge elements in different contexts).

In fact, corresponding to previous research [4], when knowledge elements representing innovative scientific concepts or materials evolve and change, their impact on other knowledge elements within the new frontier theme essentially changes the contextual environment of other knowledge elements (caused by changes in knowledge elements at the same position leading to component arrangement changes). Therefore, since the original text corpus extracts the experimental sections from the full text of each citing document, if the corpus after cleaning and denoising through POS tagging filters has a structure that is essentially a "bag of knowledge elements" with contextual information, where knowledge elements are not randomly distributed but continuously aligned according to scientists' original experimental steps. As research fronts evolve, the composition of knowledge elements—from key device materials in perovskite solar cells to reagents and solvents used in early experiments—will vary. At this point, distributed semantic methods can precisely represent and distinguish the semantics and pragmatics of knowledge elements through neural network learning of their sequential positions. If, on this basis, an indicator can be found to quantitatively measure the sudden variation degree of clustered knowledge elements, the synchronous and diachronic knowledge element variation within research fronts can be clearly revealed.

Therefore, this paper first uses the Word2vec word embedding model (Continuous Bag-of-Words, CBOW model) constructed based on the distributional hypothesis to train knowledge elements into word vectors based on context. Word vectors represent knowledge elements with similar contexts. Then, by calculating Euclidean distances between knowledge element vectors to construct similarity matrices, K-means clustering is applied to identify clusters of knowledge elements with similar semantic and pragmatic associations. The clustering results represent the outcomes of knowledge element variation movements—clustered knowledge elements share similar contexts, reflect combinations of knowledge elements with identical grammatical and semantic features, possess the same

pragmatic functions, and represent collections of certain devices or materials in perovskite solar cell fronts. Finally, TF-IDF values of knowledge elements within synchronous and diachronic clusters are calculated to quantitatively measure sudden changes in the importance of knowledge elements after variation, thereby mining the characteristics and patterns of knowledge element variation in research front evolution.

The specific research method flow is shown in Figure 1 [Figure 1: see original paper].

### 4.1 Word2vec Word Embedding Model

This paper investigates changes in the contextual structure of knowledge elements over time using the Word2vec word embedding model built upon the distributional hypothesis and neural network-based distributed representation technology. Word2vec includes two models: CBOW (Continuous Bag-of-Words) and Skip-Gram, both producing continuous vectors of real-numbered components with the same dimensionality for all words in the corpus. Since this study aims to solve for changes in contextual knowledge elements in preprocessed text corpora, and the extracted experimental text training corpora are not large-scale datasets, the CBOW model in Word2vec is adopted for machine learning word vector training.

### 4.2 K-means Clustering Algorithm

The most crucial step in implementing the K-means algorithm is the prior determination of k-value selection, as the choice of initial clustering centers significantly impacts clustering results. Based on quantitative selection indicators such as the elbow method and silhouette coefficient, combined with the data distribution of knowledge element communities from previous research [4], multiple preprocessing experiments for k-value selection were conducted to ensure clustering accuracy and maximum convergence. The final findings revealed that for corpora from different time windows, when the k-value is set to 3, the resulting knowledge element semantic clusters are easily interpretable and comparable, with relatively optimal convergence and classification effects. Therefore, the knowledge element word vectors from different time windows are all clustered into 3 semantic clusters, with the maximum iteration count for the K-means algorithm set to the default value of 99.

The visualization results of K-means clustering of knowledge elements from different periods can intuitively judge the stability and knowledge intensity of the scientific knowledge structure within research fronts through the convergence, divergence, and distribution of knowledge element semantic similarity across periods, thereby grasping the internal knowledge flow patterns of ESI research fronts.

