https://doi.org/10.51408/issi2025_041

A Comparative Study on Text Multi-Features Mining for Patent Text Clustering: The Case of Graphene Sensing Technology

Xian Zhang[1], Jiahui Li[2], Shuying Li[3], Haiyun Xu[4]

[1]*zhangx@clas.ac.cn*, [2]*lijiahui222@mails.ucas.ac.cn*, [3]*lisy@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[4]*xuhaiyunnemo@gmail.com*
Business School, Shandong University of Technology, Zibo (China)

Abstract

The development of text feature extraction and measurement methods has given rise to a diversification of perspectives on text mining. However, few studies have explored the similarity, complementarity, and effectiveness of different text features. The selection of different feature combinations lacks a supporting basis. This study selected four types of text feature words from patent texts, namely, text domain feature keywords by Comprehensively Measure Feature Selection algorithm (CMFS), technical interdisciplinary keywords by the term Interdisciplinary index (TI), technical breakthrough keywords by the Kleinberg burst detection algorithm (KB) and high-frequency words (HF). A set of measurement indicators and implementation methods based on the Jaccard distance index, information entropy, and mutual information theory was designed, to determine the similarities, differences, synergies, and complementarities of the four types of text feature words. Based on comparative analysis, a comprehensive measurement index was designed, as well as feature combinations were selected. To illustrate this approach, we selected patent documents in the domain of graphene sensing and evaluated various feature combinations with different word embedding and clustering algorithms. The results show that multivariate features enhance the effectiveness of single high-frequency features in text mining tasks. There is a wide range of applicability for CMFS+KB feature combination, with the clustering effect being optimal when used with FsatText+K-means. For the specific case of HDBSCAN+FastText, the HF+CMFS+KB feature combination demonstrates superior performance. This study corroborates the information representation significance and complementarity of four types of keywords in information representation, thereby substantiating the extraction and analysis of text multi-features. Finally, we also point out the limitations of measurement methods and feature types in the research and prospects for future research.

Introduction

Text feature words refer to keywords that can represent the main theme, meaning, content and other features of the text. They are widely used in fields such as information retrieval and text classification (Chi et al., 2019). Text feature words are usually extracted directly from the text. Word frequency represents the most widely applied fundamental method for extracting text feature words. This method reveals the text topic by analyzing and describing lexical rules (Feng Guohe & Kong Yongxin, 2020; Salton, Allan, & Singhal, 1996). For example, high-frequency words often dominate in topic classification and identification (Li, Zhang, Li, Ouyang, & Cai, 2018).

However, in the field of patent text mining, topic models often tend to favor high-frequency words and have limitations in implicit semantic expression (Yu Yan & Zhao Naixuan, 2018). Cassandra L. Jacobs et al. (Jacobs, Dell, Benjamin, & Bannard, 2016) also proposed that high-frequency words are more easily recognized in cognitive processes, while low-frequency words exhibit enhanced recognizability and potentially contain more significant information. Therefore, beyond the comprehensive mining of high-frequency features, there has been a surge of interest within the academic community in the mining of selecting low-frequency words as a complement to high frequency words. In addition, multi-feature extraction has been found to be more conducive to the accuracy of machine understanding for text mining (Cheng Yong, Xu Dekuan, & Lv Xueqiang, 2019). The perspective of extracting text feature words is constantly enriched, such as revealing important features of the field, technical interdisciplinary features, and technical breakthrough features, which have been widely used in research.

On the one hand, there are few studies on how the words extracted from different feature relationships represent the text topics; the similarities and differences between these representations; the potential supplementary role of these representations for high-frequency words; and how to quantitatively measure the differences in their information meaning. On the other hand, few researchers have explored the practical effects of different combinations of multiple features in text mining tasks.

This study aims to address this gap by conducting quantitative measurement and comparative research on the text topic representation effects of different types of text feature words. The objective is to provide scientific and quantitative reference for text topic feature mining. To illustrate this, we selected four types of text feature words from patent texts, namely, text domain feature keywords (CMFS), technical interdisciplinary keywords (TI), technical breakthrough keywords (KB) and high-frequency words (HF). We then designed a set of comprehensive measurement indicators for feature combinations and implementation methods based on similarity, information entropy, and mutual information theory. A comparative study was conducted on the similarity, difference, complementarity, and synergy of the text representations of four types of text feature words. Different feature combinations were applied to three word embedding models and three clustering algorithms to explore the application effect of multi-feature combination in text clustering.

The rest of this study is as follows: Section 1, Introduction; In Section 2, we reviewed the application of word embedding and clustering algorithms, as well as the extraction and selection methods of different text feature words. Based on this design, the research framework is obtained; In Section 3, we introduced the data source and vocabulary extraction. We also propose comparative analysis methods, comprehensive indicator design, word embedding model and clustering algorithm; In Section 4, we present the empirical results and analysis; In Section 5, we summarize the characteristics and usage scenarios of multi-feature combinations based on the results; Finally, we summarize the theoretical and practical significance of this study, as well as its limitations and future research directions.

Literature review

Word embedding and text clustering

In text mining tasks, vocabulary is the core unit and the basic representation form of knowledge content in the semantic field. With the maturity of word representation technology in natural language processing, existing research has mostly used word embedding models to generate vocabulary semantic vectors to achieve more accurate vocabulary semantic analysis (Chen G., Xu, Hong, Wu, & Xiao, 2024). Commonly used word embedding models include Word2Vec, GloVe, and FastText (Borah, Barman, & Awekar, 2021). All three models use context information of words to capture the semantic relationship of words. Word2Vec optimizes the objective function to ensure that the distance between word vectors in similar contexts is close; FastText, based on Word2Vec, additionally captures structural information such as the internal character composition of words; GloVe represents semantics through the co-occurrence frequency of features in the entire corpus. Studies have shown that most word embedding models randomly initialize vectors, and the resulting semantic space is uncertain. Their default tokenizer often only performs simple word segmentation operations, and less work is done on screening feature combinations. On the same data, the word vectors generated by two trainings are different, and the semantic fields formed by the nearest neighbors of the words do not completely overlap (Kutuzov, Øvrelid, Szymanski, & Velldal, 2018; Rettenmeier, 2020). Therefore, reducing the uncertainty of word vectors is one of the key points of word embedding. For example, studies have shown that using a custom corpus can significantly improve the effect of text mining (Ercan & Cicekli, 2016; Nguyen, Billingsley, Du, & Johnson, 2015). In addition, N-gram Categories (i.e., phrases consisting of multiple words) show better performance in text classification (Kruczek, Kruczek, & Kuta, 2020).

Through the word embedding model, various text forms such as sentences, paragraphs, and documents can be represented as vectors, thereby realizing the combination with machine learning methods. One of the text clustering methods is to cluster text vectors into sentences, paragraphs, or documents through clustering methods such as K-means (Ji, Liu, Peng, & Kong, 2024). In addition to K-means, other commonly used algorithms for text clustering include Agglomerative Clustering (Enguix, Carrascosa, & Rincon, 2024), HDBSCAN (Inje, Nagwanshi, & Rambola, 2024), etc. Their clustering performance varies in different scenarios.

