https://doi.org/10.51408/issi2025_040

Research and Application on Multiple Topic Association Fusion Method Based on Neural Network and Evidence Theory

Shuying Li[1], Xian Zhang[2], Jiahui Li[3], Xin Zhang[4], Haiyun Xu[5]

[1] *lisy@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[2] *zhangx@clas.ac.cn,*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[3] *Lijh@las.mail.ac.cn*
University of Chinese Academy of Sciences (UCAS), Beijing 321000 (China)

[4] *zhangx@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[5] *xuhaiyunnemo@gmail.com*
Shandong University of Technology, Zibo City, 255000 (China)

Abstract

This paper contributes to the field of multivariate theme association analysis by proposing a novel data fusion method for patent text theme analysis. The method leverages multiple theme data association features from patent text mining. The methodology of the study involves the extraction of three thematic correlations: term co-occurrence, citation- term coupling, and patent assignee-term coupling. Corresponding matrixes are then constructed, thereby facilitating the analysis of the data. A neural network and evidence-based fusion method is then developed to generate matrixes that are enhanced with information and integrate multi-source uncertainties. Empirical validation using graphene sensing patents demonstrates the method's effectiveness, revealing complementary thematic features and enhanced information richness in fused matrix. The results reveal significant differences among the three types of thematic correlations, highlighting their complementary nature in revealing thematic features. The fused matrices exhibit enhanced information richness and reduced dispersion, effectively capturing both dominant and rare thematic associations. This study underscores the potential of the proposed method to provide comprehensive and precise tools for thematic analysis in patent texts.

Introduction

In the domain of information fusion and knowledge mining, research on multi-topic correlation fusion methods is of considerable theoretical importance and practical application value. With the rapid development of information technology and the increasing demand for multi-source data integration, the efficient extraction, integration, and analysis of multi-topic correlation information from vast amounts of patent literature have become critical for driving technological innovation, guiding industrial upgrading, and predicting technological trends. In light of the rapid advancements in big data and artificial intelligence technologies, patent literature analysis is undergoing a transformation toward greater complexity and systematization, signifying a significant shift in the research paradigm.

The core of multi-topic correlation fusion identification methods lies in the integration of information from different topic correlations and multiple objectives. This method enables the synthesis of evidence from multiple uncertain information sources to construct an information enhancement matrix that is rich in topic correlations. Consequently, it effectively addresses the limitations of single-relationship types and more accurately reflects the similarity between topics. The fundamental principle underlying this approach is the extraction of consistent information from multi-topic correlations, with the objective of addressing uncertainties caused by various factors, including domain-specific topic terms, cross-topic terms, emerging topic terms, and high-frequency topic terms. Evidence theory, as an efficient information fusion technology, has the ability to clearly distinguish between unknown and uncertain information, and is able to realize the deep integration of information in multiple dimensions(Xiao, 2023). Its powerful fault-tolerant mechanism provides a solid theoretical foundation for the information fusion process(Pan et al., 2021). However, when faced with conflicting evidence, the theory of evidence may encounter limitations, which in turn affects the accuracy of the final judgment(Hamda et al., 2023). In contrast, neural network algorithms demonstrate considerable advantages in the domain of multi-topic relevance fusion recognition due to their superior fault-tolerance performance, efficient hierarchical processing capability, powerful self-learning ability, flexible adaptability, and efficient parallel processing capability. These properties position neural network algorithms as a potent instrument for addressing complex information fusion problems.

In light of the aforementioned analysis, this study proposes a novel integration of evidence theory and neural network algorithm, with the objective of enhancing the processing of relational data in patent documents. The proposed framework encompasses a multifaceted approach, addressing critical aspects such as data source processing, feature-level fusion, decision-level fusion, and data-level fusion. A comprehensive evaluation of conflicting evidence information is achieved through the framework's application, resulting in several notable optimizations. Primarily, the framework enhances the weight allocation for correctly identified evidence. Secondly, it effectively reduces the impact of ambiguous evidence and outlier evidence deviating from the overall level. Finally, it significantly improves the precision and reliability of the system. This innovative approach provides a new technical pathway for patent analysis and offers valuable insights for research in the field of multiple topic association fusion.

Literature Review

In patent analysis, the multi-topic correlation fusion method has shown a wide range of application potential. For example, in technological innovation assessment, this method can accurately identify key technologies and core patents (Huailan Liu et al., 2022); in industrial competition analysis, it helps to reveal competitors' technology layout and market strategies (Song et al., 2023); in policy making, it provides the government with scientific technology trend forecasts and industrial development suggestions (Yan et al., 2024). In the future, with the continuous development of technologies such as big data and artificial intelligence, the application of multi-topic correlation fusion methods in patent document analysis will be more in-depth and extensive.

Multi-Source Information Fusion

Multi-Source Information Fusion (MSIF) is a comprehensive interdisciplinary field involving multiple disciplines and technologies. In recent years, MSIF has made significant progress in theory and application, but it still faces some key issues and challenges, such as information processing and fusion system design, fusion model and method classification, etc. The application fields of MSIF include but are not limited to military, meteorology, medical, transportation, etc. Tan et al.(2022) designed a multi-source fusion positioning and navigation algorithm based on adaptive filters to integrate the advantages of multiple sensors and provide high-precision and high-reliability positioning and navigation services. Zhu et al.(2024) fused the rolling multi-source heterogeneous information of wind turbines and combined it with the improved PCR6 method to enhance the recognition performance of Rolling Bearing Fault Diagnosis. Li et al.(2024) introduced a multi-source object association method in the radar camera fusion scheme, which significantly improved the vehicle detection accuracy under various adverse conditions and achieved accurate traffic parameter estimation.

In the field of information science, multi-source information fusion can make full use of different information features and internal relationships to achieve information dimension reduction, information integration, information unification, and reduce information uncertainty. Zhang and Lin (2025) proposed a data fusion hierarchical framework that adapts to multi-source and multi-scale schemes, using information gain to aggregate heterogeneous data sources and refine data sets, improving the robustness and effectiveness in processing complex multi-source and multi-scale data environments. Qian et al.(2023) used a variety of generalized multi-granularity rough set models to fuse and utilize multi-source information from multiple perspectives, and adaptively obtained the threshold pairs corresponding to the knowledge granularity through a parameter compensation coefficient, making the model more flexible in practical applications and making decisions more reasonable. Lin et al.(2025)used the information fusion enhanced domain adaptive attention network (IF-EDAAN) to reduce potential feature conflicts, and achieved effective extraction and alignment of temporal and spatial features without domain invariance to improve the efficiency of metastasis diagnosis.

Multivariate Relationship Fusion

In patent analysis, multi-source information may come from different information subjects, and the various multi-relationships between different subjects complement a single relationship. Therefore, patent multi-source information fusion focuses on multi-relationship fusion. The multi-relationship fusion method extracts and integrates multiple related relationships in patent documents, such as subject co-occurrence relationships, citation relationships, patent owner cooperation relationships and other multi-relationship information, to achieve in-depth identification and analysis of patent themes. This method not only helps to reveal the inherent structure and development context of the technology field but also predicts the emergence of technology fusion and emerging fields. For example, Zhang X. et al.(2024) integrated citation connection relationships, subject association relationships and citation motivation relationships, and proposed a main path identification method for multi-relationship fusion, which effectively improved the identification effect of technology evolution paths in the empirical field. Liu et al.(2024) constructed correlation indicators based on the subject citation relationships, subject relationships, content relationships, and cross-relationships between papers, patents and products, thereby identifying the evolution path of quantum communication technology. In addition, the multi-resource integration model based on the theme graph provides a new perspective and method for the visualization and in-depth mining of patent information. For example, Liu et al.(2022) constructed a hierarchical interactive multi-channel graph neural network based on four relationships: high-order interactions, co-occurrence, hierarchy, and technical knowledge flow to achieve technical knowledge flow prediction. Zhai et al.(2023) constructed a knowledge graph of traditional Chinese medicine based on multi-source heterogeneous data, and used deep learning information, string matching, frequency analysis, association rule Apriori algorithm, etc. to assist researchers in conducting innovative research in the field of traditional Chinese medicine.