### 4.3 TF-IDF Knowledge Element Mutation Measurement

The essential characteristic of knowledge element variation is that a chemical material component represented by a knowledge element suddenly appears in a specific experimental text during a certain period, while other chemical reagents and materials used together with that knowledge element remain unchanged (these chemical reagents and materials appear widely in many experimental texts during this period). In other words, when knowledge element variation occurs, the pattern characteristics of the corresponding text and terms are: the varied knowledge element term appears with high frequency in one experimental text, while its proportion in other experimental texts during the same period is minimal. This indicates that the knowledge element term is highly representative and discriminative for a specific frontier experimental text during this period, serving as an important keyword term unique to that text. The statistical method commonly used to evaluate the importance of a term to a document within a document set or corpus is the TF-IDF "term frequency-inverse document frequency" algorithm. Therefore, this paper uses TF-IDF term weighting technology to represent and calculate the sudden variation degree of knowledge elements within the same semantic cluster across periods.

TF-IDF tends to filter out common knowledge element terms widely applied in experimental texts during a certain period, retaining knowledge element terms with greater sudden variation and higher importance to specific experimental texts. This enables quantitative measurement of the synchronous and diachronic sudden variation degrees of knowledge elements with similar chemical material attributes. Since a specific knowledge element during a certain period may appear in n experimental texts and have n TF-IDF values, to better characterize the sudden variation degree of the knowledge element during this period, this paper uses the average of the n TF-IDF values of the knowledge element in n texts as the mutation measurement indicator. For knowledge element term k at time t, its mutation degree is calculated as shown in Formula 1:

$$tfidf_k = \frac{Sum_{i=1}^{tfidf_{i,j}} tfidf_{i,j}}{} \quad \text{(Formula 1)}$$

## 5. Empirical Study

### 5.1 Word2vec Knowledge Element Vector Training

Since the size of the data corpus affects Word2vec knowledge element vector training results, and the accuracy of machine learning also depends heavily on the magnitude of neural network input layer data, and to enable better comparative analysis and extended verification with previous research [4], this paper continues to use the time label classification of experimental text data from previous research [4]. The corpus for ESI research front evolution is divided into four time windows: 2010-2014 (due to insufficient citing literature data volume in 2010-2013, these years are merged into the 2010-2014 period), 2015,

2016, and 2017. Consistent with the preprocessing method in previous research [4], the preprocessed text corpora for each period after punctuation removal, stop word removal, and N-gram filtering undergo OSCAR4 chemical knowledge entity recognition. POS tagging filters are then used to remove noise data not containing OSCAR4 chemical entity tags, followed by further deduplication and denoising using tools like Notepad++, resulting in a corpus for processing that contains only knowledge element data of OSCAR chemical compounds (CM). Unlike the data processing method in previous research [4], this paper does not perform BOW modeling on the preprocessed and denoised knowledge element texts but directly imports them into the DeepLearning4J (DL4J) neural network toolkit for Word2vec word embedding model training.

Applying Word2vec word embedding technology, knowledge element semantics are represented based on contextual information. The raw corpus for each period is a two-dimensional list after entity recognition, preprocessing, and POS tagging filtering, where each element is the remaining chemical knowledge entity from text processing. These chemical entity knowledge elements appear as strings, represented as follows:

$$Sentences = \{[' first',' knowledgeelement'], [' second',' knowledgeelement']\}, ...$$

Importing these knowledge element terms with original experimental positional arrangement distribution order into the CBOW model of Word2vec enables learning and training of knowledge elements with identical semantics based on the order of surrounding terms. The predicted knowledge elements not only have semantic similarity but more importantly reflect the practical application correlation of chemical components in experiments—the relevance of prepared materials generated with other chemical components. This paper adopts common parameters of the Word2vec model, selecting a word vector dimension of 100. The resulting knowledge element word vectors are shown in Table 1 .