Text multi- feature extraction and seletion

High-frequency words have achieved rich application results in fields such as text topic classification and recognition (Qaiser & Ali, 2018; Tseng, Lin, & Lin, 2007). For example, the Vector Space Model (VSM) mainly uses word frequency to represent feature vectors and derives indicators such as TF-IDF, which is the most widely used (Choi, Oh, Choi, & Yoon, 2018). In addition, from the perspective of statistical features, the following two categories of text feature words are of particular concern. The first is to examine the ability of feature words to represent the characteristics of the technical domain from a global perspective, measure the

core influence and representativeness of feature words in the domain, and extract appropriate feature words with domain characteristics. The domain characteristic indicators used are mostly selected based on metrological and statistical features, mainly including TF-IDF (Chawla, Kaur, & Aggarwal, 2023), information gain (Yu, Ju, & Shang, 2022), Gini coefficient (Mengle & Goharian, 2009), and Comprehensively Measure Feature Selection (CMFS) (Yang, Liu, Zhu, Liu, & Zhang, 2012). They are mostly based on one kind of feature, among which CMFS integrates the comprehensive measurement of domain characteristics within and outside the class and has relatively good domain representation. The second is multi-feature, with more attention paid to technical interdisciplinary features (Yao, Wang, Wu, Xu, & Zhang, 2023) and technical breakthrough features (Jia et al., 2021; Liu Yahui, Xu Haiyun, Wu Huawei, Liu Chunjiang, & Wang Haiyan, 2023). Their effective identification methods are mostly achieved through relevant quantitative measures. Among them, interdisciplinary feature indicators mainly include Citation Outside Category index (COC) (Porter & Chubin, 1985), Weighted Citation Outside Category index (WCOC) (K. Chen & Chiung-fang, 2004), Shannon-Wiener Index (SWI) (Shannon, 1948), Brillouin's Index (BI) (Chang & Huang, 2012) and Terms Interdisciplinarity index (TI) (Xu, Guo, Yue, Ru, & Fang, 2016). Their main idea is to measure the degree of cross-integration between features. For example, the TI index considers cross-domain features and influence. So, its scalability is comprehensively good. Breakthrough features often have the characteristics of novelty, foresight, uncertainty, and nonlinearity. Scholars often start with the attributes of the technology itself or combine complex network methods for identification. For example, the identification method based on word frequency growth rate (Feng, Wu, & Mo, 2020), the identification method based on TRIZ theory (Vicente-Gomila, Artacho-Ramirez, Ting, & Porter, 2021), and the burst monitoring algorithm proposed by Kleinberg (Kleinberg, 2002). Among them, the Kleinberg burst detection algorithm is widely recognized by the academic community.

In the research on multi-feature technology topics mining, there is a significant impact on text mining results by feature selection (Büyükkeçeci & Okur, 2023). The results obtained by using features of different indicators and methods may be completely different(Zhang, Sun, Chinchilla-Rodríguez, Chen, & Huang, 2018). Scholars usually conduct comparative analysis of features from two major perspectives. First, from the information perspective, by comparing the differences in information content, richness, and synergy between different features, the similarity between different features is obtained based on indicators such as information entropy, mutual information, and information gain(Wang, Lu, & Tai, 2015), and the feature weights are assigned(Prabowo & Thelwall, 2006). The second is to explore the intrinsic connections between different features at the semantic level from a semantic perspective, thereby achieving topic clustering(Zhao, Guo, & Wu, 2024), feature fusion(Tien, Le, Tomohiro, & Tatsuya, 2019), etc. By analyzing the differences between different features, it is helpful to select appropriate features and apply them to text mining tasks such as topic representation. However, in the current research on text multi-feature, the academic community focuses more on how to

integrate features, and less on the selection of features combination and the influence of their mutual influence.

The word embedding model can convert the text feature words into vectors. It is one of the important steps in text clustering and is widely used in patent text analysis. Existing research starts from the perspective of multiple features, such as high frequency, field characteristics, technical intersections and technological breakthroughs. However, the default word embedding model is often implemented through simple tokenizer, lacks feature selection, and the certainty of the model needs to be improved. Scholars mostly compare and analyze different features from the information and semantic levels and rarely select multiple feature combinations. Therefore, existing research is insufficient in exploring the invisible relationships between different features and has not fully explained the complementary effects and coupling relationships of different features. In terms of the application of multiple features, it focuses on feature fusion but lacks feature selection methods. Therefore, explaining the specific complementary effects and coupling relationships between features, providing a basis for feature selection, and improving the certainty of the word embedding model is a key issue in improving the text clustering effect.

Methodology

Research Framework

Since patent documents are effective carriers of a large amount of world science and technology information, this study conducted research on patent texts. In view of the characteristics of technical themes, four types of patent text feature words, namely, text domain feature keywords (CMFS), technical interdisciplinary keywords (TI), technical breakthrough keywords (KB) and high-frequency keywords (HF), are selected as research objects. To fully explain the problem of complementary effects and coupling relationships between different features, based on the characteristics of patent text with strong technicality, obvious interdisciplinary features, and fast information changes, we combined the semantic and information levels, and selected the similarity, difference, complementarity and synergy between different features as the analysis target. In view of the problem of missing feature selection methods, the Jaccard distance is selected to measure the similarity and difference between features, and the information entropy and mutual information theory are combined to measure the information difference and synergy between features. The information difference and change between different feature words are compared and analyzed, and a comprehensive indicator is designed to select feature combinations. Based on the feature word list, we used Word2Vec, GloVe, and FastText to implement word embedding, and apply K-means, Agglomerative Clustering, and HDBSCAN algorithms to cluster the patent texts. By calculating the silhouette coefficient of each clustering result, we analyze the impact of different text feature combinations on text clustering, thereby verifying the effectiveness and feasibility of this method. The research method framework is shown in Figure 1.

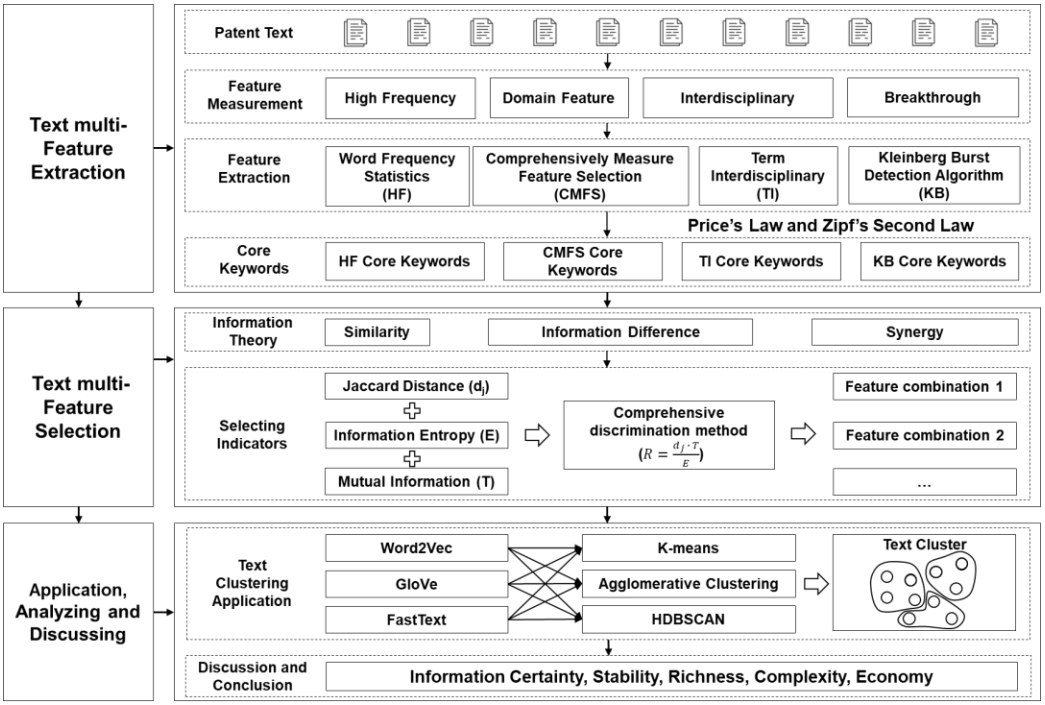

Figure 1. The research framework of clustering method by multi-text feature combination.