Application of D-S Evidence Theory

Commonly used multi-source information fusion methods include classical rough set theory, multi-granularity method, evidence theory and information entropy (Xu et al., 2023). Among them, Dempster-Shafer reasoning (D-S evidence theory) has a strong advantage in processing uncertain information. Zhang et al.(2025) introduced the support matrix based on Dempster-Shafer evidence theory, combined with the hierarchical fusion method, and conducted in-depth research on the information fusion strategy of large-scale multi-source data, and verified that the method is both efficient and effective, and has shown excellent performance in information fusion. Li et al.(2024) realized the information fusion of multi-source incomplete mixed data based on conditional information entropy and DS evidence theory, thereby improving the performance of the classification algorithm. Zhang et al.(2024) proposed a new data enhancement method based on hybrid and Dempster-Shafer reasoning, combined with training deep neural networks to complete recognition or classification tasks, to achieve more effective data enhancement effects and further improve the performance of deep neural networks.

In existing research, the DS method focuses on application scenarios such as intelligent decision-making, while the application of information fusion with patent documents is relatively rare. Patent documents often contain ambiguous, incomplete or contradictory information, and the wide applicability and good robustness of the DS method make it unique in patent document analysis. The D-S evidence theory quantifies this uncertain information through Basic Belief Assignment (BBA) and uses combination rules to achieve effective fusion of multi-source information. This helps to accurately identify technology trends, evaluate technology maturity, and predict potential technology breakthroughs.

Existing research has primarily applied DS approaches to scenarios such as intelligent decision-making. Nevertheless, the utilization of DS approaches for patent semantic fusion remains comparatively limited. The primary research gaps and shortcomings in this domain can be categorized into three aspects: first, the modeling capability of dynamic topic association is limited. The majority of current methodologies are confined to static topic associations, failing to incorporate effective modeling tools for dynamic topic associations that undergo changes over time (e.g., selection of feature words, adjustment of weights, evolution, etc.). The existing methods demonstrate clear limitations when it comes to capturing and predicting dynamic associations. Second, the extraction of semantic information remains inadequate. Most extant research relies on statistical features or shallow semantic information (e.g., term frequency, co-occurrence relationship, etc.), while the ability to mine deep semantic associations (e.g., semantic similarity of technical concepts, citation text coupling associations, patentee text coupling, etc.) is limited. This may result in the exclusion of significant semantic information during the process of topic association fusion, consequently impacting the precision of the fusion outcomes. In addition, the robustness of evidence conflict processing requires enhancement. Patent text information often contains ambiguous, incomplete, or contradictory content, which further complicates the assessment of evidence quality and the fusion of heterogeneous data. The integration of evidence theory and neural network methodologies has been demonstrated to enhance the balance between the robustness of evidence processing and the accuracy of fusion outcomes. However, these prevailing methodologies continue to fall short in this regard. D-S evidence theory quantifies such uncertain information through Basic Belief Assignment (BBA) and utilizes combinatorial rules to achieve effective fusion of multi-source information. Nevertheless, further enhancement of its robustness remains necessary when addressing conflicts among evidence sources.

In addressing the aforementioned deficiencies, the present study has developed a fusion processing framework for relational data found in patent documents, which integrates neural networks with evidence theory. The framework addresses the limitations of existing methods by enhancing the robustness of evidence conflict processing through the reinforcement of weight assignment to correctly identified evidence. Additionally, the framework improves processing capability for heterogeneous data from multiple sources through multi-level fusion modules. Nonetheless, there is a necessity for further exploration and improvement in dynamic topic modeling, cross-domain adaptation, and deep semantic mining. Future research directions could include the integration of techniques such as graph neural networks and knowledge graphs. This integration has the potential to further enhance the performance and applicability of the multivariate topic association fusion method.

Methodology

Research Framework

This research introduces a novel methodology that integrates neural networks with evidence theory to systematically analyze the intricate relationships among multiple topic associations in patent documents. The proposed approach systematically incorporates three critical data dimensions to holistically capture the intricate relationships among topics. Specifically, the analysis encompasses three aspects: (1) the co-occurrence of subject terms, (2) the coupling relationships between citations and subject terms, and (3) the coupling relationships between patent assignees and subject terms. This multidimensional framework effectively overcomes the inherent limitations of traditional co-occurrence analysis, particularly its susceptibility to loss of information, while simultaneously mitigating the identification inaccuracies that frequently arise from irregular citation patterns in conventional citation-based analysis methods.

The proposed methodology establishes a comprehensive data framework (Fig. 1) through the construction of three distinct types of subject association relationships: subject term co-occurrence, citation-subject term coupling, and patent applicant-subject term coupling, along with their corresponding association matrices. The fusion process employs a weighted allocation strategy that prioritizes evidence information with high reliability while reducing the influence of ambiguous or biased evidence. This approach achieves integration across three levels: feature-level, decision-level, and dataset-level. The result is a robust multivariate topic association fusion model.

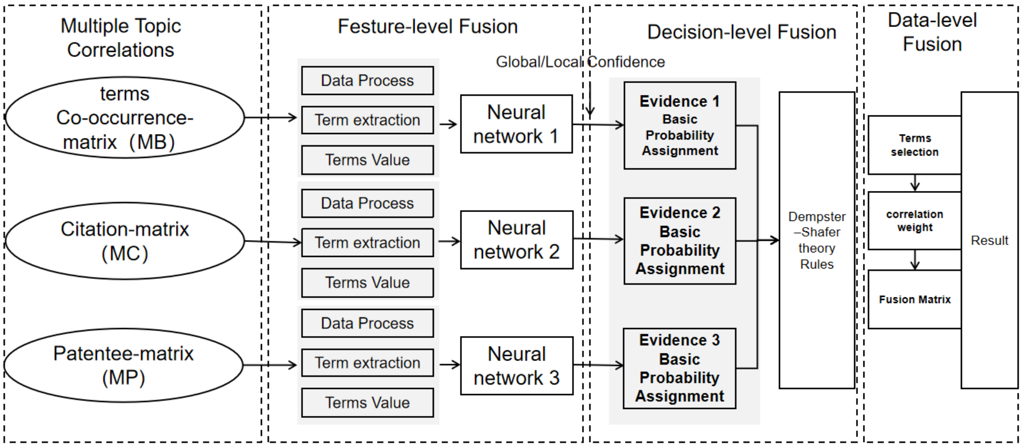

Figure 1. A Neural Network and Evidence Theory-Based Framework for Multi-Topic Association Relationship Integration.

To validate the effectiveness of the proposed method, an empirical study was conducted within the graphene sensing domain as a representative research context. The study successfully demonstrated the integration of the three types of topic association matrices, leveraging the adaptive learning capabilities of neural networks to refine the fusion process. The incorporation of credibility from diverse evidence sources and the optimization of the fusion mechanism are critical aspects of the proposed method. This framework offers a novel and effective approach to patent text analysis, addressing the limitations of traditional approaches and the complexities of multi-source data integration.

Multiple Topic correlations

Multiple Topic Association Relationships aims to explore in depth how topic terms in patent texts form multiple semantic connections with other measured entities (e.g., patent applicants, citations, and so on). By integrating these different types of associations, we can improve the accuracy and richness of the topic identification process in patent texts. Considering the uniqueness of patent technology innovation activities and patent text characteristics, this study focuses on analyzing diverse subject association relationships in patent texts.

Based on the synergy and inheritance between the subject (e.g., patent applicant), the object (e.g., patent documents), and their characteristics, we have identified three core types of thematic associations by utilizing the information of subject term co-occurrence, citation, and patent cooperation application in patent documents: subject term co-occurrence relationship, citation-subject term coupling relationship, and patent applicant-subject term coupling relationship. The specific definitions of these three relationships are as follows:

(1) Basic association: Terms Co-occurrence Matrix (MB)

This refers to the relationship in which the subject term Ti and the subject term Tj directly co-occur in the same patent document Pm. It reflects the most direct semantic proximity between the subject terms and thus forms the basis for the fusion of multiple relationships in this study.

(2) Extended association: Citation-terms Coupling Matrix (MC)

This relationship indicates that the subject term Ti and the subject term Tj, although distributed in different patent documents, have formed an enhanced association because they are jointly cited by a cited document Ci. The citation of Ci strengthens the association between subject matter Ti and Tj.

(3) Additional Association: Patent Assignee-terms Coupling Matrix (MP)

This refers to the fact that although subject matter Ti and subject matter Tj do not appear in the same patent document or do not have a common patent applicant, a new association path has been formed due to the existence of a cooperative application for patent Pm by their respective corresponding patent applicants Ai and Aj. This relationship connects the originally independent subject terms Ti and Tj through patent applicant Ai, patent document Pm, and patent applicant Aj.