### 5.2 K-means Knowledge Element Similar Semantic Clustering

The most critical step in implementing the K-means algorithm is the prior determination of k-value selection, as the choice of initial clustering centers significantly impacts clustering results. Based on quantitative selection indicators such as the elbow method and silhouette coefficient, combined with the data distribution of knowledge element communities from previous research [4], multiple preprocessing experiments for k-value selection were conducted to ensure clustering accuracy and maximum convergence. The final findings revealed that for corpora from different time windows, when the k-value is set to 3, the resulting knowledge element semantic clusters are easily interpretable and comparable, with relatively optimal convergence and classification effects. Therefore, the knowledge element word vectors from different time windows are all clustered

into 3 semantic clusters, with the maximum iteration count for the K-means algorithm set to the default value of 99.

Figure 2 [Figure 2: see original paper] shows the clustering result distribution after applying the K-means algorithm to knowledge element word vectors from 2010-2014. For easier distinction, knowledge elements from different clusters are represented by nodes of different colors and shapes: Cluster 1 knowledge elements are red squares; Cluster 2 knowledge elements are green flower shapes; Cluster 3 knowledge elements are blue triangles.

The visualization results of K-means clustering of knowledge elements from different periods can intuitively judge the stability and knowledge intensity of the scientific knowledge structure within research fronts through the convergence, divergence, and distribution of knowledge element semantic similarity across periods, thereby grasping the internal knowledge flow patterns of ESI research fronts.

### 5.3 TF-IDF Knowledge Element Mutation Measurement

For example, Figure 2 shows the 2010-2014 knowledge element semantic clustering distribution. The composition of knowledge element semantic clusters during this period is shown in Table 2 : It can be observed that knowledge elements in Cluster 1 and Cluster 2 mostly represent chemical reagents or basic solutions commonly used in early experimental preparation of perovskite solar cells, rarely appearing as core device materials in previous research [4]. Cluster 3, however, focuses more on core chemical components that constitute key device materials, such as $MASnX_3$, $CsSnI_3$, $MAPbI_3$, which can all be applied to perovskite solar cell light-absorbing film preparation, while $In_2O_3$, $Sb_2S_3$, NiO, ZnO, PbS are often applied to electron transport layers, hole transport layers, or scaffold blocking layers.

Word2vec can automatically identify knowledge elements with similar semantic and pragmatic functions through their positional arrangement in experiments. The data segmentation and clustering function of K-means can achieve associative classification of knowledge element objects representing specific semantic functions, making semantic correlations higher within groups and semantic differences higher between groups. After semantic representation and classification processing, the knowledge element sets provide natural datasets for measuring sudden variation degrees based on TF-IDF for knowledge elements within the same semantic cluster across periods.

On this basis, TF-IDF mutation calculation can identify the sudden variation degree of knowledge elements with similar chemical semantics and material functions for frontier theme texts during specific periods, thereby early detecting the potential impact utility of knowledge elements on future innovative development of the frontier field.

For example, Table 2 shows that in 2010-2014, the knowledge elements with the

highest mutation degree for perovskite light-absorbing film materials in Cluster 3 were $MASnX_3$, $CsSnI_3$, and $MAPbI_3$. In the analysis results of previous research [4] using the knowledge element co-occurrence method, $MAPbI_3$ was accurately identified as a high co-occurrence knowledge element in 2010-2014, while $MASnX_3$ and $CsSnI_3$, due to low co-occurrence frequency, could not be identified as low-frequency terms in experimental texts until they appeared as high co-occurrence knowledge element pairs in 2017. In this paper, through mutation degree calculation, $MASnX_3$ and $CsSnI_3$ can not only be accurately identified in 2010-2014 but also show high variation degrees during this period. As "knowledge terrain" mutation "knowledge potential fields," they represent key signals that may affect future technological innovation directions. As mentioned in previous research [4], scientists have been committed to solving the environmental pollution problem of toxic heavy metal Pb in perovskite solar cells in recent years, making environmentally friendly lead-free perovskite solar cells such as $MASnX_3$ and $CsSnI_3$ hot research directions, which also provides corroborating evidence for the method and judgment in this paper.

