ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

A Multimodal Data Fusion Framework for Knowledge Graph Construction and Data Mining in Information Science

Ming Li [1] *and Yongjia Xu*[1,*]

[1] Business School of Hohai University, Nanjing 211100, Jiangsu, China

SUMMARY: *To solve the problem of scattered storage of information science objects among paper texts, citation paths, author institutions, source journals, topic tags and time periods, a multimodal knowledge graph empirical sample based on open academic metadata has been established and its function in knowledge organization and data mining has been tested. The sample includes records of information science and related directions from 2014 to 2024. After processing, 5200 papers, 13172 entities and 105148 relationships remain, forming seven kinds of nodes: papers, authors, institutions, sources, keywords, topics and years, and relationships such as authorship, institutional affiliation, source publication, keyword association, topic attribution, citation linkage and co-occurrence. In terms of methodology, the titles, abstracts and keywords of the papers, the citation network, the author-institution network, the source-topic distribution and the annual period were encoded into five kinds of modal features and compared in four tasks: entity alignment, relationship completion, cross-modal retrieval and topic mining: Text only, Citation GCN, Late Fusion, MKG-BERT and Proposed MDF-KG. The results show that the average Macro-F1 of Proposed MDF-KG in four types of tasks is 0.8535, which is higher than the 0.8213 of MKG-BERT and the 0.7880 of Late Fusion; the Mean Reciprocal Rank (MRR) reaches 0.912 in the relationship completion task and the normalized discounted cumulative gain at 10 (nDCG@10) reaches 0.858 in cross-modal retrieval. The results of the ablation study indicate that the text modality has the highest average contribution and the performance decreases by 0.0553 after deleting it; the citation modality decreases by 0.0415 and the source topic modality decreases by 0.0318. The robustness test shows that when the field is missing or the noise ratio reaches 0.5, the average Macro-F1 of the Proposed MDF-KG is still 0.8080. The research findings suggest that the multimodal field organization can enhance the connectivity, ranking quality and error correction capability of information science knowledge graphs, but author alias, topic granularity and weak citation context are still the main obstacles for further development.*

KEYWORDS: *Informatics; Knowledge graph; Multi modal data fusion; Data mining; Academic metadata*

1 Introduction

The literature object faced by scientific and technological intelligence work has expanded from bibliographic information composed of titles, abstracts, and keywords to knowledge units composed of papers, authors, institutions, source journals, topic tags, citation paths, and annual changes. The retrieval of digital libraries, the analysis of disciplinary trends, the identification of research frontiers, and the evaluation of scientific research all require the

*xuyongjia0427@163.com
https://doi.org/10.65102/is20261284
Published: 10 June, 2026

simultaneous inclusion of paper content and academic relationships in the analysis. Open academic metadata provides a verifiable underlying condition for this type of research. OpenAlex organizes objects such as papers, authors, institutions, sources, and concepts into an open index; Crossref provides community maintenance metadata such as DOI, publication source, and publication time [2]; S2ORC supplements semantic records, abstracts, references, and structured clues throughout the paper [3]; SemOpenAlex transforms OpenAlex objects into large-scale RDF knowledge graphs, enabling academic objects to be connected in a queryable semantic space [4]. These resources have changed the problems of sample irreproducibility, unstable source caliber, and difficult alignment of relationship fields in intelligence research, and have also provided an executable foundation for empirical research on knowledge graphs based on public data.

In the knowledge organization field, a single document is not a separate text. The research theme, author cooperation, affiliation with institutions, journal sources, citation routes and topic tags of a paper jointly decide its place in the academic community. The relevant research of Open Research Knowledge Graph indicates that academic knowledge should be changed from the traditional form of documents to machine-readable structured objects [5]. Furthermore, the study on building knowledge graphs based on academic communication shows that rule extraction, machine learning, natural language processing and semantic normalization are the main technical links in constructing academic knowledge graphs [6]. In the aspect of open science, the research assessment suggests that the heterogeneity among publications, data, software, institutions and citation relationships will have an impact on the reliability and applicability of academic knowledge graphs [7]. Hence, the graph construction mentioned in this paper is not only converting paper records into nodes and edges, but also converting multi-source academic disciplines into empirical objects which can be used for retrieval, completion, alignment and topic mining.