Data source and tokenizers

We took the field of graphene sensing technology as an example to carry out experimental research, extracting text domain feature keywords, technical interdisciplinary keywords, technical breakthrough keywords and high-frequency keywords, comparing the topic representation effects of the four types of characteristic keywords and their relationships. The reason for selecting graphene sensing technology for empirical research is that, firstly, there are strong interdisciplinary features in this technology, covering multiple technical fields such as materials, information, and biology; secondly, the breakthrough technology features and active technological innovations in this field are prominent, which have good practical significance for this study.

We selected the Derwent Innovation Index database as the data source. With the assistance of domain experts, we identified patent search strategies that are highly relevant to the topic of graphene sensing technology. The search date is October 31, 2022, and the search strategy is shown in Table 1. There were 974 items obtained after preliminary screening and elimination. Using the Derwent Data Analyzer (DDA) platform to perform NLP word segmentation processing based on the title and abstract text fields of 974 patent records, we obtained 20,036 original n-gram feature words (groups), where n ranges from 1 to 10. Then, the feature words were cleaned, using DDA's built-in stop words list, thesaurus, etc. The cleaning content includes removing common meaning stop words, formatting and grammatical terms of patent documents, DWPI description format abbreviations, compound name

specifications, British and American spelling specifications, etc. After cleaning, 16,604 feature words (groups) were obtained. Finally, manual cleaning is carried out. Personnel skilled in the field conduct manual interpretation, merge synonyms, and eliminate common feature words that are not closely related to substantive research, such as include, use, etc., as well as general experimental tool names, material names, etc. After cleaning, 7873 feature words (groups) were obtained as candidate feature items.

Table 1. The retrieval strategy for Grapheny Sensing Technology.

No.	*Search strategy*
# 1	TS=(sensor* or transducer* or (sensing same (element* or devic* or unit* or organ* or apparatus* or system*)) or (sense same organ*) or Photosensor* or microsensor* or chemosensor* or multisensory* or hypersensor*) database =Cderwent, Ederwent, Mderwent Timespan =2003-2022
# 2	TS=(graphene*) database =Cderwent, Ederwent, Mderwent Timespan =2003-2022
# 3	PN=(US*) database =Cderwent, Ederwent, Mderwent Timespan =2003-2022
# 4	#1 and #2 and #3 database =CDerwent, EDerwent, MDerwent Timespan =2003-2022

Text multi-feature extraction and evaluation

We selected word frequency, CMFS, TI, and KB as the feature extraction indicators, as shown in Table 2. Among them, further combining with the technical features of patent documents, when calculating TI, the IPC classification number is used to measure the technical intersection. At this time, the distribution degree *d* of the feature is the number of technical categories containing the feature, and *tf* is the frequency of the feature.

Table 2. Patent Text multi-feature measurement index.

Target	*Indicator*	*Methods*
High-Frequency Keywords	Word Frequency (Yan, Shuliang, Xiaochao, Yuhui, & Yafei, 2016)	$F = \sum_{i=0}^{N} f_i$ Measure the sum of word frequencies of the word in *N* documents.
Domain Feature Keywords	CMFS (Yang et al., 2012)	$CMFS_{avg}(t_k) = \sum_{i=1}^{\lvert C \rvert} \frac{P(t_k, c_i)(tf(t_k, c_i) + 1)^2}{(tf(t_k) + \lvert C \rvert)(tf(t, c_i) + \lvert V \rvert)}$

Technical Interdisciplinary Keywords	TI (Xu et al., 2016)	$tf(t_k, c_i)$represents the frequency of feature t_k in the i-th category c_i; $tf(t_k)$ represents the frequency of feature t_k in the entire training set; $tf(t, c_i)$ represents the sum of the frequencies of all features in category c_i; $\|C\|$ represents the number of categories; $\|V\|$ represents the number of features in the original vector space. $P(t_k, c_i)$ represents the frequency of feature t_k in the i-th category c_i as a percentage of the frequency of all categories $\|C\|$. $TI = d * \ln(tf)$ d is the degree of the feature words' distribution, and tf is the frequency of the feature words.
Technical Breakthrough Keywords	KB (Kleinberg, 2002)	$\sigma(i, r_t, d_t) = -\ln\left[\binom{d_t}{r_t} p_i^{r_t}(1-p_i)^{d_t - r_t}\right]$ $weight = \sum_{t=t_1}^{t_2} (\sigma(0, r_t, d_t) - \sigma(1, r_t, d_t))$ r_t is the frequency of target i at time t, d_t is the number of events in time t, and p_i is the frequency of the target in events.

Similarity and difference

The Jaccard similarity principle is used to calculate the similarities and differences between the comparison word lists. We used the complementary index of the Jaccard coefficient, the Jaccard distance d_j. The larger the Jaccard distance, the higher the difference between the sets. It is defined as follows(Jaccard, 1912):

$$\mathrm{d}_j(A,B) = 1 - \mathrm{Jaccard}(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (1)$$

To facilitate the quantification of the differences between multiple sets, all two sets that do not repeat are taken from the multiple sets, and their Jaccard coefficients are calculated respectively. Then, the average value of the Jaccard coefficients of all two sets is calculated, which is defined as Equation 2, where n is the number of sets.

$$\mathrm{d}_j(A,B,\ldots,N) = \frac{\sum \mathrm{d}_j(x,y)}{C(n,\ 2)} \quad (x,y \in (A,B,\ldots,N), x \neq y) \quad (2)$$

Information difference and synergy

We also introduced the concept of information entropy and mutual information measurement method, places the feature words in the context of sentences, and quantitatively detects the amount of information revealed by the feature words list and the degree of overlap and fusion between them.

The uncertainty of information corresponds to information entropy. Shannon borrowed the concept of thermodynamics and defined the mathematical expectation of self-information as "information entropy" to measure the amount of information. Combined with linguistic improvements, the probability P_x in the formula is expressed as the relative frequency of a certain feature (that is, the ratio of the feature frequency to the total number of all feature frequencies), and the information entropy calculation formula for measuring the amount of information is obtained as follows (Shannon, 1948):

$$E_x = -\sum_x P_x \lg(P_x) \quad (3)$$

The smaller the information entropy, the more information the text information is concentrated on one or some features; the larger the information entropy, the more information it carries, and the richer or more variable its features are, and the greater its uncertainty. For two-dimensional events, the information entropy E is as follows: Where P_{xy} is the joint probability distribution of event x and event y.

$$E_{xy} = -\sum_x \sum_y P_{xy}\, lg\left(P_{xy}\right) \quad (4)$$

Mutual information is the amount of information about another random variable contained in a random variable, that is, the uncertainty of a random variable reduced by knowing another random variable. It can measure the uncertainty transfer degree between subsystems, that is, the synergy relationship. Abramson (Abramson, 1963) used the mutual information measure of subsystem variables to define the mutual information transfer degree of two interacting subsystems and three interacting subsystems as follows:

$$T_{xy} = E_x + E_y - E_{xy} \quad (5)$$

$$T_{xyz} = E_x + E_y + E_z - E_{xy} - E_{xz} - E_{yz} + E_{xyz} \quad (6)$$

Based on the mutual information measurement theory and application research, we constructed four types of vocabulary synergy. According to the chain rule of mutual information, as follows:

$$T(x_1, x_2, \cdots, x_n; y) = E(x_1, x_2, \cdots, x_n) - E(x_1, x_2, \cdots, x_n|y) \quad (7)$$

Then, specifically for the four vocabularies of HF, CMFS, TI, and KB, their synergy T_{hctk} can be defined as:

$$T_{hctk} = E_h + E_c + E_t + E_k - E_{hc} - E_{ht} - E_{hk} - E_{ct} - E_{ck} - E_{tk} + E_{hct} + E_{hck} + E_{htk} + E_{ctk} - E_{hctk} \quad (8)$$

As a quantitative indicator, T_{hctk} measures the uncertainty of the interaction between the four types of vocabularies, thereby reflecting the degree of information fusion and interaction between the four types of features. T_{hctk} is a positive indicator. The larger the T_{hctk} value, the stronger the interaction and synergy of the four features.