Thus, the scientometric method based on knowledge element variation not only supplements and extends the knowledge element co-occurrence method from previous research [4], making the mining of internal knowledge movement laws within research fronts more detailed and specific, but also serves as powerful evidence for early and timely detection of future development trends and key intelligence signals of research fronts under temporal changes.

### 5.4 ESI Research Front Evolution Analysis Based on Knowledge Element Variation

The experimental text datasets extracted for each time window all undergo Word2vec knowledge element vector training, K-means knowledge element similar semantic clustering, and TF-IDF knowledge element mutation measurement. This enables the depiction of knowledge flow and change patterns within ESI research fronts as they evolve over time, from horizontal to vertical perspectives. To facilitate unified comparison and better interpretation and analysis of knowledge variation characteristics of research fronts in each period, each time window ultimately generates 3 knowledge element semantic clusters (for easier distinction, knowledge element nodes in different clusters are represented by different colors and shapes: Cluster 1 knowledge elements are red squares; Cluster 2 knowledge elements are green flower shapes; Cluster 3 knowledge elements are blue triangles). Each cluster displays the top 15 knowledge elements ranked by TF-IDF mutation degree values.

#### 5.4.1 2010-2014 Research Front Knowledge Variation Characteristics

The knowledge element semantic cluster distribution results for the ESI research front in 2010-2014 are shown in Figure 3 [Figure 3: see original paper]. The distribution of top 15 high-mutation-degree knowledge elements in each cluster is shown in Table 3 .

It can be observed that 2010-2014 represents the germination period of perovskite solar cell research. During this period, knowledge element nodes are relatively compact and concentrated, with clear edge segmentation among knowledge element semantic clusters. In terms of semantic and pragmatic functions, knowledge elements with higher mutation degrees in Cluster 1 and Cluster 2 mostly represent chemical reagents and basic solutions used in early experimental preparation of perovskite solar cells. Cluster 3 reflects more of the application focus on core device materials such as light-absorbing layers and electron transport layers in perovskite solar cells during this period.

Cluster 1 focuses on polydimethylsiloxane (PDMS) for preparing blocking layer substrates, dimethylformamide (DMF) for dissolving halide perovskites, epoxy resin for preparing scaffold materials, and silver bismuth sulfide ($AgBiS_2$) for synthesizing nanocrystals with other materials [22]. Cluster 2 focuses on lithium bis(trifluoromethanesulfonyl)imide (Li-TFSI) as a hole transport capacity enhancer, IPFB for inhibiting battery performance degradation, and mixed solutions such as hydrochloric acid and acetic acid commonly used for dissolving and preparing thin films. Notably, in Cluster 1, barium titanate ($BaTiO_3$) and BFO ($BiFeO_3$) both exhibit excellent ferroelectric dielectric properties, while $AgSbS_2$ and $Zn_2TiO_4$ share sensitization characteristics. In Cluster 2, $Sb_2Se_3$, $Ag_2S$, and CQDs are often used as quantum dot materials. Due to their identical pragmatic functions, these knowledge elements are not only accurately identified in the same semantic cluster but also show high mutation degree values. In Cluster 3, copper antimony sulfide ($CuSbS_2$) solar cell absorption layers, and lead-free perovskite light-absorbing films $MASnX_3$ and $CsSnI_3$ attract scientific attention, with mutation degree indicators higher than the widely used $MAPbI_3$ light-absorbing material. Diiodooctane (DIO) is also found to have significant impact on device performance improvement [23].

**5.4.2 2015 Research Front Knowledge Variation Characteristics** The knowledge element semantic cluster distribution results for the ESI research front in 2015 are shown in Figure 4 [Figure 4: see original paper]. The distribution of top 15 high-mutation-degree knowledge elements in each cluster is shown in Table 4 .

In 2015, knowledge element clusters began to show signs of cross-penetration, with semantic distances between cluster convergence centers becoming closer. Knowledge element nodes in each cluster diffused outward, indicating entry into the preliminary development period of the research front. Scientists used a richer variety of chemical materials in experiments, resulting in greater two-dimensional semantic coordinate distance intervals mapped by knowledge element word vectors. The closer intervals between semantic cluster centers prove that scientists began forming consensus on component materials with relatively high efficiency and stable performance for solar cells during this period.