Feature-level integration of multiple thematic correlations

The utilization of feedback mechanisms inherent in neural network algorithms facilitates the implementation of feature-level fusion through the adjustment of weights. This process entails the consideration of diverse combinations of various features, thereby ensuring the effective integration of multi-topic relational data. The fundamental principle underpinning this process is the continuous adjustment of

weights and thresholds within the network through the backward propagation of errors, a process that continues until the sum of squared errors at the output layer of the network is minimized. In this study, the BP neural network model was adopted to preprocess three types of multi-topic relational data, from which representative feature vectors were extracted as inputs for the BP neural network. Based on previous research, four target classifications were identified, namely, domain-specific terms, technology cross-over terms, technology burst terms, and high-frequency terms. Subsequently, eight measurement indicators were selected as the feature values to be fused for these four target classifications, including High-Frequency (HF) (Qaiser and Ali, 2018; Tseng et al., 2007), Term Frequency-inverse Document Frequency (TFIDF) (Chawla et al., 2023), Comprehensively Measure Feature Selection (CMFS) (Yang et al., 2012), Information Gain (IG) (Yu et al., 2022), Term Interdisciplinary index (TI) (Xu et al., 2016), Shannon-Wiener Index (SWI) (Shannon, 1948), Kleinberg burst (KB) (Kleinberg, 2002), and growth rate (GR) (Feng et al., 2020).
A BP neural network model (Fig.2) is constructed for each topic association relationship. The strong generalization and nonlinear mapping capability of the neural network algorithm enables the association of multiple eigenvalues, thereby facilitating the identification and classification of the target type by each topic-associated relationship. The output of the neural network can be utilized as evidence of the efficient utilization of multiple eigenvector changes.

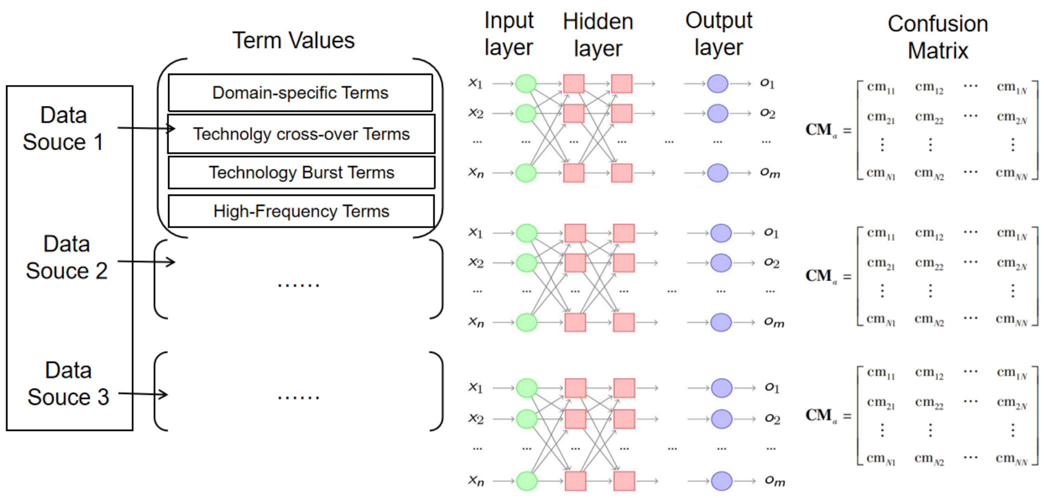

Figure 2. BP neural network model.

The confusion matrix, a statistical tool used for understanding and interpreting data, is a crucial component of this analysis. It delineates the ability of each subject association to recognize the target, and consequently, the global and local credibility of each subject association is calculated. The local credibility is weighted and fused with a posteriori probability output to construct the basic probability distribution function, providing a comprehensive framework for understanding the relationship between credibility and prediction accuracy.

The Confusion Matrix (CM) developed from the BP neural network model classification indicates that the recognition capability of each association relationship between topics varies. These associations include domain feature topic words, technology cross-topic words, technology burst topic words, and high frequency topic words, among others. Theoretically, topic word co-occurrence, serving as the base relationship, exhibits a potential enhancement in recognition performance for both domain feature topics and high frequency feature topics. Similarly, citation-topic word coupling, operating as the reinforcement relationship, is predicted to demonstrate an improved recognition capability for technology burst features. Furthermore, patentee-topic word coupling, functioning as the additional relationship, is expected to show enhanced performance in recognizing technology burst features. It is conceivable that the citation-topic word coupling as a reinforcement relationship would be more efficacious for recognizing technical cross features; alternatively, the patentee-topic word coupling as an additional relationship may be more effective for recognizing technical emergent features. The confusion matrix includes T samples, each containing N distinct target types, with a sample count of Ti (i=1,2,…,N) for each target. The formula is as follows:

$$CM_a = \begin{Bmatrix} \mathrm{cm}_{11} & \mathrm{cm}_{12} & \mathrm{cm}_{1N} \\ \vdots & \vdots & \vdots \\ \mathrm{cm}_{N1} & \mathrm{cm}_{N2} & \mathrm{cm}_{NN} \end{Bmatrix} \qquad (1)$$

Where *a* is the number of neural networks; the row subscripts of *cm* in the set of confusion matrices are the true target types; the column subscripts are the target types recognized by the neural network, representing the proportion of the number of samples with target type *i* recognized by the neural network as type *j* to the proportion of samples of type *i*; and the diagonal elements are the percentage of elements of each target type that can be correctly recognized by the neural network.

The BP neural network is employed to evaluate the target recognition classification ability of multivariate topic association by constructing a realistic basic probability assignment function. To this end, it is imperative to calculate the probability that the test sample of class i is classified to class j by the neural network based on the confusion matrix as expressed in equation (2):

$$W_{ij} = cm_{ij} \qquad (2)$$

Second, the local credibility of the j^{th} target in the a^{th} neural network is calculated using the following equation（3）：

$$W_{\mathrm{Local}_{\mathrm{aj}}} = cm_{jj} \Big/ \sum_{i=1}^{N} cm_{ij} \qquad (3)$$

The global credibility of the neural network is ultimately determined by the following equation(4)：

$$W_{\mathrm{Global}_{aj}} = \sum_{i=1}^{4} Wii \Big/ N \qquad (4)$$

Decision-level integration of multiple thematic correlations

Evidence theory, also known as Dempster-Shafer (DS) theory, is a theoretical framework that has found wide application in the fields of multi-source information fusion and decision analysis. The core advantage of this theory lies in its ability to effectively integrate multiple pieces of uncertain information evidence. Through synthesis or reasoning processes, it standardizes and combines information from multi-source data, thereby enhancing the reliability and accuracy of fusion recognition. In this study, "evidence" is defined as uncertain information data involved in target recognition, including domain-specific topic terms, technology crossover topic terms, technology burst topic terms, and high-frequency topic terms. Meanwhile, "DS combination" refers to the process of synthesizing information represented by multi-source data through combination rules. By fusing and reasoning data sources under different topic relationships, it ultimately outputs decision inputs or decision results. The employment of DS evidence theory for decision-level fusion facilitates comprehensive observation of local feature values provided by multi-entity relationships, thereby enhancing the accuracy and reliability of decision-making.

Specifically, the basic probability assignment at the decision level can be achieved based on the global credibility and local credibility of the a^{th} BP neural network, combined with the posterior probability estimates provided by the algorithm classification results. The calculation of the basic probability assignment involves weighting and fusing the local credibility output by the BP neural network with the posterior probability, followed by normalization. $\mathrm{P}_{aj'}$ is the output of neural network test samples, and its calculation (5) is as follows:

$$P_{aj}' = W_{aj}P_{aj} \Big/ \sum_{i=1}^{4} W_{aj}P_{aj} \qquad (5)$$

The basic probability assignment is defined as the sum of all subset likelihood calculations for that S hypothesis to be true, expressing the level of confidence in the event that S is hypothesized to be true, with the formula as in (6):

$$m_a(S_1, S_2, \cdots S_j, \Theta,\ \emptyset) = (W_{Global_{aj}} * P_{a1}',\ W_{Global_{aj}} * P_{a2}',\ \cdots,\ 1 - W_{Global_{aj}},\ 0) \qquad (6)$$

Where $\Theta = (S_1,\ S_2,\ \cdots,\ S_j,\ \Phi)$ and S is each hypothesis in the recognition framework for the jth target type, i.e., as in Equation (7):

$$\mathrm{m}(\emptyset) = 0, \qquad (7)$$

$$m_a(S_1) + m_a(S_2) + \cdots + m_a(S_j) + m_a(\Theta) = 1$$

The subsequent step involves the utilization of the DS combination rule to adjust the basic probability assignments, which are defined as *n* mutually independent variables. The underlying assumption of this study is that the three relations *MB*, *MC*, and *MP* possess three trust functions within the same identification framework. It is further postulated that *m1*, *m2*, *m3*, and *m4* represent the fundamental probability assignments of their respective features. The combination rule, as in equation 8, is then employed to derive the subsequent results.

$$\mathrm{m(a)} = \frac{1}{\mathrm{K}} \sum_{\mathrm{a1 \cap a2 \cap a3 = a}} m_1(\mathrm{a1}) m_2(\mathrm{a2}) m_3(\mathrm{a3}) m_4(\mathrm{a4}) \quad (8)$$

where K is the normalization factor:

$$\mathrm{K} = \sum_{\mathrm{A1 \cap A2 \cap A3 = \emptyset}} m_1(\mathrm{a1}) m_2(\mathrm{a2}) m_3(\mathrm{a3}) m_4(\mathrm{a4}) \quad (9)$$

Therefore, based on the fused trust decision to derive which category of features the evidence data (subject terms in this study) belongs to, the feature with the greatest trust is selected as the verdict. If $\mathrm{m}(a_1) = max\{m(a_i),\ a_i \in \Theta\}$, then *a1* is the categorization result.