Comprehensive Evaluation

Mutual information is often used for feature selection, and especially has good performance in feature dimensionality reduction(Gandhi & Prabhune, 2017). However, mutual information is difficult to filter out information redundancy. The academic community often combines information entropy to maximize mutual information and minimize information entropy to comprehensively select features, while reducing feature dimensionality, filtering out information redundancy(Liu & Wen, 2023). Therefore, this study calculated the ratio of the two to balance information entropy and mutual information. On this basis, feature combinations with good complementarity and low repetition rate are preferred. Therefore, the Jaccard distance is added to the numerator of the ratio fraction, and the final selection measurement index of the feature combination is obtained as follows.

$$R = \frac{d_j \cdot T}{E} \quad (9)$$

The larger the R is, the more likely it is that the feature combination has higher information certainty, higher information synergy, and better complementarity among the features within this combination compared to any other combination.

Word embedding and text clustering

We used Word2Vec, GloVe, and FastText models to implement word embedding for different feature combinations. Since most pre-trained models would converge after the word vector dimension reaches 300, this study set the word vector dimension to 300 and uses the weighted average of all word vectors in the document (Equation 10) to represent the document. We applied three algorithms: K-means, Hierarchical Clustering (Agglomerative Clustering), and HDBSCAN to cluster patent texts. Through combination, 9 different text clustering models can be obtained. By calculating the silhouette coefficient of each clustering result, the influence of different feature combinations on the text clustering effect is measured. The silhouette coefficient is a clustering performance evaluation index that objectively reflects the outline clarity of each clustering cluster. Its calculation formula is shown in Equation 11(Bagirov, Aliguliyev, & Sultanova, 2023).

$$v_W = \sum v_i * p_i \quad (10)$$

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (11)$$

Among them, v_W represents the vector of document W, v_i represents the vector of feature i, p_i is the frequency of feature i in document W, a_i is the average distance

between each data point i and all other points in the same cluster, and b_i is the average distance between each data point i and all points in the nearest non-self-cluster. The value range of the silhouette coefficient s_i is [-1, 1]. Close to 1 means that the data point is very similar to other points in its own cluster and has obvious differences from data points in other clusters, and the clustering effect is good; while close to 0 means that the data point is on the boundary of two clusters, and the clustering effect is average; close to -1 means that the data point may be mistakenly assigned to the wrong cluster, and the clustering effect is poor.

Results

Text multi-feature extraction result

For 7873 feature words (groups), the Comprehensively Measure Feature Selection (CMFS), Term Interdisciplinary index (TI), Kleinberg burst detection algorithm, and word frequency statistics were used to measure four types of text features. We extracted four types of feature words through programming. According to the measurement results of the feature value of each feature word, the CMFS keywords, TI keywords, KB keywords, and HF keywords in the field of graphene sensing were obtained. Taking the top 20 words as an example, the results of four types of feature-word lists are shown in Table 3.

Next, it is necessary to determine the effective threshold for each feature value, in order to select the appropriate amounts of core keywords of HF, CMFS, TI, and KB respectively, and form a thesaurus with effective topic representation meaning. For threshold determination, this study applied the ideas of Price's law and Zipf's second law. Price's formula was first used to determine highly cited literature and then determine the core authors in a certain research field. It is a scientific method for selecting thresholds and has gradually been applied by scholars in different research fields. Here, we used Price's formula to determine the threshold for core keywords, with the independent variable N_{max} representing the maximum value of the keyword's frequency, TI and KB, to obtain the core keywords threshold for each word list. The calculation formula is as follows (Price, 1963):

Table 3. The extraction results of keywords of HF, CMFS, TI and KB in graphene sensing field (TOP 20).

NO.	HF Keywords	Frequency	CMFS Keywords	CMFS (×10-7)	TI Keywords	TI	KB Keywords	KB
1	method	1380	three dimensional image	2.28	sensor	639.32	method	72.23
2	layer	914	sensor mounting	2.18	method	571.16	surface	50.32
3	sensor	899	conductive membrane	2.13	detecting	489.49	sensor	49.63
4	surface	804	manganese content	2.12	patient	474.78	device	42.80
5	patient	802	toilet seat	2.10	device	467.13	chemistry	39.86
6	device	791	spring structure	2.09	system	442.08	patient	38.28
7	detecting	746	transducer assembly	2.08	surface	441.51	substrate	37.17
8	substrate	671	ultrasound transducer element	2.05	material	425.27	electrode	34.78
9	electrode	660	nano-cone structure polypyrrole	2.05	substrate	410.05	layer	34.57
10	analyte	594	carbon particle	2.02	graphene	394.30	graphene	32.80
11	system	553	enzyme solution	2.01	chemistry	368.09	glucose	29.40
12	sample	531	temperature difference	2.01	layer	361.35	system	27.60
13	glucose	499	carbon nanowalls-based breath sensor	2.00	electrode	318.12	concentration	23.56
14	protein	475	sol-gel silicon film	2.00	solution	310.66	protein	22.03
15	solution	442	lithium ion battery	1.99	polymers	310.57	polymers	21.83
16	material	435	body sensor network	1.99	glucose	291.99	electrodes	19.08
17	graphene	431	elastic container	1.99	concentration	282.80	metal	13.74
18	antibody	412	cervical cancer	1.98	electrodes	270.50	working electrode	12.63
19	binding	397	carbon quantum dot	1.98	polymer	262.96	data	11.90
20	nanoparticles	381	metal hydroxide quantum dots	1.97	protein	234.21	sensitivity	10.35

$$M = 0.749\sqrt{N_{max}} \quad (12)$$

Due to the significant scale difference between the CMFS feature values and the other three types of feature values, the sensitivity of the Price formula in distinguishing the core words of CMFS is poor. So, we applied Zipf's second law to calculate the threshold of CMFS core keywords. The calculation formula is as follows (Donohue, 1973), where I is set to the maximum value of CFMS.

$$T = \frac{1}{2}(-1 + \sqrt{1 + 8 * I}) \quad (13)$$

According to the calculation results of the core keywords thresholds of each words list, the results of four types of core keywords are shown in Table 4.

Table 4. The threshold and number of core keywords of HF, CMFS, TI and KB.

Keywords	*HF*	*CMFS*	*TI*	*KB*
Core Keyword Threshold	27.82	1.69	18.94	6.37
Number of Core Keywords	305	185	608	29

Text multi-feature combination discrimination results

Similarity and difference

The overlap and Jaccard distance between the CMFS, TI, KB core keywords and the HF core keywords are calculated respectively, as shown in Table 5.

Table 5. The differences between core keywords of the CMFS, TI, KB and HF.

Feature	*CMFS vs. HF*	*TI vs. HF*	*KB vs. HF*
Overlaps Number	11	285	29
Overlaps Rate	$\frac{\text{CMFS} \cap \text{HF}}{\text{HF}} = 3.60\%$	$\frac{\text{TI} \cap \text{HF}}{\text{HF}} = 93.44\%$	$\frac{\text{KB} \cap \text{HF}}{\text{HF}} = 9.51\%$
Number of Core Keywords	$\frac{\text{CMFS} \cap \text{HF}}{\text{CMFS}} = 5.95\%$	$\frac{\text{TI} \cap \text{HF}}{\text{TI}} = 46.88\%$	$\frac{\text{KB} \cap \text{HF}}{KB} = 100\%$

The results show that: (1) The overlap rate between the CMFS core keywords and the HF core keywords is the lowest, and the Jaccard distance is the largest, that is, the difference between the CMFS and HF keywords is the largest. This shows that in terms of text feature representation, CMFS core keywords can reveal important thematic features that HF cannot reflect and may play a complementary role for the HF vocabulary. (2) The overlap rate between the TI core keywords and the HF core keywords is high, and the Jaccard distance is relatively small. The technical interdisciplinary has the characteristics of a wide range, but not completely overlap.