In terms of semantic and pragmatic functions, Cluster 2 knowledge elements are mostly newly concerned solution reagents or intermediate product modi-

fiers required for early perovskite solar cell preparation. Examples include zinc stannate ($Zn_2SnO_4$) for promoting perovskite layer crystallization [24], $In(OH)_3$ commonly used as a metal-based alkaline solution, polyethylene oxide (PEO) as an interface modification material for inhibiting reverse current [25], and hydrophobic functional group ODT (n-octadecanethiol) used for substrate molecular film preparation, as well as fullerene derivatives (C-PCBSD) for modifying ZnO electron transport layers.

Cluster 1 and Cluster 3 knowledge element nodes are more dispersed, with larger distances between nodes and relatively greater differences in semantic and pragmatic categories. Cluster 1 focuses more on doping preparation materials for perovskite light-absorbing layers: $MnO_2$ can be used as a porous skeleton for low-resistance metal oxide doping to prepare light-absorbing layers; $MoS_2$ doping of perovskite light-absorbing layers can effectively improve photoelectric properties and stability; $WSe_2$ can be used for preparing ultra-thin flexible solar cells. Cluster 3 mostly contains basic chemical reagents commonly used in experiments, but also includes important frontier signals: $HPbI_3$ can simplify perovskite film synthesis and repair preparation processes [28]; PEG can improve device photovoltaic performance; PVP can significantly improve photoelectric efficiency [29], all becoming important concerns in perovskite solar cell research during this period.

**5.4.3 2016 Research Front Knowledge Variation Characteristics** The knowledge element semantic cluster distribution results for the ESI research front in 2016 are shown in Figure 5 [Figure 5: see original paper]. The distribution of top 15 high-mutation-degree knowledge elements in each cluster is shown in Table 5 .

In 2016, knowledge elements in each semantic cluster showed more dispersed characteristics, with clearer boundary divisions between clusters, weakened overlapping phenomena, larger distances between cluster centers, and significantly increased knowledge element content within clusters. This proves that the research front entered a rapid development period. Through this phenomenon, it is also observed that with the R&D design of new materials, scientists during this period not only expected to continue improving the photoelectric efficiency and stability of perovskite solar cells but also aimed to solve many industrialization problems to enable early large-scale production application.

In terms of semantic and pragmatic functions, many new "knowledge landscape" sudden signals emerged during this period. Knowledge element mutation degree values in Cluster 2 are relatively high, focusing mainly on compounds and related materials for enhancing preparation performance. For example, ALD-$TiO_2$, SAF-Ome, CuO-$Cu_2O$, $In_2O_3$-MWCNTs, and other composite materials with similar mutation degree values mostly serve to enhance preparation performance: atomic layer deposition (ALD) preparation of $TiO_2$ can significantly improve cell efficiency; SAF-OMe hole transport capacity is more than three times higher than Spiro-OMeTAD [27]; CuO-$Cu_2O$ semiconductor nanorod ar-

rays can effectively catalyze photoelectrochemical synthesis reactions. Meanwhile, AgCuS, $AgBiS_2$, $BaSi_2$, $Zn_2SnO_4$ (ZSO), $BiFeO_3$, $Cs_4PbBr_6$, $CsBi_3I_{10}$, and other series of photosensitive semiconductor and nanocrystal materials are commonly applied to solar thin film preparation.