The existing knowledge graph research offers a theoretical basis for the organization of information science data. The study on automatic knowledge graph construction focuses on extracting entities, relationships and conceptual structures from various sources of data, and updating the graph by means of fusion, inference and quality control [8]. Generally, knowledge graph research regards entities, relationships, attributes, pattern layer and instance layer as the fundamental elements of graph representation, providing a common framework for the heterogeneous modeling of objects like papers, authors, institutions and topics [9]. From the viewpoint of application, knowledge graphs have important advantages in information retrieval, recommendation, question answering and explanatory analysis, but are still limited by problems such as knowledge acquisition, knowledge fusion, relationship completion and noise spread [10]. These investigations show that the essence of intelligent knowledge graphs is not only in the capability to construct networks, but also in whether their structure can be supportive for distinct data mining tasks.

The academic object of information science has typical heterogeneous network characteristics. Papers, authors, institutions, sources, keywords, topics, and years are not homogeneous nodes, and the meaning of edges varies from authorship, affiliation, publication, citation, co-occurrence, to topic attribution. Research on heterogeneous information network analysis suggests that different types of nodes and semantic paths can affect similarity calculation, clustering, and recommendation results [11]. Research on knowledge graph embedding further indicates that entities and relationships need to be mapped to a low dimensional representation space in order to support link prediction, relationship inference, and entity alignment [12]. If these relationships are compressed into a single text similarity, the author institution, source boundary, and citation direction will be weakened; If only the citation network is retained, the semantic differences and topic

granularity in the title abstract will be difficult to enter the model.

Graph neural networks provide trainable representations for multi relational graphs. GraphSAGE supports inductive representation learning on large-scale graphs through neighborhood sampling and aggregation [13]; Relationship graph convolutional networks can learn parameterized propagation rules between different types of relationships, making them suitable for handling multi relationship knowledge graphs [14]; Graph attention networks improve the ability to utilize structural information through neighbor weight allocation [15]. These methods provide a fundamental reference for graph propagation, relationship completion, and modal fusion in this article. Unlike general social networks or product recommendation graphs, intelligence knowledge graphs have fewer node types but clearer semantic levels, and edge types are closer to the real evidence paths in academic activities. Therefore, model design needs to maintain a balance between structural propagation and academic field constraints to avoid excessive mixing of adjacent topics due to deep propagation.

Text representation remains an important foundation of the knowledge graph in information science. BERT has improved the semantic representation ability in academic texts through bidirectional context pre training [16]; Sentence BERT uses sentence vector learning for semantic similarity calculation, improving the matching efficiency between abstracts, titles, and keywords [17]; BERTopic combines topic embedding, clustering, and categorization TF-IDF, making it suitable for forming interpretable topics from paper abstracts and keywords [18]; LDA, as a classic topic model, provides a comparable baseline for estimating topic distributions in a collection of papers [19]. These textual methods can explain the research content, but the characterization of author institutions, citation directions, source boundaries, and annual evolution is still limited. This article takes text modality as the core evidence, while introducing citation, author institution, source topic, and temporal modality to reduce mismatches caused by single text judgments.

Scientific metrology research provides an evaluation tradition for graph structure analysis and result interpretation. VOSviewer transforms co word, co citation, and collaborative networks into interpretable knowledge structures [20]; CitNetExplorer is used to track the literature path and development chain in citation networks [21]; The modularity method can characterize the structure of online communities and assist in determining the boundaries of topic aggregation [22]; The PageRank idea transforms the link structure into node importance ranking, providing a foundation for citation and graph propagation analysis [23]. The nDCG metric in information retrieval evaluation emphasizes the impact of ranking position on the quality of results, and is suitable for testing whether relevant objects appear at the forefront in cross modal retrieval tasks [24]. Based on the above research, this article incorporates the quality of graph construction, mining performance, ablation contribution, robustness, efficiency, and error sources into the same data caliber, avoiding using only network display or single item accuracy evaluation of multimodal graphs.

2 Methods

2.1 Public Scholarly Metadata Collection and Multimodal Sample Organization

The research samples are arranged according to the field structure of publicly accessible academic metadata, including related areas like information science, library and information science, scientometrics, knowledge organization, digital libraries, semantic retrieval,

knowledge graphs and data mining from 2014 to 2024. The sample is constructed with papers as the fundamental unit of observation, keeping fields such as openalex_id, semantic_stcholar_i, title, abstract, publication year, literature type, source name, citation count, number of references, number of authors, number of institutions, number of keywords and number of topics. If the DOI is missing, the record is maintained to keep the consistency of the entity and to prevent a large reduction in the sample size due to the lack of sufficient DOI coverage. The original target number is 6000 records, and after eliminating the samples which do not meet the criteria of field integrity, publication type and task availability, 5200 papers will be finally selected.