Data-level integration of multiple thematic correlations

The integration of the identification framework, basic probability assignment, and combination rules of Dempster-Shafer (DS) evidence theory results in the calculation of data fusion results. Subsequently, decision criteria are employed to identify different target objects, thereby achieving data fusion and target recognition for multi-topic relationships. Specifically, the first step involves the utilization of a BP neural network to achieve feature-level fusion, which is a process of feature input and decision output. DS evidence theory is implemented to attain decision-level fusion of multi-topic relationships, a process of combining decisions as input and producing decisions as output.

As illustrated in Figure 3, the data-level fusion comprises two stages: In the initial stage of fusion, DS evidence theory is employed to optimize decision-making regarding identification results, screen topic words, and select the feature word list with the highest degree of belief, thereby forming a comprehensive word list post-fusion. In the second stage, the basic association relationship matrix MB, the enhanced association relationship matrix MC, and the newly added association relationship matrix MP are reconstructed using the method described in Section 3.1.3. The BP neural network is then utilized to calculate the global confidence level of the comprehensive word list, which serves as the fusion association weight and becomes the foundation for calculating and realizing the fusion of multi-topic relationships.

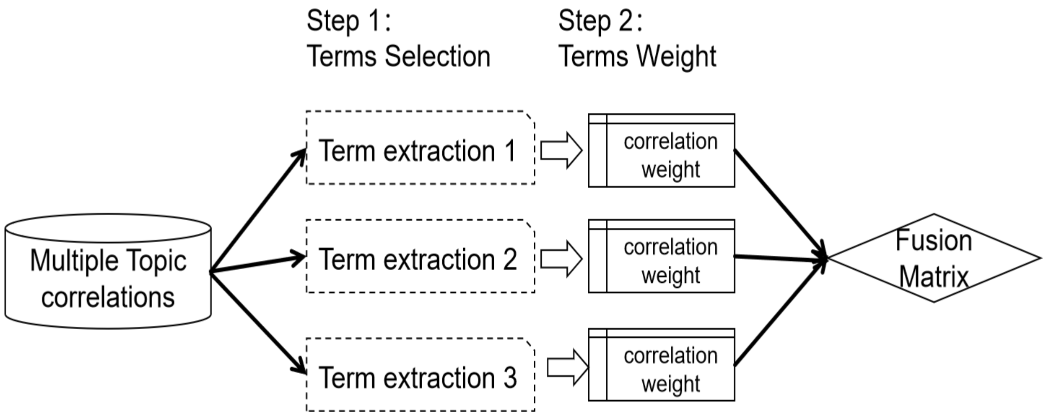

Figure 3. Multi-Relationship Matrix Fusion.

The fusion computation of the multi-relation matrix is intended to amplify the weight of evidence information that is conducive to correct identification, while minimizing the impact of ambiguous evidence information and evidence information that deviates significantly from the overall level. Therefore, this study sets the weights of three neural network relationships based on the global confidence level of the BP neural network, as shown in formulas (10), (11), and (12), respectively. This approach is predicated on the multi-objective characteristic of the feature-level fusion of the three types of matrices.

$$M_{MB}=\frac{W_{\mathrm{Global}_{MB}}}{(W_{\mathrm{Global}_{MB}}+W_{\mathrm{Global}_{MC}}+W_{\mathrm{Global}_{MP}})} \quad (10)$$

$$M_{MC}=\frac{W_{\mathrm{Global}_{MC}}}{(W_{\mathrm{Global}_{MB}}+W_{\mathrm{Global}_{MC}}+W_{\mathrm{Global}_{MP}})} \quad (11)$$

$$M_{MP}=\frac{W_{\mathrm{Global}_{MP}}}{(W_{\mathrm{Global}_{MB}}+W_{\mathrm{Global}_{MC}}+W_{\mathrm{Global}_{MP}})} \quad (12)$$

The weighted weights of the base relationship matrix MB, the enhanced relationship matrix MC, and the added relationship matrix MP are obtained according to the aforementioned method. The relationship fusion is then performed, and the fusion matrix is calculated to obtain the fusion matrix.

Empirical Study

This study selected the field of graphene sensing technology as the empirical research. A comprehensive patent analysis was conducted, yielding a set of subject feature words, including domain-specific terms, technology cross-over terms, technology burst terms, and high-frequency terms. These words were then used to construct three types of multiple topic correlations: terms co-occurrence, citation-term coupling, and patentee-term coupling. To verify the efficacy of the study's construction method, fusion calculations were performed.

In this study, graphene sensing technology is selected as the empirical technology field of the method due to its highly interdisciplinary nature, encompassing multiple technological domains such as materials, information, and biological sciences. This technology field is also characterized by its dynamism, marked by ongoing research activities, significant innovation, and rapid technological convergence. These characteristics are particularly relevant for the practical evaluation of the method outlined in this paper.

Data preprocessing

This study employs Derwent Innovation as its data source and devises a search strategy by combining the relevant concepts of graphene and sensor technology through Boolean logic. Following the filtration of search results, a collection of 974 patent families' documents was obtained, and the search for data continued up to September 22, 2022.

The feature items that are subsequently extracted from these patent data include subject terms, patent owner, and cited documents.

Of these, the patent owners and cited patent documents are directly extracted from the original data of the search results. The subject terms are extracted from the title and abstract text using text mining methods. The specific process is outlined as follows:

Step 1: The title and abstract text fields of the patent are segmented using the natural language processing (NLP) function of the Derwent Data Analyzer (DDA), resulting in a collection of 20,036 original subject terms (groups).

Step 2: The initial term sets must be rectified. The first step in the process is to convert all texts to lowercase in order to prevent errors. The second step involves the use of a built-in stop word list, thesaurus, etc., in order to remove general stop words, as well as format and grammatical terms in patent documents, DWPI catalog format abbreviations, compound name specifications, British and American spelling specifications, etc. The third step involves the use of Python's NLTK package stop word list to remove meaningless stop words and numbers, merge similar word forms, etc. Experts then perform manual cleaning, merge synonyms, and eliminate general subject words that are not closely related to substantive research, as well as conventional experimental tool names, material names, etc. After the above preprocessing operations, the pre-selected subject word set to be measured is obtained, containing 7873 words.

Step 3: It was implemented to extract feature parameters and feature vector sets. For 7873 pre-selected subject words, four types of features were obtained based on three different network of basic association MB, extended association MC and additional Association MP: namely, domain-specific terms, technology cross-over terms, technology burst terms, and high-frequency terms. Eight measurement indicators were calculated and collected respectively, which were HF, TFIDF, CMFS, IG, TI, SWI, KB and GR. The 8 parameters after standardization were used as indicators to be fused. The data examples are shown in Table 1, Table 2 and Table 3. The 7873 keywords in each type of network are divided into 5511 training samples and 2362 test samples in a ratio of 7:3.

Table 1. Top 20 Terms Eigenvalues in basic relation MB network.