This shows that the TI core keywords can effectively identify some low-frequency words with technical interdisciplinary characteristics. (3) The KB core keywords have the least number of words, and all of them belong to HF core keywords, which is consistent with the explosive growth of technology breakthrough in a short period of time. However, its supplementary role in the HF core keywords is of little significance.

Furthermore, the core keywords of CMFS, TI, and KB are compared in pairs, and their overlap rate and Jaccard distance are calculated. As shown in Table 6.

Table 6. The differences between core keywords of CMFS, TI, KB.

Feature	*CMFS vs. TI*	*TI vs. KB*	*CMFS vs. KB*
Overlaps Number	7	29	0
Overlaps Rate	$\frac{\text{CMFS} \cap \text{TI}}{\text{CMFS}} = 3.78\%$	$\frac{\text{TI} \cap \text{KB}}{\text{TI}} = 4.77\%$	$\frac{\text{CMFS} \cap \text{KB}}{\text{CMFS}} = 0\%$
Number of Core Keywords	$\frac{\text{CMFS} \cap \text{TI}}{\text{TI}} = 1.15\%$	$\frac{\text{TI} \cap \text{KB}}{\text{KB}} = 100\%$	$\frac{\text{CMFS} \cap \text{KB}}{\text{KB}} = 0\%$

The results show that: (1) All KB core keywords overlap with the TI core keywords, which displays that most technical breakthrough words may also have technical intersection attributes. (2) The CMFS core keywords do not overlap with the KB core keywords list at all, and the overlap rate with the TI core keywords is very low. (3) The overlap rate of TI core keywords and CMFS is extremely low. Overall, three types of core keywords show good text feature complementarity, especially the CMFS keywords and the TI keywords.

Information difference and synergy

To explore the features at the sentence level, this study segmented the patent document text into sentences. We selected sentences containing HF core words, CMFS core words, TI core words, and KB core words, and classified them into four types of patent text sets. There are 11 different combinations of the four types of features, resulting in 11 types of text sets. The information entropy and mutual information of each text set are calculated as shown in Table 7. Figure 2 intuitively presents the changes in information entropy and mutual information of texts with different feature words.

Table 7. The Information Entropy and Mutual Information of text sets of HF, CMFS, TI and KB.

NO.	*Feature*	*Extraction Rules*	*Sentences Number*	*Information Entropy*	*Mutual Information*
#1	HF	HF core keyword appears in the sentence.	9502	0.072	-
#2	CMFS	CMFS core keyword appears in the sentence.	1215	0.102	-
#3	TI	TI core keyword appears in the sentence.	9930	0.059	-
#4	KB	KB core keyword appears in the sentence.	6182	0.146	-
#5	HF + CMFS	HF, CMFS core keyword appears in the sentence together.	1011	0.092	0.082
#6	HF + TI	HF, TI core keyword appears in the sentence together.	9475	0.073	0.058
#7	HF + KB	HF, KB core keyword appears in the sentence together.	6182	0.146	0.072
#8	CMFS + TI	CMFS, TI core keyword appears in the sentence together.	1041	0.094	0.068
#9	CMFS + KB	CMFS, KB core keyword appears in the sentence together.	614	0.067	0.181
#10	TI + KB	TI, KB core keyword appears in the sentence together.	6182	0.146	0.059
#11	HF + CMFS + TI	HF, CMFS and TI core keyword appears in the sentence together.	1006	0.092	0.067
#12	HF + CMFS + KB	HF, CMFS and KB core keyword appears in the sentence together.	614	0.067	0.082
#13	HF + TI + KB	HF, TI and KB core keyword appears in the sentence together.	6182	0.146	0.058
#14	CMFS + TI + KB	CMFS, TI and KB core keyword appears in the sentence together.	614	0.067	0.068
#15	HF + CMFS + TI + KB	HF, CMFS, TI and KB core keyword appears in the sentence together.	614	0.067	0.067

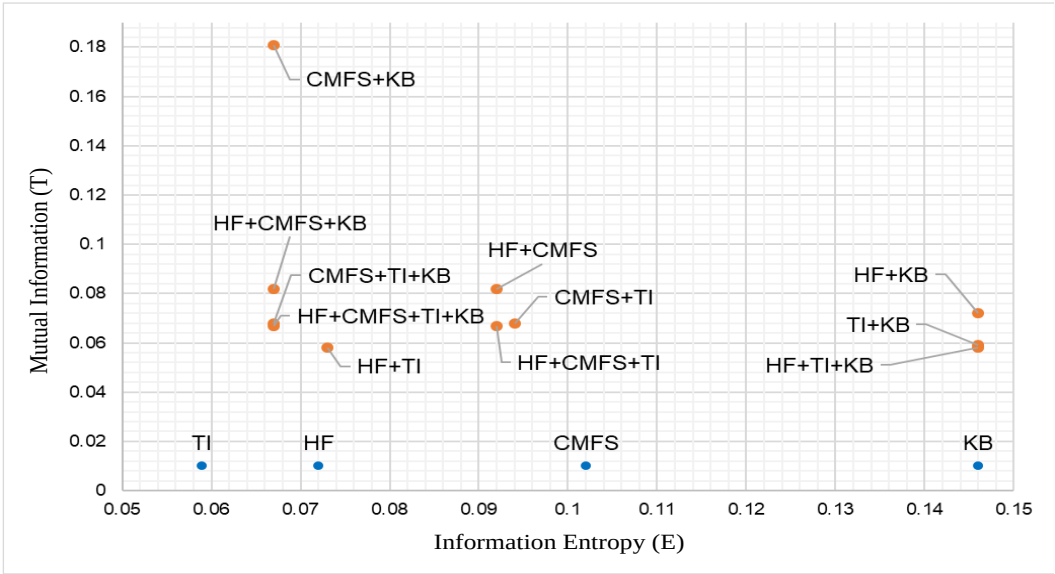

Figure 2. The Information Entropy and Mutual Information of HF, CMFS, TI, and KB texts.

The comparative analysis results of the Information entropy suggest that: (1) The descending order of the number of sentences containing the four types of core keywords is: TI > HF > KB > CMFS. The order of information uncertainty from high to low is KB > CMFS > HF > TI (#1 to #4). (2) The information entropy of TI text (#3) is lowest, while the KB text (#4) is highest, showing that the text feature concentration of the technology interdisciplinary is the highest, and the text feature complexity of the technology breakthrough is the highest. (3) The topic complexity of HF+KB texts is higher than HF texts, while HF+CMFS slightly increases topic complexity, and HF+TI has a smaller change (#1, #5, #6, #7). (4) The information entropy of CMFS text and KB text is relatively high (#2, #4). When either of them is combined with HF or TI features, it can improve the information entropy of the original HF or TI features (#5, #7, #8, #10). When CMFS and KB features appear at the same time (#9), the information entropy of the text decreases significantly. When CMFS+KB features are combined with other features at the same time, the information entropy of the text decreases significantly relative to other features (#12, #14). Therefore, the information uncertainty of CMFS and KB features is high individually, but when they are used simultaneously, the uncertainty is greatly reduced. (5) The four types of texts, namely KB, HF+KB, TI+KB, and HF+TI+KB features texts, have the same number of sentences and information entropy (#4, #7, #10, and #13), which shows that KB features are always accompanied by HF and TI features. This reveals to a certain extent that technological breakthroughs often occur when the development of technology accumulates to a certain extent and intersects. (6) CMFS features significantly reduce the information richness of the HF+TI+KB text (#11, #12, #13, #14, #15).