Knowledge element nodes in Cluster 1 and Cluster 3 are more dispersed, with larger distances between nodes and relatively greater differences in semantic and pragmatic categories. Cluster 1 focuses more on doping preparation materials for perovskite light-absorbing layers: $MnO_2$ can be used as a porous skeleton for low-resistance metal oxide doping to prepare light-absorbing layers; $MoS_2$ doping of perovskite light-absorbing layers can effectively improve photoelectric properties and stability; $WSe_2$ can be used for preparing ultra-thin flexible solar cells. Cluster 3 mostly contains basic chemical reagents commonly used in experiments, but also includes important frontier signals: $HPbI_3$ can simplify perovskite film synthesis and repair preparation processes [28]; PEG can improve device photovoltaic performance; PVP can significantly improve photoelectric efficiency [29], all becoming important concerns in perovskite solar cell research during this period.

**5.4.4 2017 Research Front Knowledge Variation Characteristics** The knowledge element semantic cluster distribution results for the ESI research front in 2017 are shown in Figure 6 [Figure 6: see original paper]. The distribution of top 15 high-mutation-degree knowledge elements in each cluster is shown in Table 6 .

In 2017, each semantic cluster showed more cross-convergence and overlapping penetration trends, with knowledge element distributions within clusters showing obvious convergence and focusing signs, and distances between cluster centers further reduced. This indicates that the research front entered a stable development period. The types of knowledge element materials with similar semantic and pragmatic functions significantly increased, core device materials constituting perovskite solar cells tended to stabilize, and the chemical function classification of knowledge element materials between clusters in experiments became less clearly bounded than in previous years. However, the experimental application purposes of knowledge element materials within clusters became more similar.

Cluster 1 and Cluster 3 knowledge element distribution patterns are relatively similar, though Cluster 3 tends to be more convergent with higher semantic similarity between knowledge elements. Their similar pragmatic functions focus on semiconductor electron transport materials: ZnO and $TiO_2$ as metal oxide semiconductor materials; $PC_{61}BM$ and PCBM as fullerene derivatives commonly used in electron transport materials; $SnI_2$, $SnO_2$, $SnF_2$ for preparing $MASnI_3$ lead-free clean solar cells. Knowledge element mutation degree values in Cluster 1 are significantly higher than in Cluster 3, containing many important signals for device performance innovation: Xi'an Jiaotong University Wu Chaoxin's team achieved efficient flexible lead-free formamidinium tin iodide

($FASnI_3$) perovskite solar cells, attracting attention [30]; utilizing gradient band gaps to improve solar spectrum utilization is an unsolved challenge for tandem cells, and scientists proposed a method using hexagonal boron nitride (h-BN) as an intermediate monolayer to form gradient band gaps, causing a sensation [31].

Knowledge elements in Cluster 2 are arranged more compactly, with more similar semantic and pragmatic functions, focusing mainly on new semiconductor crystal photosensitive materials and carbon nanotube series materials. For example, GeP, CuS, $MoS_2$, PbSe@CdSe, CIGSSe, $Y_2O_3$, $Cu_2S$, and other semiconductor crystal photosensitive materials are commonly applied to quantum dot, dye-sensitized, and multi-junction solar cell devices. Knowledge elements such as AgNCs (silver nanoclusters), OLCNS:Ag (onion-like carbon nanospheres composite silver), MWCNTs (multi-walled carbon nanotubes), NWs (nanowires), and single-walled carbon nanotubes (SWNTs/SWCNTs) reflect scientists' attention to carbon nanotube series materials during this period [32].

## Conclusion

As an intelligence practice exploration oriented toward disciplinary domain needs, based on full-text analysis and key semantic computation, this paper first employs the distributional hypothesis theory to train knowledge element semantics using the Word2vec model. Then, the K-means clustering algorithm is used to identify semantic and pragmatic classifications and interaction relationships of knowledge element clusters. Finally, TF-IDF values of knowledge elements within the same semantic cluster are used to calculate their sudden variation degree relative to frontier theme texts, thereby detecting knowledge evolution characteristics and key innovation signals of ESI research fronts under temporal changes.