After cleaning, there were 3130 journal articles in the sample, accounting for 60.19%; 1057 reviews, accounting for 20.33%; 1013 conference papers, accounting for 19.48%. The average citation count of a single paper is 17.47, the average number of references is 31.49, the average number of authors is 2.74, the average number of institutions is 2.67, the average number of keywords is 4.61, and the average number of topic tags is 1.34. This distribution indicates that the sample is not only composed of highly cited papers or a single literature type, but can also cover stable journal articles, topic reviews, and conference communication records. In terms of annual structure, the sample has expanded from 283 in 2014 to 659 in 2024, with a time window sufficient to observe the growth and changes of information science objects in knowledge graph and semantic retrieval related directions.

Sample cleaning is divided into four actions: field filtering, duplicate merging, task availability tagging, and relationship pre extraction. Field filtering requires that the title length be no less than 20 characters, the abstract length be no less than 80 characters, the keywords be no less than 2, and the source field be recognizable. Priority is given to using DOI for repeated merging, and when DOI is missing, the similarity of the word set normalized by the title is used, with a threshold set at 0.92. Task availability tagging does not force all records to be included in the same task. Records with less than 3 references can enter text and topic tasks, but not citation tasks. In the relationship pre extraction stage, only author, institution, source, keywords, topic, and year fields that can be materialized are retained, and weak evidence fields are not directly entered into the graph.

Multimodal fields are organized into five categories of evidence. The text modality consists of title, abstract, and keywords, used to express research content and conceptual boundaries; The citation mode consists of references, citation relationships, and connections between papers, used to describe knowledge sources and influence paths; The author institution modality consists of the author, institution, and collaborative relationship, used to identify the research production subject; The source topic modality consists of journals, conferences, topic tags, and keyword clusters, which are used to constrain disciplinary boundaries; The time mode consists of the distribution of publication year and theme year, and is used to identify the stage of theme evolution. The five modalities are aggregated around the nodes of the paper, and subsequent entity alignment, relationship completion, retrieval, and topic mining are all derived from the same cleaning sample.

The sample construction process is shown in Figure 1.