The comparative analysis results of the mutual information suggest that: (1) the synergy of the CMFS+KB (#9) text is the highest, indicating that there is a certain

information sharing between CMFS and KB features, that is, when one type of feature appears in a sentence, the certainty of the other type of feature will further increase. (2) The mutual information of HF+TI, TI+KB, and HF+TI+KB is relatively small (#6, #10, #13), indicating that the interactivity and synergy of HF, TI, and KB are relatively low. (3) The mutual information of the TI+KB text is almost the smallest and the information entropy is the largest (#10), but the information entropy of the TI text is the lowest (#3) while the information entropy of the KB text is the highest (#4). So, the synergy of the TI and KB features is relatively low. Combined with the quantitative characteristics of TI and KB core keywords (Table 6), the KB core keywords are less in number than the TI core keywords, but the information content is richer. So, the information gain effect of the TI core keywords on the KB core keywords is relatively small.

Taken together, when only a certain type of feature needs to be extracted, HF features and TI features involve rich sentences and the information certainty is highest, so they can be preferred as basic word lists. KB features contain a large amount of information but high levels of uncertainties. Although there are many sentences involved, the number of core words is small. Therefore, KB features can be used in combination with HF features and TI features to enhance the information richness of the latter two. CMFS features can enhance the stability of HF features and TI features. CMFS features and KB features are highly synergistic in text, and their combined use can significantly reduce topic uncertainty. From the perspective of computational complexity, when giving priority to information richness, the most economical choice is HF+KB core words. When giving priority to information certainty, the most economical choice is CMFS+KB core words. When considering a compromise between the two, HF+CMFS+KB core words are a suitable choice

Comprehensive discrimination method

To comprehensively balance the differences, information certainty and information synergy between different features in the feature combination, the comprehensive discrimination index R of each feature combination is calculated. The results are shown in Table 8.

Table 8. The Comprehensive discrimination index of each feature combination.

NO.	*Feature*	*Information Entropy*	*Mutual Information*	*Jaccard distance*	*Comprehensive discrimination*
#5	HF + CMFS	0.092	0.082	0.977	0.871
#6	HF + TI	0.073	0.058	0.546	0.434
#7	HF + KB	0.146	0.072	0.905	0.446
#8	CMFS + TI	0.094	0.068	0.99	0.716
#9	CMFS + KB	0.067	0.181	1	**2.701**
#10	TI + KB	0.146	0.059	0.95	0.384
#11	HF + CMFS + TI	0.092	0.067	0.838	0.610
#12	HF + CMFS + KB	0.067	0.082	0.961	**1.176**
#13	HF + TI + KB	0.146	0.058	0.800	0.318
#14	CMFS + TI + KB	0.067	0.068	0.980	**0.995**
#15	HF + CMFS + TI + KB	0.067	0.067	0.895	0.895

Based on Table 8, the three combinations with the highest comprehensive discrimination indexes are selected, namely, CMFS + KB, HF + CMFS + KB, and CMFS + TI + KB. The number of KB features used alone is small, and it is difficult to obtain valuable information. While CMFS+KB significantly reduces topic uncertainty and provides information supplements for the scarce breakthrough features, the CMFS+KB focuses on the characterization of technology breakthrough features. Since the repetition rate between the KB and the TI and HF is high, the HF+CMFS+KB focuses on the characterization of technology domain features, while the CMFS+TI+KB focuses more on the characterization of interdisciplinary. So far, we have selected three sets of feature combinations with better comprehensive representation capabilities.

Patent text clustering based on multi-feature combination

To further explore the impact of different feature combinations on text clustering, we applied 9 different text clustering models to the three sets of features combinations and calculated the silhouette coefficient of each clustering algorithm as shown in Table 9. This patent dataset covers 8 major IPC categories, so in the algorithm that requires the input of the number of clusters, the default number of clusters is 8. For easy comparison, this study also applied the clustering model to the HF features and used principal component analysis (PCA) to reduce the dimension of each text to 2 dimensions and visualized it as shown in Figure 3.

Table 9. Silhouette coefficients of text clustering based on different feature combinations.

Feature Combinations	*Word Embedding*	*K-means*	*Agglomerative Clustering*	*HDBSCAN*	*Mean*
HF	GloVe	0.066	0.038	0.019	0.041
	Word2Vec	0.161	0.141	-0.059	0.081
	FastText	0.204	0.158	0.089	0.150
	Mean	0.144	0.112	0.016	Overall mean 0.091
HF + CMFS + KB	GloVe	0.071	0.035	0.009	0.038
	Word2Vec	0.166	0.132	0.001	0.100
	FastText	0.203	0.164	**0.107**	**0.158**
	Mean	0.147	0.110	**0.039**	Overall mean **0.099**
CMFS + TI + KB	GloVe	0.064	0.022	0.018	0.035
	Word2Vec	0.160	0.135	0.008	**0.101**
	FastText	0.176	0.152	0.092	0.140
	Mean	0.116	0.082	0.016	Overall mean 0.071
CMFS + KB	GloVe	0.131	0.083	0.010	**0.075**
	Word2Vec	0.178	0.123	-0.133	0.056
	FastText	**0.216**	**0.170**	0.064	0.150
	Mean	**0.175**	**0.125**	-0.020	Overall mean 0.094

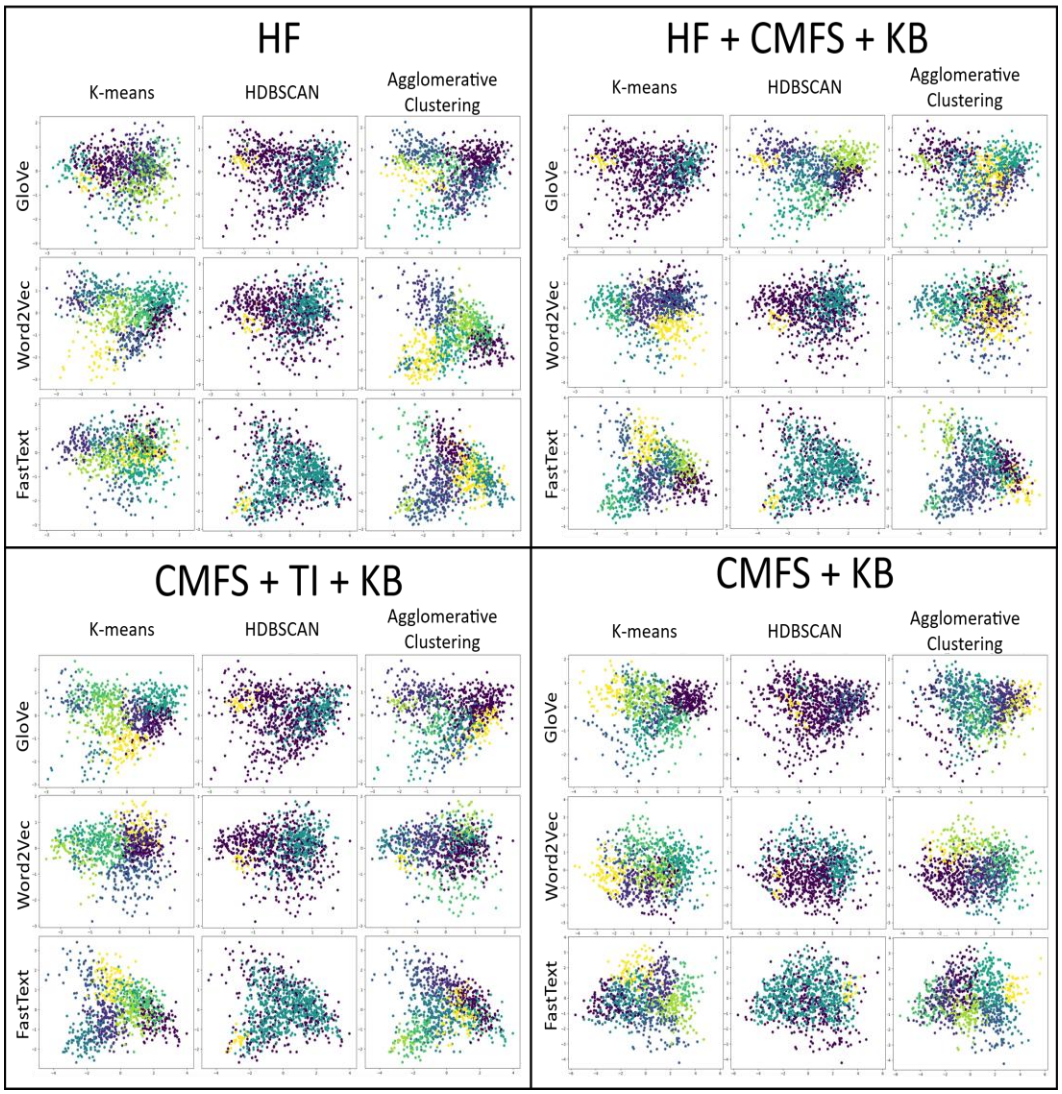

Figure 3. Clustering scatter plots of different feature combinations by PCA.