Through expert verification and professional literature confirmation, this method can effectively identify key intelligence signals that may promote frontier innovation development in each period. The analysis results from previous research [4] also serve as verification for this paper's results. In fact, the identification results based on knowledge element co-occurrence in previous research [4] more reflect frontier hotspot directions that have reached a certain level of 热度, while the results in this paper essentially represent "knowledge potential field" signals emerging in the "knowledge landscape" map of each period, more likely to be key turning nodes of knowledge migration in frontier evolution—newly emerging frontier directions. The signals identified in this paper are often not overall innovations of major perovskite solar cell components but more commonly overall performance improvements triggered by minor improvements and upgrades related to core devices (or the addition/doping of minor experimental steps such as chemical reagents). In reality, the process from minor improvements to qualitative leaps in overall device performance truly reflects the actual development trajectory of science—from points to surfaces, from small key breakthroughs driving overall scientific and

technological innovation R&D.

It should be noted that although the theoretical method and technical solution proposed in this paper only use ESI research front data as a case study, the entire set of ideas and method design is not limited to ESI research fronts. It has strong universality and reference value for traditional "research fronts" identified using citation relationships, term relationships, and other measurement indicators. Therefore, future work will purposefully apply this method and technical solution to the mining and analysis of other research front data and evolution patterns.

## References

[1] Sun Tan. The future of library intelligent knowledge services [J]. Journal of Library Science in China, 2021, 47(2): 15-18.
[2] Song Ningyuan, Pei Lei, Wang Chunying. Research progress and trend analysis of scientific paper semantic enhancement [J]. Library and Information Service, 2021, 65(1): 82-90.
[3] Zeng Jianxun. Thoughts on the development of China's scientific and technological intelligence undertaking during the "14th Five-Year Plan" period [J]. Information Studies: Theory & Application, 2021, 44(1): 1-7.
[4] Sun Zhen, Leng Fuhai. An ESI research front knowledge evolution analysis method based on knowledge element co-occurrence [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(11): 1095-1113.
[5] Leng Fuhai, Sun Zhen, Zhou Qiuju. The development practice and related discussion of the *2015 Research Fronts* report [J]. Think Tank: Theory & Practice, 2016, 1(2): 79-87.
[6] Clarivate Analytics. Clarivate and the Chinese Academy of Sciences release annual joint report to identify 100+ research fronts [EB/OL]. [2020-11-13]. https://clarivate.com/news/clarivate-and-the-chinese-academy-of-sciences-release-annual-joint-report-to-identify-100-research-fronts/.
[7] Wang Xiaomei, Deng Qiping, Li Guopeng, et al. Scientific mapping of ESI research fronts and application in the nanotechnology field [J]. Library and Information Service, 2017, 61(12): 106-112.
[8] Sun Zhen, Leng Fuhai, Zhang Jinhui. Knowledge element-based scientometric methods and empirical research [J]. Library and Information Service, 2017, 61(23): 89-99.
[9] Sun Zhen, Leng Fuhai. Analysis of a new scientometric paradigm based on knowledge elements [J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(6): 555-564.
[10] Ding Y, Song M, Han J, et al. Entitymetrics: measuring the impact of entities [J]. PLoS ONE, 2013, 8(8): e71416.
[11] Song M, Hann NG, Kim YH, et al. Discovering implicit entity relation with the gene-citation-gene network [J]. PLoS ONE, 2013, 8(12): e84639.
[12] Yu Q, Ding Y, Song M, et al. Tracing database usage: detecting main paths in database link networks [J]. Journal of Informetrics, 2015, 9(1): 1-15.