Terms	*Frequency Features*		*Domain Feature*		*Interdisciplinary Feature*		*Breakthrough Feature*		*Cate-gory*
	HF	*TF-IDF*	*CMFS*	*IG*	*TI*	*SWI*	*KB*	*Max (GR)*	
analyte	0.430	0.040	0.680	0.299	0.340	0.411	0.000	0.015	1
nanopore	0.118	0.125	0.700	0.440	0.072	0.431	0.013	0.029	1
binding member	0.067	0.929	0.767	0.667	0.007	0.154	0.000	0.000	1
response data	0.040	0.550	0.692	0.557	0.019	0.433	0.000	0.000	1
gip agonist peptide	0.033	1.000	0.671	0.555	0.006	0.154	0.000	0.000	1
sensing layer	0.020	0.054	0.684	0.486	0.016	0.262	0.000	0.004	2
photolumine scent nanostructur e	0.020	0.092	0.716	0.518	0.015	0.212	0.000	0.008	2
ionic liquid	0.019	0.050	0.670	0.371	0.010	0.236	0.000	0.000	2
graphene channel	0.018	0.068	0.658	0.455	0.015	0.253	0.000	0.004	2
balloon	0.017	0.052	0.702	0.470	0.010	0.178	0.000	0.003	2
detecting	0.541	0.010	0.635	0.198	0.766	0.462	0.000	0.041	3
sample	0.385	0.022	0.655	0.232	0.353	0.436	0.000	0.016	3
solution	0.320	0.018	0.681	0.300	0.486	0.413	0.000	0.013	3
material	0.315	0.014	0.654	0.236	0.665	0.598	0.000	0.078	3
antibody	0.299	0.026	0.666	0.200	0.330	0.533	0.000	0.065	3
method	1.000	0.011	0.670	0.182	0.893	0.531	1.000	0.075	4
layer	0.662	0.021	0.701	0.266	0.565	0.495	0.479	0.044	4
sensor	0.651	0.013	0.669	0.249	1.000	0.514	0.687	0.013	4
surface	0.583	0.013	0.658	0.175	0.691	0.533	0.697	0.092	4
patient	0.581	0.018	0.694	0.306	0.743	0.554	0.530	0.011	4

Table 2. Top 20 Terms Eigenvalues in extended relation MC network.

Terms	*Frequency Features*		*Domain Feature*		*Interdisciplinary Feature*		*Breakthrough Feature*		*Cate-gory*
	HF	*TF-IDF*	*CMFS*	*IG*	*TI*	*SWI*	*KB*	*Max (GR)*	
analyte	0.685	0.063	0.707	0.257	0.359	0.340	0.000	0.135	1
nutritional substance	0.370	0.078	0.809	0.686	0.285	0.556	0.000	0.005	1
binding member	0.198	0.929	0.791	0.672	0.016	0.157	0.000	0.000	1
nanopore	0.181	0.165	0.715	0.451	0.092	0.302	0.000	0.030	1
gip agonist peptide	0.096	1.000	0.692	0.572	0.013	0.157	0.000	0.000	1

composite material	0.038	0.030	0.692	0.466	0.030	0.225	0.000	0.020	2
electrode array	0.036	0.106	0.754	0.628	0.039	0.273	0.000	0.000	2
sensing device	0.032	0.051	0.682	0.416	0.028	0.144	0.000	0.007	2
enzyme-free glucose sensor	0.030	0.062	0.710	0.662	0.009	0.157	0.000	0.002	2
intermediate body	0.030	0.087	0.742	0.752	0.018	0.268	0.000	0.000	2
method	1.000	0.029	0.711	0.178	1.000	0.453	0.000	0.092	3
layer	0.528	0.038	0.694	0.260	0.496	0.335	0.000	0.016	3
device	0.487	0.029	0.686	0.257	0.676	0.560	0.000	0.061	3
detecting	0.438	0.019	0.640	0.187	0.626	0.449	0.000	0.039	3
sample	0.398	0.039	0.660	0.236	0.380	0.432	0.000	0.015	3
sensor	0.679	0.022	0.689	0.246	0.977	0.491	1.000	0.020	4
surface	0.587	0.026	0.674	0.135	0.797	0.563	0.759	0.053	4
substrate	0.534	0.029	0.680	0.219	0.688	0.534	0.539	0.118	4
binding	0.460	0.056	0.688	0.194	0.372	0.443	0.285	0.063	4
amino acid	0.436	0.220	0.928	0.186	0.387	0.278	0.000	1.000	4

Table 3. Top 20 Terms Eigenvalues in additional relation MP network.

Terms	*Frequency Features*		*Domain Feature*		*Interdisciplinary Feature*		*Breakthrough Feature*		*Cate-gory*
	HF	*TF-IDF*	*CMFS*	*IG*	*TI*	*SWI*	*KB*	*Max (GR)*	
nutritional substance	0.815	0.191	0.889	0.595	0.437	0.649	0.000	0.069	1
analyte	0.556	0.200	0.795	0.350	0.217	0.653	0.000	0.000	1
patient	0.407	0.177	0.872	0.476	0.232	0.544	0.000	0.000	1
target analyte	0.370	0.365	0.794	0.583	0.161	0.594	0.000	0.000	1
nanopore	0.284	1.000	0.842	0.722	0.119	0.758	0.000	0.000	1
cnt	0.049	0.174	0.819	0.747	0.039	0.593	0.000	0.000	2
diabetes	0.037	0.059	0.706	0.350	0.010	0.211	0.000	0.000	2
conductive polymer	0.037	0.130	0.746	0.534	0.021	0.409	0.000	0.000	2
sensor chamber	0.037	0.130	0.744	0.527	0.021	0.409	0.000	0.000	2
pressure sensor	0.037	0.130	0.763	0.595	0.010	0.211	0.000	0.000	2
compound	0.506	0.135	0.768	0.206	0.458	0.824	0.000	0.000	3
graphene	0.457	0.095	0.751	0.212	0.308	0.594	0.000	0.130	3
sensor	0.444	0.054	0.753	0.230	0.510	0.558	0.000	0.042	3
substances	0.383	0.080	0.755	0.339	0.326	0.473	0.000	0.000	3
organoleptic	0.358	0.095	0.774	0.392	0.351	0.455	0.000	0.000	3
method	1.000	0.075	0.736	0.145	1.000	0.711	1.000	1.000	4
device	0.580	0.083	0.744	0.199	0.365	0.636	0.478	0.111	4
system	0.506	0.095	0.761	0.335	0.563	0.731	0.351	0.093	4
detecting	0.494	0.060	0.704	0.174	0.490	0.576	0.569	0.333	4
sample	0.481	0.148	0.752	0.334	0.174	0.510	0.199	0.083	4

The Confusion matrix and target identification results

The application of the aforementioned method entails the substitution of training samples into the neural network to facilitate preliminary target recognition and classification. Subsequent to the execution of this operation, the fundamental probability assignment derived from fuzzy processing is modified. Concurrently, the evidence theory space is formed, integrating the previously obtained data. The confusion matrix corresponding to the multi-topic relationships are obtained according to the classification results of the entire set.

MB confusion matrix:

$$\begin{bmatrix} 0.9333 & 0.0018 & 0 & 0 \\ 0.0333 & 0.9866 & 0.0077 & 0.1316 \\ 0.1 & 0.0018 & 0.9957 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

MC confusion matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.9849 & 0.0138 & 0 \\ 0.1429 & 0.0022 & 0.9842 & 0.3333 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

MP confusion matrix:

$$\begin{bmatrix} 0.9767 & 0.0159 & 0 & 0 \\ 0 & 0.8889 & 0.0464 & 0 \\ 0 & 0.0159 & 0.9404 & 0.5333 \\ 0 & 0 & 0.0066 & 0.9333 \end{bmatrix}$$

The global and local credibility of the multi-topic relationship are calculated using the confusion matrix, as illustrated in Table 4. The global credibility indicates that the classification recognition rate of the basic relationship MB is high, while the classification recognition efficiency of the newly added relationship MP is relatively low. It is evident that misjudgments occur in the recognition and classification of the three topic relationships. Specifically, the basic relationship MB demonstrates a high recognition rate for domain features, cross features, and burst features, while the enhanced relationship MC exhibits a high recognition rate for domain features, cross features, and word frequency features. Notably, the newly added relationship MP shows a high recognition rate for word frequency features, domain features, and cross features. Consequently, each multi-topic relationship exhibits distinct recognition accuracy rates for various indicators, thereby effectively addressing the issue of substantial confidence disparities during target recognition under a single topic relationship.

Table 4. The retrieval strategy for Grapheny Sensing Technology.