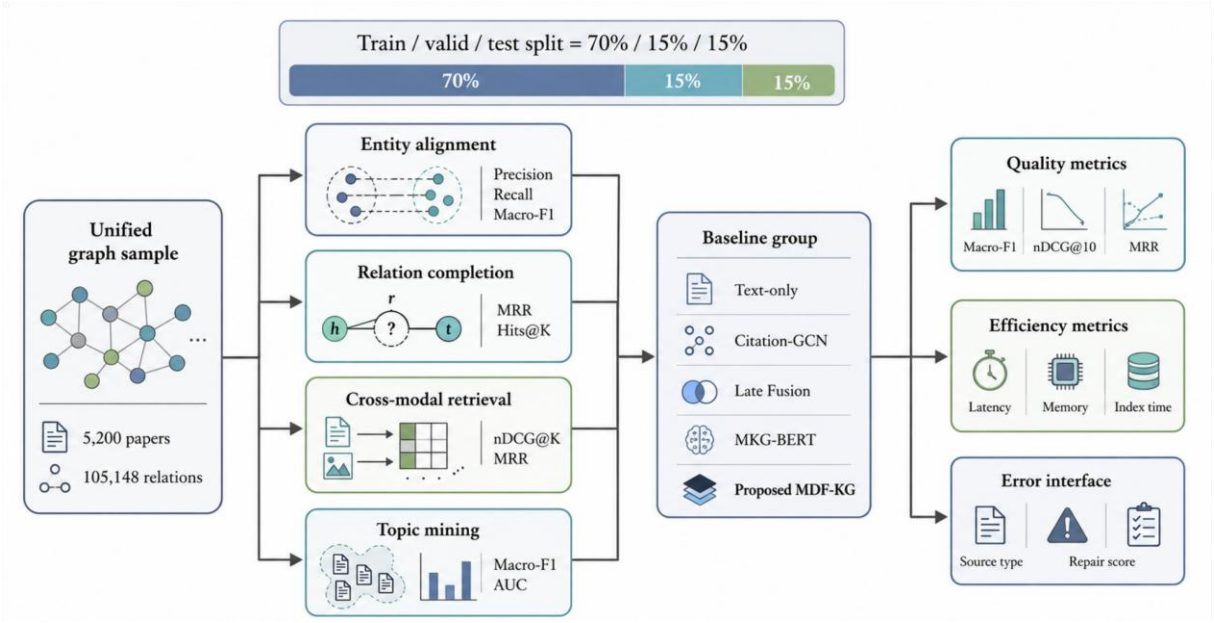

Figure 1: Organization of Public Academic Metadata Samples and Multi modal Field Coverage Mechanism

In Figure 1, the three types of public metadata fields are first imported into the cleaning sample pool, and then the availability of text, citation, institution, topic, and year fields is determined through the field coverage matrix. Clean the samples and continue to convert them into heterogeneous knowledge graphs and task samples, ensuring that graph statistics, model inputs, and experimental results use the same data caliber. Table 1 lists the sample size and coverage of major fields.

Table 1: Sample Size and Multimodal Field Coverage

Project	Value	Description
Sample Period	2014—2024	Continuous 11 years
Cleaned Papers	5200	Articles: 3130; Reviews: 1057; Proceedings: 1013
Average References	31.49	Core fields for citation tasks
Average Authors	2.74	Core fields for author institution modality
Average Keywords	4.61	Core fields for text and source theme modality
Text Field Coverage	1.00	Title, abstract, and keywords available
Citation Field Coverage	0.92	References and citation relationships available
Institution Field Coverage	0.88	Author institution relationships available
Theme Field Coverage	0.95	Theme tags or keyword clusters available

2.2 Multimodal Knowledge Graph Construction and Feature Fusion

The knowledge graph is represented using a heterogeneous graph structure. The node set includes seven types of entities: papers, authors, institutions, sources, keywords, topics, and years, with a total of 13172 entities. 7800 author nodes, accounting for 59.22%; There are 5200 paper nodes, accounting for 39.48%; 40 institutions, 61 keywords, 30 sources, 30 themes, and 11 years. The entity scale presents a clear dual subject structure of papers and authors, with a relatively small number of institutions, sources, and topic nodes, but it undertakes cross paper connection and semantic constraint functions. If we only judge the information value based on the number of nodes, it is easy to underestimate the role of small-

scale entities in graph connectivity and task constraints.

The relationship set includes authored_by、affiliated_with、published_in、has_keyword、belongs_to_topic、published_in_year、cites The total number of relationships with the eight types of edges co_occursw_ith is 105148. 33742 citation relationships, accounting for 32.09%; 23979 keyword relationships, accounting for 22.80%; There are 14224 articles of authorship and affiliation with institutions, accounting for 13.53% respectively; 6979 thematic attribution relationships, accounting for 6.64%; 5200 articles on source publication and year relationship, accounting for 4.95% respectively; There are 1600 keyword co-occurrence relationships, accounting for 1.52%. The above relationship constitutes a structural network in which content, source, subject, and time participate together. The relationship between citations and keywords provides the main information channel, the relationship between author institutions and source themes provides identity and disciplinary boundaries, and the relationship between years is used to limit the time frame of thematic interpretation.

The weight of the relationship is calculated individually according to the kind of evidence. The authorship, publication source and publication year are classified as deterministic edges with a weight of 1; The keyword relationships are normalized considering the significance of the terms in the paper and the sparseness among different papers; The topic attribution relationship is decided by the strength of topic tags and the similarity of keywords; The co-occurrence relationships are normalized based on the frequency of keywords appearing in the same paper; The citation relationship maintains directed links for relationship completion and ordering. The relationship weight is defined as above.

$$w_{ij}^{(r)} = confidence_{ij}^{(r)} \cdot norm\left(freq_{ij}^{(r)}\right) \tag{1}$$

In the formula, $w_{ij}^{(r)}$ represents the edge weight of entity i and entity r on relationship type r, $confidence_{ij}^{(r)}$ represents the confidence of the relationship source, $freq_{ij}^{(r)}$ represents the frequency of relationship occurrence, and $norm(\cdot)$ represents the normalization function. The $confidence$ of the deterministic edge is set to 1, and the confidence of keywords, topics, and institutional relationships is given based on field integrity and normalization results. This design avoids all relationships being equally weighted, allowing high-frequency weak relationships and low-frequency strong relationships to have different impacts in the model.