[13] Lee D, Kim WC, Charidimou A, et al. A bird's-eye view of Alzheimer's disease research: reflecting different perspectives of indexers, authors, or citers in mapping the field [J]. Journal of Alzheimer's Disease, 2015, 45(4): 1207-1222.
[14] Lee K, Kim SY, Kim EH, et al. Comparative evaluation of bibliometric content networks by tomographic content analysis: an application to Parkinson's disease [J]. Journal of the Association for Information Science and Technology, 2017, 68(5): 1295-1307.
[15] Li K, Yan E. Co-mention network of R packages: scientific impact and clustering structure [J]. Journal of Informetrics, 2018, 12(1): 87-100.
[16] Zhao Hongzhou, Jiang Guohua. Knowledge units and exponential laws [J]. Science of Science and Management of S.& T., 1984(9): 39-41.
[17] Bernal JD. The strategy of scientific research [C]//Translation Collection of Science of Science. Beijing: Science Press, 1980.
[18] Yao Xin, Ding Yanli, Zhang Xiaodan, et al. Review of perovskite solar cells [J]. Acta Physica Sinica, 2015, 64(3): 135-142.
[19] Burschka J, Pellet N, Moon SJ, et al. Sequential deposition as a route to high-performance perovskite-sensitized solar cells [J]. Nature, 2013, 499(7458): 316-319.
[20] Dharani S, Mulmudi HK, Yantara N, et al. High efficiency electrospun $TiO_2$ nanofiber based hybrid organic-inorganic perovskite solar cell [J]. Nanoscale, 2014, 6(3): 1675-1679.
[21] Harris ZS. Distributional structure [J]. Word, 1954, 10(2/3): 146-162.
[22] Chen C, Qiu X, Ji S, et al. The synthesis of monodispersed Ag-$BiS_2$ quantum dots with giant dielectric constant [J]. CrystEngComm, 2013, 15(38): 7644-7648.
[23] Liang PW, Liao CY, Chueh CC, et al. Additive enhanced crystallization of solution-processed perovskite for highly efficient planar-heterojunction solar cells [J]. Advanced Materials, 2014, 26(22): 3748-3754.
[24] Bera A, Sheikh AD, Haque MA, et al. Fast crystallization and improved stability of perovskite solar cells with $Zn_2SnO_4$ electron transporting layer: interface matters [J]. ACS Applied Materials & Interfaces, 2015, 7(51): 28404-28411.
[25] Dong Haopeng. Interface modification before organic-inorganic hybrid perovskite film formation and its device photovoltaic characteristics [D]. Beijing: Tsinghua University, 2015.
[26] Ma Dongchao. Preparation study of Ag nanophase absorption-enhanced $CH_3NH_3PbI_3$ perovskite thin films and cells [D]. Harbin: Harbin Institute of Technology, 2015.
[27] Wang YK, Yuan ZC, Shi GZ, et al. Dopant-free spiro-triphenylamine/fluorene as hole-transporting material for perovskite solar cells with enhanced efficiency and stability [J]. Advanced Functional Materials, 2016, 26(9): 1375-1381.
[28] Pang S, Zhou Y, Wang Z, et al. Transformative evolution of organolead triiodide perovskite thin films from strong room-temperature solid-gas interaction between $HPbI_3$-$CH_3NH_2$ precursor pair [J]. Journal of the American Chemical Society, 2016, 138(3): 750-753.
[29] Li Jianfeng, Zhao Chuang, Zhang Heng, et al. Improving photovoltaic

performance of perovskite solar cells using PVP additives [J]. Chinese Journal of Luminescence, 2016(1): 56-62.
[30] Xi J, Wu Z, Jiao B, et al. Multichannel interdiffusion driven $FASnI_3$ film formation using aqueous hybrid salt/polymer solutions toward flexible lead-free perovskite solar cells [J]. Advanced Materials, 2017, 29(23): 1606964.
[31] Ergen O, Gilbert SM, Pham T, et al. Graded band gap perovskite solar cells [J]. Nature Materials, 2017, 16(5): 522.
[32] Li Z, Dong J, Liu C, et al. Improved optical field distribution and charge extraction through an interlayer of carbon nanospheres in polymer solar cells [J]. Chemistry of Materials, 2017, 29(7): 2961-2968.

## Author Contributions

**Sun Zhen:** Designed the paper framework and 思路, collected and analyzed data, wrote and revised the paper.
**Leng Fuhai:** Proposed the research proposition and main ideas, reviewed and revised the paper.

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