BP Neural Network	*Global Credibility*	*Local Credibility*			
		Frequency Features	*Domain Feature*	*Interdisciplinary Feature*	*Breakthrough Feature*
MB	0.937	0.875	0.996	0.9923	0.884
MC	0.902	0.875	0.998	0.986	0.75
MP	0.887	1	0.966	0.947	0.636

The Feature Terms Selection

In order to mitigate the discrepancy between the output of the neural network and the actual classification of feature words, the output results of the BP network are normalized. Subsequently, the fusion evaluation is performed using the evidence theory (see Table 5) to mitigate the impact of uncertain factors. The core subject word list is selected and merged based on the output fusion classification results and combined with expert opinions. After deduplication, the comprehensive subject word list used in this experiment is obtained, which contains 887 core terms (groups).

Table 5. The Top terms of Basic probability distribution function assignment and fusion recognition results.

Terms	*Network*	$m_{Frequency}$ *(S1)*	m_{Domain} *(S2)*	$m_{Interdisciplinary}$ *(S3)*	$m_{Breakthrough}$ *(S4)*	$m_a(\Theta)$	*Identification Results*
arterial pressure	MB	0.000	0.909	0.000	0.000	0.091	S2
	MC	0.000	0.854	0.000	0.000	0.146	S2
	MP	0.000	0.854	0.008	0.007	0.128	S2
conductivity	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.001	0.003	0.846	0.001	0.150	S3
hemorrhage	MB	0.000	0.909	0.000	0.000	0.091	S2
	MC	0.000	0.854	0.000	0.000	0.146	S2
	MP	0.878	0.000	0.000	0.000	0.122	S1
expression	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.001	0.837	0.023	0.000	0.139	S2
	MP	0.046	0.305	0.478	0.020	0.150	S3
synthetic compounds	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.010	0.651	0.151	0.037	0.151	S2
metal	MB	0.000	0.000	0.000	0.909	0.091	S4
	MC	0.036	0.000	0.000	0.768	0.195	S4
	MP	0.000	0.032	0.579	0.224	0.164	S3
cholesterol	MB	0.006	0.000	0.008	0.895	0.091	S4
	MC	0.006	0.000	0.963	0.023	0.008	S3
	MP	0.002	0.031	0.903	0.040	0.025	S3
conditioner	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.001	0.007	0.836	0.004	0.152	S3
nerve cell	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3

	MP	0.005	0.004	0.839	0.002	0.150	S3
polymers	MB	0.000	0.000	0.000	0.909	0.091	S4
	MC	0.025	0.000	0.000	0.782	0.193	S4
	MP	0.001	0.004	0.998	0.001	-0.004	S3
logistic transport	MB	0.000	0.899	0.010	0.000	0.091	S2
	MC	0.001	0.850	0.002	0.000	0.147	S2
	MP	0.874	0.001	0.002	0.000	0.122	S1
stem cell	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.005	0.004	0.839	0.002	0.150	S3
glucose concentrati on	MB	0.000	0.000	0.010	0.899	0.091	S4
	MC	0.080	0.000	0.060	0.662	0.198	S4
	MP	0.002	0.003	0.997	0.001	-0.003	S3
calibration temperature sensor	MB	0.000	0.908	0.001	0.000	0.091	S2
	MC	0.000	0.845	0.008	0.000	0.147	S2
	MP	0.007	0.184	0.646	0.008	0.155	S3
amyotrophic lateral sclerosis	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.003	0.845	0.000	0.152	S3
	MP	0.004	0.077	0.754	0.011	0.154	S3
high reactivity	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.023	0.825	0.000	0.152	S3
	MP	0.004	0.021	0.817	0.005	0.152	S3
progression	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.848	0.000	0.152	S3
	MP	0.716	0.009	0.146	0.005	0.124	S1
beverage consumptions	MB	0.000	0.881	0.027	0.000	0.091	S2
	MC	0.001	0.846	0.007	0.000	0.146	S2
	MP	0.001	0.750	0.098	0.014	0.138	S2
pressure sensor	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.008	0.013	0.808	0.012	0.159	S3
	MP	0.429	0.443	0.012	0.001	0.115	S2
transmembr ane pore	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.003	0.412	0.438	0.001	0.147	S3
	MP	0.000	0.000	0.851	0.000	0.149	S3
substrate platform	MB	0.205	0.478	0.174	0.052	0.091	S2
	MC	0.155	0.612	0.034	0.012	0.188	S2
	MP	0.926	0.000	0.007	0.000	0.066	S1
transistor	MB	0.003	0.000	0.000	0.906	0.091	S4
	MC	0.001	0.000	0.984	0.014	0.002	S3
	MP	1.047	0.000	0.000	0.000	-0.047	S1
high reliability	MB	0.000	0.000	0.908	0.001	0.091	S3
	MC	0.002	0.495	0.358	0.000	0.145	S2
	MP	0.004	0.077	0.754	0.011	0.154	S3
antibody	MB	0.000	0.000	0.909	0.000	0.091	S3
	MC	0.000	0.000	0.843	0.004	0.153	S3
	MP	0.002	0.079	0.405	0.209	0.305	S3

The Matrix extraction and fusion calculation

Based on the 887 terms (groups) in the comprehensive term list, the terms co-occurrence, the citation-terms coupling, and patent assignee-terms coupling were extracted, and 84,958 groups of multivariate relationships were obtained. Based on these relationships, three types of association matrix were calculated and constructed: MB, MC, MP.

The 887 comprehensive subject terms (groups) were substituted into the neural network for target classification. According to the classification results, the MB_{core}, MC_{core}, and MP_{core} confusion matrix corresponding to the multiple topic associations were obtained, and the global credibility and local confidence were calculated based on the confusion matrix. Therefore, the fusion weights of the multivariate subject association network were judged to be 0.361, 0.333, and 0.305, respectively (Table 6).

MB_{core} confusion matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.9767 & 0 & 0.0556 \\ 0.0833 & 0.0233 & 0.9227 & 0.7222 \\ 0.0833 & 0 & 0 & 0.9444 \end{bmatrix}$$

MC_{core} confusion matrix:

$$\begin{bmatrix} 0.8333 & 0 & 0.0065 & 0 \\ 0.1667 & 0.8519 & 0.013 & 0.0909 \\ 0 & 0.2222 & 0.9351 & 0.3636 \\ 0.1667 & 0 & 0.0065 & 0.8182 \end{bmatrix}$$

MP_{core} confusion matrix:

$$\begin{bmatrix} 0.8333 & 0 & 0.0143 & 0.1111 \\ 0.1667 & 0.625 & 0.0143 & 0 \\ 0.3333 & 0.375 & 0.8429 & 0.4444 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Table 6. BP neural network core keyword target recognition results.

BP Neural Network	*Global Credibility*	*Local Credibility*				*Fusion Weight*
		Frequency Features	*Domain Feature*	*Interdisciplinary Feature*	*Breakthrough Feature*	
MB	0.846	0.857	0.977	1	0.548	0.361
MC	0.781	0.714	0.793	0.973	0.643	0.333
MP	0.715	0.625	0.625	0.967	0.643	0.305

Discussion

Theme correlation matrix fusion effect

According to the ranking of the fusion matrix value results, the top 50 terms(group) associations were selected, and their situations in the fusion matrix and the three types of subject association matrices were analysed. Table 7 shows the 50 groups of subject associations with the highest fusion weights. Among them, the top three groups of subject associations with the highest fusion association weights are (method, solution), (method, patient) and (method, detecting), and their fusion matrix association weights are 0.361, 0.359 and 0.35 respectively. Among them, the subject word with the strongest co-occurrence association is (method, solution), the citation-subject word coupling association with the strongest association is (semiconductor, multi-walled carbon), and the patent applicant-subject word coupling association with the strongest association is (resonator pattern, expandable element) and (resonator pattern, flexible circuit assembly).

Table 7. Correlation of graphene sensing field fusion matrix (TOP 50).