Multimodal representation converges around paper nodes. The text vector is encoded from the title, abstract, and keywords; The citation vector consists of citation, citation, PageRank, and graph propagation embedding; The author's institutional vector is encoded by the collaborative network neighborhood and institutional distribution; The source topic vector is composed of journal, conference, and topic label distributions; The time vector is composed of the publication year and the intensity of the topic year. The definition of fusion representation is as follows:

$$h_p = \alpha_t h_p^t + \alpha_c h_p^c + \alpha_a h_p^a + \alpha_v h_p^v + \alpha_y h_p^y \tag{2}$$

In the formula, h_p is the fused representation of paper p, h_p^t, h_p^c, h_p^a, h_p^v, and h_p^y correspond to text, citation, author institution, source topic, and temporal modal representations, respectively. α_t, α_c, α_a, α_v, and α_y are modal weights. The weights are adjusted through the validation set and corresponding modalities are removed item by item in the ablation experiment to confirm the source of performance gain.

The relationship completion task uses a triplet rating of head entity, relationship type, and tail entity. Given candidate triplets (h, r, t), the model calculates the matching degree of the

head tail entity fusion vector in the relational space. This setting refers to the explicit modeling approach for relationship patterns in knowledge graph link prediction, especially the representation requirements for symmetric, antisymmetric, inverse, and combinatorial relationships [25].

$$s(h,r,t) = \sigma\left(h_h^T W_r h_t + b_r\right) \tag{3}$$

In the formula, $s(h,r,t)$ represents the probability of a triplet being formed, h_h and h_t represent the head and tail entities, W_r is the parameter matrix of relationship type r, b_r is the bias term, and σ is the Sigmoid function. Candidate tail entities are sorted by score and used to calculate MRR Hits@1 The Hits@3 and Hits@10 Figure 2 presents the mapping relationships between the graph pattern layer, instance layer, and five types of modalities.

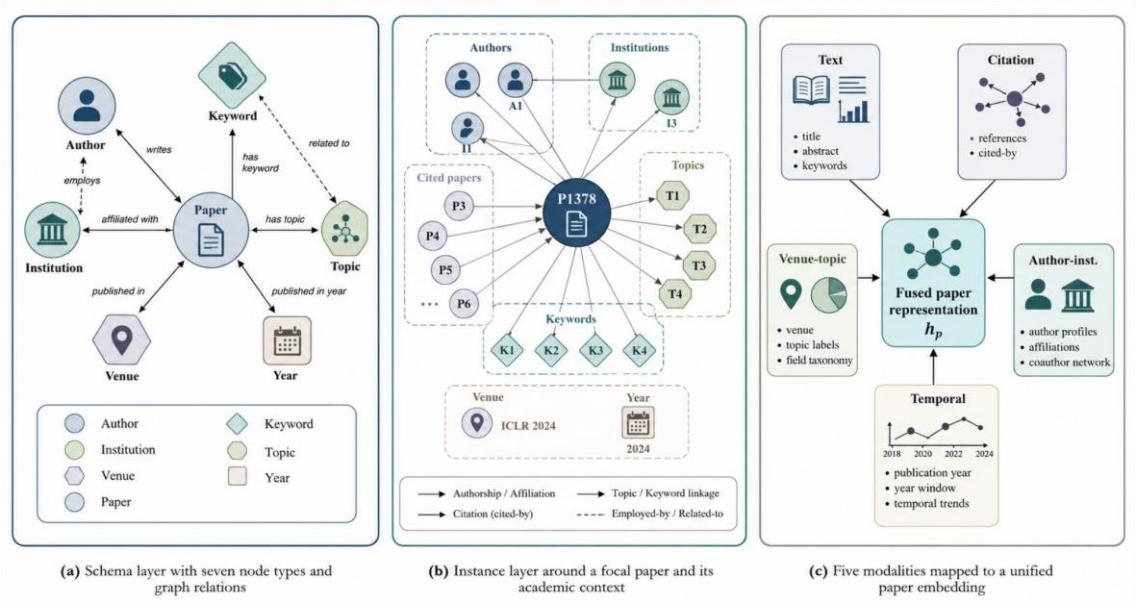

Figure 2: The Pattern Layer and Instance Layer of the Multimodal Knowledge Graph in Information Science

Figure 2 (a) describes the pattern constraints for seven kinds of entities and eight kinds of relationships, Figure 2 (b) displays the local network among authors, institutions, citations, keywords and topics for a single paper, and Figure 2 (c) indicates the integration of the five modes into a unified representation of the paper.