As shown in Table 9, overall, the overall average silhouette coefficient of feature combinations HF + CMFS + KB and CMFS + KB is greater than HF in the 9 models, which effectively improves the patent text clustering effect. For different word embedding models, HF + CMFS + KB performs best in FastText, CMFS + TI + KB performs best in Word2Vec, and CMFS + KB performs best in GloVe. For different clustering algorithms, HF + CMFS + KB performs best in HDBSCAN, and CMFS + KB performs best in K-means and Agglomerative Clustering.

Combining the comparative analysis results of the differences, information certainty and synergy of different feature combinations, the information certainty, synergy and stability of CMFS + KB feature combination is high. It achieved good results in multiple models and algorithms, especially when used with FsatText+K-means. But for Word2Vec and HDBSCAN, the advantage is not obvious; HF + CMFS + KB slightly improves the information certainty compared to CMFS + KB, but the

synergy decreases significantly. It is more suitable for the specific HDBSCAN+FastText model.

Discussion

Based on the above research and analysis, the main discussions are as follows:

(1) From the perspective of feature morphological differences, the complementarity of three feature words of CMFS, TI, and HF is good in general. Among them, the CMFS features have the best supplementary effect on the HF features and can supplement some features with domain characteristics but without high frequency. The CMFS features do not overlap with the KB features at all, and the overlap between other features is relatively low. The overlap between TI features and the KB or HF features is relatively high, but TI features can effectively identify some non-high frequency and interdisciplinary characteristics, which has a certain supplementary significance for the HF features. KB features completely overlap with HF and TI features, of which have both technical breakthrough and technical interdisciplinary characteristics. It has the weakest supplementary significance for the HF features and can only realize the selection of features with technical breakthrough characteristics from HF features.

(2) From the perspective of text information, the significance of CMFS features in revealing text topic characteristics is stronger than other features. The information concentration of TI features is the highest; the complexity of KB features is the highest. The CMFS and KB features are strongly uncertainty, but when they appear at the same time, the uncertainty of information is greatly reduced. Although HF+KB features are rich in information, they have a high repetition rate and are difficult to supplement the HF features.

(3) From the perspective of the synergy of feature texts, the synergy of CMFS and KB features is highest, and information certainly is better. There is good information sharing between CMFS+KB. For other feature combinations, the mutual information performance is relatively low, but the information certainly exhibits a variety. Among them, HF+CMFS+KB feature combination is a compromise between the synergy and information certainty.

(4) Considering the complementarity, information certainty, and synergy, the combined features of CMFS + KB, HF + CMFS + KB, and CMFS + TI + KB are better in representation.

(5) From the application effect of text clustering, the CMFS + KB is widely applicable in text clustering tasks of various word embedding and clustering algorithms. When used with FsatText + K-Means, the text clustering effect is the best. For the specific FastText + HDBSCAN model, FastTex incorporates word substring information, while HDBSCAN constructs a distance matrix and a directed weighted graph, resulting in higher computation complexity of the model. In this case, HF + CMFS + KB performs better.

In general, compared with the use of HF features alone, the combination of text multi-feature effectively improves the clustering effects. Among them, the CMFS+KB feature combination greatly improves the information certainty, resulting in a good stability of the clustering model to a certain extent, and helps to

identify the target topic more accurately and quickly. For the models with relatively high complexity, the HF + CMFS + KB feature combination may be effective. This study provides a certain basis and support for the selection and combination of vocabulary in text clustering.

Conclusion

Studying the limitation of insufficient quantitative measurement of the meaning of text multi-features in current text mining, we extracted four types of text features from patent texts, including domain feature, technical interdisciplinary, technical breakthrough and high frequency. Combined with the characteristics of patent texts, such as strong technicality, obvious interdisciplinary, and fast information changes, the similarity, difference, complementarity, and synergy between different thematic features are selected as analysis targets. Employing the Jaccard distance combined with the information entropy and mutual information theory, a method for comparative analysis of information differences and changes between different features is designed. A comprehensive discrimination index for feature combination selection is proposed. Using the Jaccard distance, information entropy, and mutual information index, a quantitative comprehensive measurement of four text features representation meaning is carried out. Taking the field of graphene sensor technology as an example, the similarity, difference, synergy, and complementarity of the four text features recognition are compared. Through comprehensive discrimination indicators, three representative feature combinations are selected, and they are applied in combination with a variety of word embedding models and clustering algorithms. The research results show that compared with simple high-frequency features, text multi-features can effectively play a complementary role from different perspectives. Selecting different feature combinations for use can reduce the uncertainty of text information, enhance the richness of text information, and improve the stability of text information to a certain extent. In addition, empirical analysis based on graphene sensing technology also provides optimization inspiration for parameter fitting and feature training of various language or topic models, thereby improving the recognition accuracy of unique feature technologies. This method can meet the needs of different analysis purposes and application scenarios in actual applications and help improve the efficiency and accuracy of technological innovation research.

Limitations

This study also has certain limitations. Since the research focus is on the quantitative and selection measurement of text multi-features, the extraction methods for the features are not rich enough, which may affect the richness of the text features. In the future, we can further enrich the measurement and extraction of different text features, explore more nonlinear relationships between different text features.

Acknowledgments

This contribution is the outcome of the projects, “Research on Multiple Relationship Fusion Methods for Identification and Prediction of Technological Innovation Paths” (No.18BTQ067) supported by the National Social Science Fund of China, “Early Recognition Method of Transformative Scientific and Technological Innovation Topics based on Weak Signal Temporal Network Evolution analysis” (No.72274113) supported by the National Natural Science Foundation of China, “Youth Innovation Promotion Association (2022173)” supported by Chinese Academy of Sciences(CAS), and the Taishan Scholar Foundation of Shandong province of China (tsqn202103069 and tsqn202103070).

References

Abramson, N. (1963). Information theory and coding (First Edition.). New York: McGraw Hill.

Bagirov, A. M., Aliguliyev, R. M., & Sultanova, N. (2023). Finding compact and well-separated clusters: Clustering using silhouette coefficients. Pattern Recognition, 135, 109144.

Borah, A., Barman, M. P., & Awekar, A. (2021). Are word embedding methods stable and should we care about it? Proceedings of the 32nd ACM Conference on Hypertext and Social Media, HT ’21 (pp. 45–55). New York, NY, USA: Association for Computing Machinery. Retrieved January 12, 2025, from https://doi.org/10.1145/3465336.3475098

Büyükkeçeci, M., & Okur, M. C. (2023). A comprehensive review of feature selection and feature selection stability in machine learning. Gazi University Journal of Science, 36(4), 1506–1520.

Chang, Y.-W., & Huang, M.-H. (2012). A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. Journal of the American Society for Information Science and Technology, 63(1), 22–33.