No.	*Keyword1*	*Keyword2*	$M_{MB+MC+MP}$	*MB*	*MC*	*MP*
1	method	solution	0.361	1	0	0
2	method	patient	0.359	0.995	0	0
3	method	detecting	0.35	0.969	0	0
4	semiconductor	multi-walled carbon nanotubes	0.333	0	1	0
5	apparatus	discrete operative device	0.32	0	0.961	0
6	stimulation	discrete operative device	0.32	0	0.961	0
7	sensor	detecting	0.314	0.87	0	0
8	resonator pattern	expandable element	0.305	0	0	1
9	resonator pattern	flexible circuit assembly	0.305	0	0	1
10	layer	electrode	0.293	0.812	0	0
11	detecting	discrete operative device	0.285	0	0.854	0
12	reservoir	discrete operative device	0.285	0	0.854	0
13	patient	device	0.285	0.788	0	0
14	patient	system	0.28	0.775	0	0
15	layer	substrate	0.275	0.761	0	0
16	method	sensor	0.267	0.738	0	0
17	method	surface	0.266	0.736	0	0
18	hemorrhage	expandable element	0.259	0	0	0.848
19	hemorrhage	flexible circuit assembly	0.259	0	0	0.848
20	sensor	patient	0.255	0.706	0	0
21	method	glucose	0.254	0.703	0	0
22	nanowire	cadmium	0.252	0	0.756	0
23	method	chemistry	0.25	0.691	0	0
24	method	discrete operative device	0.249	0	0.747	0
25	cancer	discrete operative device	0.249	0	0.747	0
26	nerve	discrete operative device	0.249	0	0.747	0
27	parameters	discrete operative device	0.249	0	0.747	0
28	discrete operative device	electrical conductivity	0.249	0	0.747	0

29	discrete operative device	accurate detection	0.249	0	0.747	0
30	method	device	0.235	0.65	0	0
31	surface	detecting	0.234	0.649	0	0
32	method	electrode	0.234	0.647	0	0
33	layer	device	0.234	0.647	0	0
34	surface	device	0.233	0.645	0	0
35	resonator pattern	pressure sensor	0.232	0	0	0.759
36	method	graphene	0.231	0.638	0	0
37	electrodes	discrete operative device	0.23	0	0.689	0
38	method	sample	0.23	0.636	0	0
39	layer	sensor	0.227	0.627	0	0
40	substrate	lumen	0.22	0	0.66	0
41	surface	electrode	0.219	0.607	0	0
42	method	substrate	0.219	0.607	0	0
43	layer	discrete operative device	0.217	0	0.65	0
44	substrate	discrete operative device	0.217	0	0.65	0
45	device	detecting	0.217	0.6	0	0
46	resonator pattern	camera	0.216	0	0	0.708
47	sensor	substrate	0.216	0.597	0	0
48	material	expandable element	0.215	0	0	0.704
49	material	flexible circuit assembly	0.215	0	0	0.704
50	layer	surface	0.214	0.593	0	0

Comparison of three types of topic associations

To further study whether the three types of subject associations have complementary significance to each other, this study uses Jaccard similarity analysis (Jaccard, 1912) to compare the similarity between the three types of subject associations. The values of subject associations are divided into 0 and non-0 categories, and the similarities between the three relationships of terms co-occurrence, citation-terms coupling, and patent assignee-terms coupling are calculated respectively. The Jaccard distance is employed to quantify the similarity between the three sets of subject associations. It is noteworthy that the magnitude of this value directly corresponds to the extent of dissimilarity between the sets.

The repetition rate of non-0 elements in the three subject associations and the calculation results of Jaccard distance are shown in Table 8. The calculation results reveal that there is no overlap between the subject word co-occurrence association and the other two associations, and there is a very small amount of overlap between the citation-subject word coupling association and the patent applicant-subject word coupling association. The Jaccard distance between the three types of relationships is extremely high. It is evident that the subject features revealed by these three types of associations differ significantly and are highly complementary. Therefore, the integration of these three types of associations is conducive to enriching the clues of subject feature and enriching and deepening the significance of text mining.

Table 8. Similarity comparison of three types of topic association matrices in the field of graphene sensing.

Theme Relationship	*MB VS. MC*	*MB VS. MP*	*MC VS. MP*
Overlapping Relationship	0	0	278
Overlapping Rate	0%	0%	$\frac{\mathrm{MC} \cap \mathrm{MP}}{\mathrm{MC}}=1.506\%$ $\frac{\mathrm{MC} \cap \mathrm{MP}}{\mathrm{MP}}=13.385\%$
Jaccard Distance	1	1	0.986

The significance of multi-relationship integration

Based on the weighted fusion method of network credibility, this study obtained the fusion matrix of three types of associations. The variance and sparsity comparison of each matrix is shown in Table 9. The results show that the variance of the fusion matrix $M_{MB+MC+MP}$ is 0.013, which is smaller than the variance of the MB, MC, and MP matrix. At the same time, the sparsity of the fusion matrix is also notably lower than that of the three types of topic association matrix. The fusion matrix demonstrates a substantial degree of richness in information and exhibits a minimal degree of discreteness.

Table 9. Variance and sparsity of three types of topic association and fusion matrix in graphene sensing field.

Theme Relationship	*MB*	*MC*	*MP*	$M_{MB+MC+MP}$
Variance	0.029	0.049	0.092	0.013
Sparsity	0.836	0.953	0.995	0.784

Furthermore, as illustrated in Table 9, the MP matrix exhibits the highest degree of sparsity, indicating that the extracted relationship is relatively weak. The MC matrix demonstrates the second highest sparsity, while the MB matrix exhibits the lowest sparsity. Concurrently, as illustrated by Table 6, the fusion matrix allocates minimal importance to the MP matrix due to the limited number of feature relationships and the low level of information accuracy. A comparison of Tables 7 and 9 reveals the fusion matrix assigns a relatively low weight to infrequent relationships. However, it also attains 0.305. Therefore, infrequent relationships, despite their relative weakness, are represented to a certain extent in the fusion matrix. To elucidate this assertion, the association relationship of resonator pattern, expandable element and resonator pattern, flexible circuit assembly can be utilized as an instance. It is not expressed in the MB and MC matrices. However, it is observed that the fusion matrix (MP) assigns a weight of 1 to this relationship, signifying its significance. The weight value in the fusion matrix is 0.305. Within the 84,680 non-zero value relationships present within the fusion matrix, it is ranked 8th, indicating its significant contribution to the fusion matrix. This observation signifies that the fusion matrix does indeed consider weak relationships to a certain extent. It is evident that the fusion matrix can not only effectively represent the primary attributes in multivariate relations but also possess a satisfactory degree of expressiveness for rare topic-related attributes.

Conclusions

This paper researches and proposes a multiple topic-association fusion method suitable for patent analysis. The proposed method is founded on the association features of multi-source topic data obtained in patent text mining. First, the paper extracts multiple topic correlations based on a number of multiple topic association relationships. Next, it combines a neural network and an evidence theory approach, creating a fusion method for multiple thematic correlations. The objective of this fusion method is to generate an information enhancement matrix that contains more comprehensive topic association relationships. Empirical study was conducted on the patent data in the domain of graphene sensing technology. It aimed to validate the efficacy of an integrated method of multiple thematic correlations. The method involves the learning of the weight distribution of different topic-associated relations through neural networks, the use of evidence theory to model and fuse the uncertainty of multi-source information, and the final generation of the topic-associated enhancement matrix. The empirical study demonstrated significant disparities among the three categories of subject association relationships: subject term co-occurrence, citation-subject term coupling, and patent applicant-subject term coupling. Their subject features manifested stronger complementarity. The fused matrix is characterized by enhanced informational content and reduced discreteness, resulting in a more comprehensive characterization of the primary subject association attributes. Additionally, the fused matrix is capable of expressing rare topic-weakly associated attributes with high efficacy. The integration of information from multiple sources through the fusion method, which is based on a neural network and evidence theory, has been shown to enhance the characterization of the association relationship between topics.

The following advantages and innovations of the proposed method are evident. Firstly, multi-source information fusion is achieved by combining neural networks and evidence theory, effectively combining three types of theme association relations: terms co-occurrence, citation-terms coupling, and patent assignee-terms coupling. This combination overcomes the limitations of a single association relation and significantly improves the comprehensiveness and accuracy of the theme association analysis. Secondly, uncertainty modelling is employed, which is another innovation. The modelling of uncertainty in multi-source information by evidence theory enhances the method's reliability and credibility. Furthermore, this enhanced uncertainty modelling provides a more robust foundation for analysing complex thematic association relationships. Enhancement of theme features: The enhanced theme association matrix, resulting from the fusion process, can effectively capture not only the primary theme attributes but also those of less prevalent theme weak associations. Consequently, this multifaceted approach provides a richer array of clues for the identification of technology themes and potential technological innovations.