2.3 Experimental Tasks, Baseline Models, and Evaluation Protocol

Four types of tasks are set up in the experiment to avoid using only a single indicator to evaluate the quality of the graph. The entity alignment task verifies the standardization effect of authors, institutions, sources, and keywords, with metrics of Precision, Recall, and Macro-F1; The relationship completion task tests the edge prediction ability of the knowledge graph, with indicators such as MRR Hits@1 The Hits@3 and Hits@10 The cross modal retrieval task tests the matching ability between text queries, citation evidence, author institution clues, and topic tags, with the indicator being nDCG@1 The nDCG@3 The nDCG@10 And MRR; The theme mining task tests the ability to identify paper themes and identify annual changes in themes, with indicators of Macro-F1 and AUC. A total of 11078 training samples, 2426 validation samples, and 2296 testing samples were generated for the four types of tasks, with

fixed ratios of 70%, 15%, and 15% for training, validation, and testing.

The comparative models include Text only, Citation GCN, Late Fusion, MKG-BERT, and Proposed MDF-KG. Text only uses title, abstract, and keywords, making it suitable as a baseline for text field strength; Citation GCN only uses citation networks and graph convolution representations to examine the performance of structural relationships in the absence of textual semantics; Late Fusion encodes each modality independently and concatenates them to compare the differences between later concatenation and intra graph fusion; MKG-BERT connects text encoding with knowledge graph embedding, representing the common route of combining text deep encoding with graph vectors; Proposed MDF-KG integrates text, citations, author institutions, source themes, and temporal information at paper nodes, and completes sorting, completion, and classification in a unified graph space. The efficiency experiment retains the idea of lightweight graph propagation as a reference for computational cost, avoiding performance improvement based on unconstrained model complexity [26].

Sorting indicators are calculated based on the ranking of real objects among candidate objects. with nDCG@K For example, the model generates a ranking list based on the scores of candidate objects. The higher the correct object, the higher the cumulative loss gain

$$nDCG@K = \frac{DCG@K}{IDCG@K} \tag{4}$$

In the equation, $DCG@K$ Represents the cumulative loss gain of the first K candidate results, $IDCG@K$ Represents the maximum cumulative loss gain under ideal sorting. Entity alignment, relationship completion, and topic mining use fixed test partitioning to report final values, and cross modal retrieval provides additional curve results for K values of 1, 3, 5, 10, and 20. The efficiency experiment constructs a scale ladder based on 10000 to 120000 triplets, recording indexing time, training time, inference delay, memory usage, Macro-F1, and nDCG@10 The robustness experiment set up four types of perturbations: missing abstracts, missing citations, keyword noise, and missing institutions, with the noise ratio increasing from 0 to 0.5. Figure 3 shows the correspondence between tasks, baseline models, and evaluation interfaces.