Chawla, S., Kaur, R., & Aggarwal, P. (2023). Text classification framework for short text based on TFIDF-FastText. Multimedia Tools and Applications, 82(26), 40167–40180.

Chen G., Xu Z., Hong S., Wu J., & Xiao L. (2024). A Study on the Stability of Semantic Representation of Entities in the Technology Domain-Comparison of Multiple Word Embedding Models. Journal of the China Society for Scientific and Technical Information, 43(12), 1440–1452.

Chen, K., & Chiung-fang, L. (2004). Disciplinary interflow of library and information science in taiwan. Journal of Library and Information Studies, 2.

Cheng Yong, Xu Dekuan, & Lv Xueqiang. (2019). Automatically grading text difficulty with multiple features. Data Analysis and Knowledge Discovery, 3(7), 103–112.

Chi, J., Ouyang, J., Li, C., Dong, X., Li, X., & Wang, X. (2019). Topic representation: Finding more representative words in topic models. Pattern Recognition Letters, 123, 53–60.

Choi, H., Oh, S., Choi, S., & Yoon, J. (2018). Innovation topic analysis of technology: The case of augmented reality patents. IEEE Access, 6, 16119–16137. Presented at the IEEE Access.

Donohue, J. C. (1973). Understanding scientific literatures: A bibliometric approach. The MIT Press, 28 Carleton Street, Cambridge, Massachusetts 02142 ($12.

Enguix, F., Carrascosa, C., & Rincon, J. (2024). Exploring federated learning tendencies using a semantic keyword clustering approach. Information, 15(7), 379.

Ercan, G., & Cicekli, I. (2016). Topic segmentation using word-level semantic relatedness functions. Journal of Information Science, 42(5), 597–608.

Feng, G., Wu, J., & Mo, X. (2020). Research on detection and verification of burst words with multiple measures. Library and Information Service, 64(11), 67–76.

Feng Guohe & Kong Yongxin. (2020). Subject hotspot research based onWord frequency analysis of time-weighted keywords. Journal of the China Society for Scientific and Technical Information, 39(1), 100–110.

Gandhi, S. S., & Prabhune, S. S. (2017). Overview of feature subset selection algorithm for high dimensional data. 2017 International Conference on Inventive Systems and Control (ICISC) (pp. 1–6). Presented at the 2017 International Conference on Inventive Systems and Control (ICISC). Retrieved January 9, 2025, from https://ieeexplore.ieee.org/document/8068599

Inje, B., Nagwanshi, K. K., & Rambola, R. K. (2024). An efficient document information retrieval using hybrid global search optimization algorithm with density based clustering technique. Cluster Computing, 27(1), 689–705.

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. New Phytologist, 11(2), 37–50.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. Journal of Memory and Language, 87, 38–58. San Diego: Academic Press Inc Elsevier Science.

Ji D., Liu Y., Peng R., & Kong H. (2024). K-means text clustering algorithm based on the center point of subject word vector. Computer Applications and Software, 41(10), 282–286, 318.

Jia, W., Xie, Y., Zhao, Y., Yao, K., Shi, H., & Chong, D. (2021). Research on disruptive technology recognition of China's electronic information and communication industry based on patent influence. Journal of Global Information Management (JGIM), 29(2), 148–165. IGI Global.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02 (pp. 91–101). New York, NY, USA: Association for Computing Machinery. Retrieved January 14, 2024, from https://doi.org/10.1145/775047.775061

Kruczek, J., Kruczek, P., & Kuta, M. (2020). Are n-gram categories helpful in text classification? In V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, & J. Teixeira (Eds.), Computational Science – ICCS 2020 (Vol. 12138, pp. 524–537). Presented at the 20th Annual International Conference on Computational Science (ICCS), Cham: Springer International Publishing. Retrieved September 16, 2024, from https://link.springer.com/chapter/10.1007/978-3-030-50417-5_39

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018, June 13). Diachronic word embeddings and semantic shifts: A survey. arXiv. Retrieved January 13, 2025, from http://arxiv.org/abs/1806.03537

Li, X., Zhang, A., Li, C., Ouyang, J., & Cai, Y. (2018). Exploring coherent topics by topic modeling with term weighting. Information Processing & Management, 54(6), 1345–1358.

Liu Y., & Wen Y. (2023). Feature Selection Based on Maximizing Joint Mutual Information and Minimizing Joint Entropy. Advances in Applied Mathematics, 12, 1451.

Liu Yahui, Xu Haiyun, Wu Huawei, Liu Chunjiang, & Wang Haiyan. (2023). Scientific breakthrough topics identification in an early stage using multiple weak linkage fusion. Journal of the China Society for Scientific and Technical Information, 42(1), 19–30.

Mengle, S. S. R., & Goharian, N. (2009). Ambiguity measure feature-selection algorithm. Journal of the American Society for Information Science and Technology, 60(5), 1037–1050.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 3, 299–313.

Porter, A., & Chubin, D. (1985). An indicator of cross-disciplinary research. SCIENTOMETRICS, 8(3–4), 161–176. Amsterdam: Elsevier Science Bv.

Prabowo, R., & Thelwall, M. (2006). A comparison of feature selection methods for an evolving RSS feed corpus. Information Processing & Management, Special Issue on Informetrics, 42(6), 1491–1512.

Price, D. (1963). Little science, big science. Columbia University Press.

Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181.

Rettenmeier, L. (2020, July 23). Word embeddings: Stability and semantic change. arXiv. Retrieved January 13, 2025, from http://arxiv.org/abs/2007.16006

Salton, G., Allan, J., & Singhal, A. (1996). Automatic text decomposition and structuring. Information Processing & Management, 32(2), 127–138.

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423. Presented at the The Bell System Technical Journal.

Tien, N. H., Le, N. M., Tomohiro, Y., & Tatsuya, I. (2019). Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity. Information Processing & Management, 56(6), 102090.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. Information Processing & Management, Patent Processing, 43(5), 1216–1247.

Vicente-Gomila, J. M., Artacho-Ramirez, M. A., Ting, M., & Porter, A. L. (2021). Combining tech mining and semantic TRIZ for technology assessment: Dye-sensitized solar cell as a case. Technological Forecasting and Social Change, 169, 120826.

Wang X., Lu L., & Tai W. (2015). Research of a New Algorithm of Words Similarity Based on Information Entropy. Computer Technology and Development, 25(9), 119–122.

Xu, H., Guo, T., Yue, Z., Ru, L., & Fang, S. (2016). Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series. SCIENTOMETRICS, 106(2), 583–601.

Yan, L. U. O., Shuliang, Z., Xiaochao, L. I., Yuhui, H. a. N., & Yafei, D. (2016). Text keyword extraction method based on word frequency statistics. Journal of Computer Applications, 36(3), 718.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Information Processing & Management, 48(4), 741–754.

Yao, R., Wang, J., Wu, J., Xu, Q., & Zhang. (2023). Research on potential interdisciplinary topic identification method. Library and Information Service, 67(15), 80–93.

Yu, Y., Ju, P., & Shang, M. (2022). Research on the evaluation method of patent keyword extraction algorithm based on information gain and similarity. Library and Information Service, 66(6), 108–117.

Yu Yan & Zhao Naixuan. (2018). Weighted topic model for patent text analysis. Data Analysis and Knowledge Discovery, 2(4), 81–89.

Zhang, L., Sun, B., Chinchilla-Rodríguez, Z., Chen, L., & Huang, Y. (2018). Interdisciplinarity and collaboration: On the relationship between disciplinary diversity in departmental affiliations and reference lists. Scientometrics, 117(1), 271–291.

Zhao, M., Guo, J., & Wu, X. (2024). A large group emergency decision-making approach on HFLTS with public preference data mining. Journal of Global Information Management (JGIM), 32(1), 1–22.