Despite the findings of this study, there remain several issues that require further investigation. Specifically, there is a need for further refinement of the multiple topic association fusion method and its application. Optimization of computational efficiency is essential for addressing the challenges posed by the growth in size of the topic terms. This growth results in exponential expansion of the dimensionality of the matrix, necessitating high-performance computing capabilities. In future research, we will explore more efficient methods for dimensionality reduction processing of thematic feature terms (e.g., techniques based on graph embedding or sparse representation) to improve computational efficiency and reduce resource consumption. The verification of method universality is imperative. The present study principally focuses on the integration of three types of thematic associations, which can be further extended to more types of thematic associations (e.g., technological efficacy associations, technological evolution associations, etc.) in the future to verify the universality and robustness of the constructed method. The Dynamic Theme Modelling method will be employed to explore the evolving nature of the theme association relationships. The integration of dynamic theme modelling will facilitate the exploration of the temporal evolution of the theme association relationship, thereby providing a more profound foundation for the prediction of technological development trends and the prospective research of technological innovation. The method is applied to other technological fields (e.g., artificial intelligence, biomedicine, etc.) to verify its applicability and effectiveness in different fields and further expand the application scope of the method.

The multivariate topic association fusion method based on a neural network and evidence theory proposed in this study provides novel concepts and methodological support for patent text topic mining. As evidenced by the experimental findings, the proposed method has the capability of effectively integrating multi-source information, thereby enhancing the characterization of topic-association relationships. This, in turn, provides a powerful tool for technology topic analysis, technology trend prediction, and related fields. Subsequent research endeavours will focus on enhancing the method's computational efficiency, validating its universality, and incorporating dynamic topic modelling. These efforts aim to broaden the application of the method in diverse scenarios, thereby providing more intelligent and precise support for technological innovation and patent analysis.

Acknowledgments

The research is the outcome of the projects, “Youth Innovation Promotion Association (2022173)” “Light of West China program”, “Intelligence Service for Research Institutes on Science, Technology and Innovation (E3291106)”, “Research on Emerging Technology Direction Identification Method based on Technology Fusion (E3Z0000803)”, supported by Chinese Academy of Sciences(CAS), “Study on the Multi-Relation Data Fusion Methods for Identification and Prediction of Technology Innovation Paths”(No. 18BTQ067) supported by National Social Science Fund of China and “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. Additionally, the special contribution of Xian Zhang as Corresponding Author (zhangx@clas.ac.cn) is noteworthy.

References

Chawla, S., Kaur, R., Aggarwal, P., 2023. Text classification framework for short text based on TFIDF-FastText. Multimed. Tools Appl. 82, 40167–40180. https://doi.org/10.1007/s11042-023-15211-5

Feng, G., Wu, J., Mo, X., 2020. Research on detection and verification of burst words with multiple measures. Library and Information Service 64, 67–76. https://doi.org/10.13266/j.issn.0252-3116.2020.11.008

Hamda, N.E.I., Hadjali, A., Lagha, M., 2023. Multisensor data fusion in IoT environments in dempster-shafer theory setting: An improved evidence distance-based approach. Sens. 23, 5141. https://doi.org/10.3390/s23115141

Jaccard, P., 1912. The distribution of the flora in the alpine zone.1. New Phytologist 11, 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Kleinberg, J., 2002. Bursty and hierarchical structure in streams, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02. Association for Computing Machinery, New York, NY, USA, pp. 91–101. https://doi.org/10.1145/775047.775061

Li, K., Dai, Z., Zuo, C., Wang, X., Cui, H., Song, H., Cui, M., 2024. Scene adaptation in adverse conditions: a multi-sensor fusion framework for roadside traffic perception. J. Intell. Transp. Syst. https://doi.org/10.1080/15472450.2024.2390844

Li, Z., Zhang, Q., Liu, S., Peng, Y., Li, L., 2024. Information fusion and attribute reduction for multi-source incomplete mixed data via conditional information entropy and D-S evidence theory. Appl. Soft Comput. 151, 111149. https://doi.org/10.1016/j.asoc.2023.111149

Lin, C., Kong, Y., Han, Q., Chen, K., Geng, Z., Wang, T., Dong, M., Liu, H., Chu, F., 2025. IF-EDAAN: an information fusion-enhanced domain adaptation attention network for unsupervised transfer fault diagnosis. Mech. Syst. Signal Process. 224, 112180. https://doi.org/10.1016/j.ymssp.2024.112180

Liu, Huijie, Wu, H., Zhang, L., Yu, R., Liu, Y., Liu, C., Li, M., Liu, Q., Chen, E., 2022. A hierarchical interactive multi-channel graph neural network for technological knowledge flow forecasting. Knowl. Inf. Syst. 64, 1723–1757. https://doi.org/10.1007/s10115-022-01697-2

Liu, Huailan, Zhang, R., Liu, Y., He, C., 2022. Unveiling evolutionary path of nanogenerator technology: a novel method based on sentence-BERT. Nanomaterials 12, 2018. https://doi.org/10.3390/nano12122018

Liu J., Zhong Y., He X., Li Z., Zhao Z., Wang H., Song S., 2024. Research on Scientific-Technological-Industrial Association Patterns Based on Multi-relationships Fusion. Journal of Modern Information 44, 67–81.

Pan, X., Wang, Y., He, S., 2021. The evidential reasoning approach for renewable energy resources evaluation under interval type-2 fuzzy uncertainty. Information Sciences 576, 432–453. https://doi.org/10.1016/j.ins.2021.06.091

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. https://doi.org/10.5120/ijca2018917395

Qian J., Tong Z., Yu Y., Hong C., Miao D., 2023. Multi-source information fusion through generalized adaptive multi-granulation. CAAI Transactions on Intelligent Systems 18, 173–185.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Song X., Chen M., Lyu G., Shen Y., 2023. A Research on Competitor Identification Model Based on Crossdomain and Multi-source Information Fusion in the Context of Big Data-Based on the New Energy Automobile Industry. Journal of the China Society for Scientific and Technical Information 42, 176–188.

Tan J., Yang A., Li T., Ma X., Chen S., 2022. Multi-source Fusion Navigation Algorithm Based on Robust Adaptive Filtering. Aerospace Control 40, 22–29. https://doi.org/10.16804/j.cnki.issn1006-3242.2022.05.012

Tseng, Y.-H., Lin, C.-J., Lin, Y.-I., 2007. Text mining techniques for patent analysis. Information Processing & Management, Patent Processing 43, 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

Xiao, F., 2023. GEJS: A generalized evidential divergence measure for multisource information fusion. IEEE Trans. Syst., Man, Cybern., Syst. 53, 2246–2258. https://doi.org/10.1109/TSMC.2022.3211498

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, 583–601. https://doi.org/10.1007/s11192-015-1792-2

Xu W., Huang X., Cai K., 2023. Review of Multi-source Information Fusion Methods Based on Granular Computing. Journal of Data Acquisition and Processing 38, 245–261. https://doi.org/10.16337/j.1004-9037.2023.02.002

Yan J., Li H., Ma Y., Liu Z., Zhang D., Jiang Z., Duan Y., 2024. Multi-source Heterogeneous Data Fusion Technologies and Government Big Data Governance System. Computer Science 51, 1–14.

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, 741–754. https://doi.org/10.1016/j.ipm.2011.12.005

Yu, Y., Ju, P., Shang, M., 2022. Research on the evaluation method of patent keyword extraction algorithm based on information gain and similarity. Libr. Inf. Serv. 66, 108–117. https://doi.org/10.13266/j.issn.0252-3116.2022.06.012

Zhai D., Lou Y., Kan H., He X., Liang G., Ma Z., 2023. Constructing TCM Knowledge Graph with Multi-Source Heterogeneous Data. Data Analysis and Knowledge Discovery 7, 146–158.

Zhang, Q., Zhang, P., Li, T., 2025. Information fusion for large-scale multi-source data based on the dempster-shafer evidence theory. Inf. Fusion 115, 102754. https://doi.org/10.1016/j.inffus.2024.102754

Zhang, X., Lin, J., 2025. Scalable data fusion via a scale-based hierarchical framework: adapting to multi-source and multi-scale scenarios. Inf. Fusion 114, 102694. https://doi.org/10.1016/j.inffus.2024.102694

Zhang X., Zeng R., Li S., Li J., 2024. Patented Technology Evolution Path Identification Based on Multi-relation Data Fusion. Library and Information Service 68, 71–84. https://doi.org/10.13266/j.issn.0252-3116.2024.03.007

Zhang, Z., Wang, H., Geng, J., Deng, X., Jiang, W., 2024. A new data augmentation method based on mixup and dempster-shafer theory. IEEE Trans. Multimedia 26, 4998–5013. https://doi.org/10.1109/TMM.2023.3330106

Zhu, J., Deng, A., Xing, L., Li, O., 2024. Rolling bearing fault diagnosis based on multi-source information fusion. J. Fail. Anal. Prev. 24, 1470–1482. https://doi.org/10.1007/s11668-024-01935-5