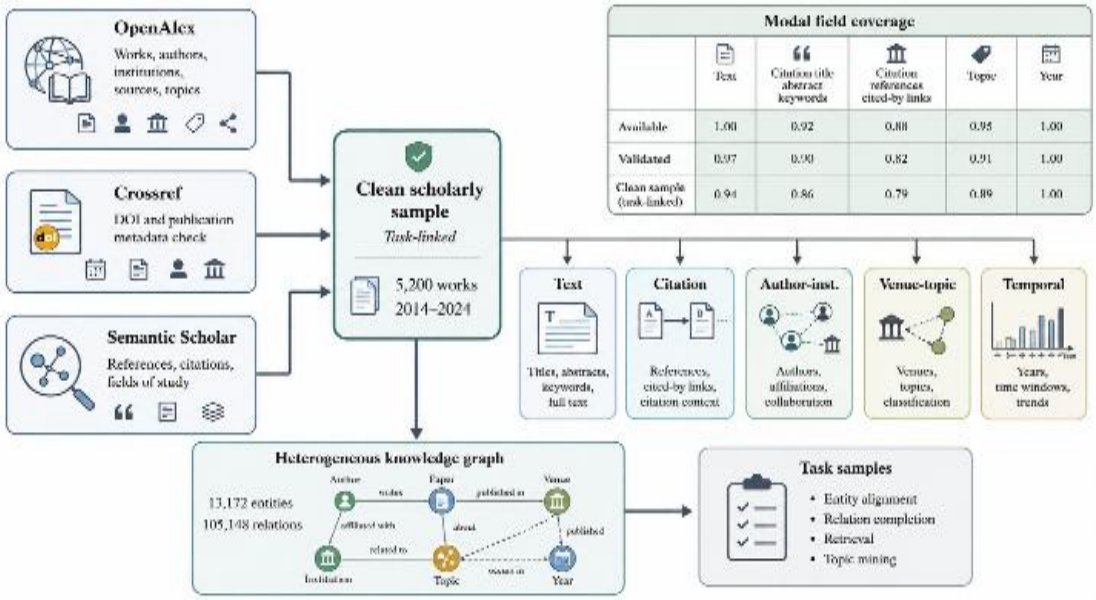

Figure 3: Experimental task, baseline model, and evaluation protocol

3 Results and Discussion

3.1 Knowledge Graph Construction Quality and Structural Characteristics

The quality of knowledge graph construction is first reflected in sample retention, entity structure, relationship types, and annual connectivity. After cleaning, 5200 papers were retained as samples, covering 11 consecutive years from 2014 to 2024. The annual number of papers increased from 283 in 2014 to 659 in 2024, the cumulative number of entities increased from 1154 to 11907, and the cumulative number of relationships increased from 5624 to 105148. This growth does not only mean an increase in the number of records, but also indicates a continuous increase in the number of citations, keywords, author institutions, and topic relationships that can be extracted between papers.

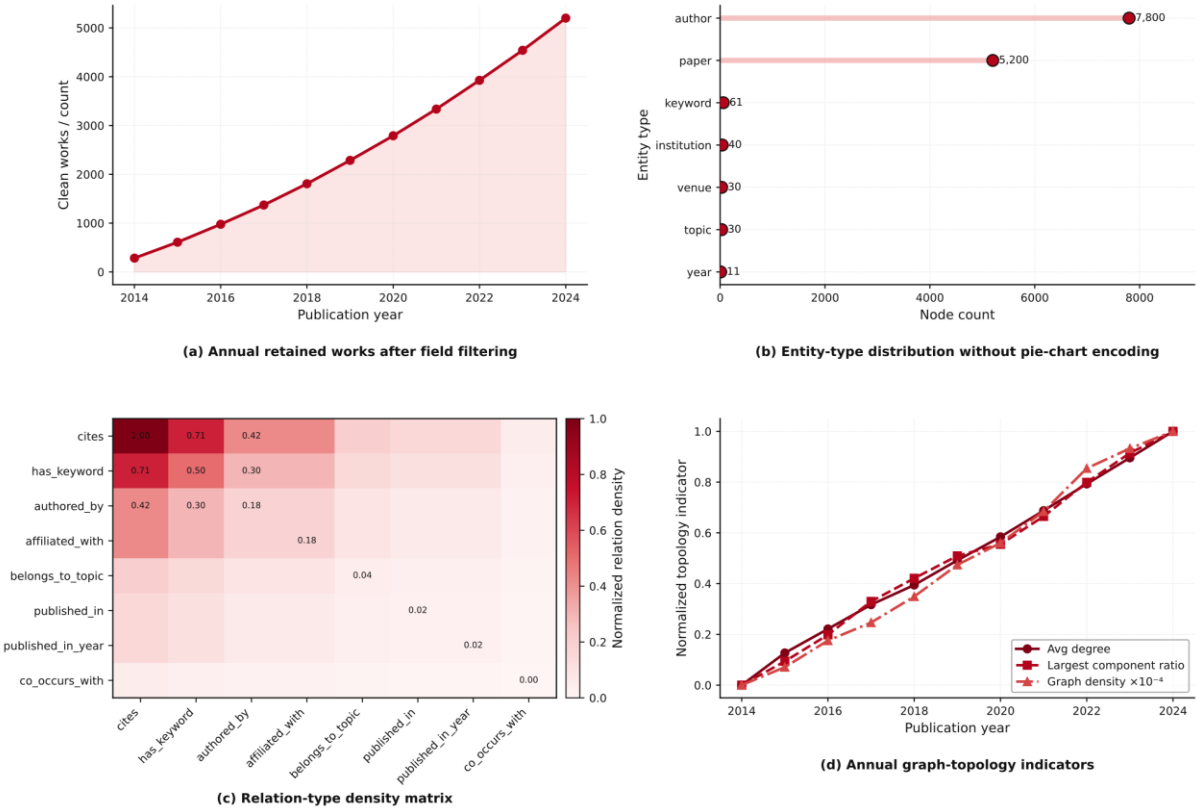

Figure 4: Quality and structural characteristics of knowledge graph construction

Figure 4 (a) shows that the annual retention curve has risen faster after 2020, which is consistent with the expansion of topics related to knowledge graphs, semantic retrieval, and open academic data.

The entity structure presents the dual center characteristics of the paper and the author. There are 7800 author nodes and 5200 paper nodes, accounting for 98.70% of the total entity count; The number of institutions, sources, keywords, themes, and year nodes is relatively small, but it determines whether the graph can surpass the text index at the level of a single paper. Institutional nodes are used to connect multiple papers of the same research subject, source nodes are used to constrain the publication space of disciplines, topic nodes are used for stable retrieval and tag mining, and year nodes are used to distinguish the development stages of topics. Figure 4 (b) presents the distribution of entity types using a dotted line graph

to avoid misjudgment of proportions caused by area or pie charts on small-scale entities.

The relationship distribution indicates that the information channels of the graph have hierarchical differences. 33742 citation relationships, which is the highest proportion of edge types; 23979 keyword relationships reflect the connection between the content of the paper and conceptual words; Author authorship and institutional affiliation are 14224 each, providing the research subject and organizational structure; Topic attribution, source publication, year relationship, and co-occurrence relationship bear semantic aggregation and time constraints. The density matrix of relationship types in Figure 4 (c) shows that citation, keyword, and author institution relationships constitute the main connecting skeleton. Although the number of source topic relationships is limited, they provide stable boundaries in topic mining and cross modal retrieval.

The annual topology index indicates the change of the graph from a sparse state to a connected one. The average degree rises from 9.747 in 2014 to 17.662 in 2024, the maximum connected component ratio increases from 0.569 to 0.942, and the graph density index increases from 1.968 to 6.467. Such modifications suggest that the isolated nodes are gradually combined through multiple mode relationships, allowing papers to reach the same knowledge field via different means. For data mining, these structural changes are very significant: entity alignment can use the author, institution and source information for cross-verification; relationship completion can take advantage of the citation and topic relationships to limit the candidate set; cross-modal retrieval can improve the ranking by utilizing the common neighborhood of text, citations and topics.

Table 2: Annual Structure Indicators of Graph

Year	Number of Papers	Number of Entities	Number of Relationships	Average Degree	Maximum Connected Component Ratio	Graph Density × 10^{-4}
2014	283	1154	5624	9.747	0.569	1.968
2016	977	3411	19611	11.499	0.643	2.758
2018	1807	5665	36455	12.870	0.726	3.539
2020	2790	7852	56421	14.371	0.776	4.488
2022	3926	9903	79333	16.022	0.867	5.811
2024	5200	11907	105148	17.662	0.942	6.467

3.2 Mining Performance Across Retrieval, Completion, and Topic Tasks

After the graph structure is confirmed, the model performance needs to be tested through task-based metrics. Table 3 shows that the Macro F1 scores of Proposed MDF-KG for entity alignment, relationship completion, cross modal retrieval, and topic mining tasks are 0.858, 0.873, 0.836, and 0.847, respectively, with an average of 0.8535. The average Macro-F1 of MKG-BERT is 0.8213, Late Fusion is 0.7880, Text only and Citation GCN are 0.6963 and 0.6980, respectively. Compared with the optimal comparison model MKG-BERT, the proposed MDF-KG has Macro-F1 increments of 0.031, 0.031, 0.031, and 0.036 on four types of tasks, which are not concentrated on a single task, indicating that the five modalities provide effective evidence in different mining scenarios.

Table 3: Comparison of Model Performance for Four Types of Tasks

Model	Entity Alignment Macro-F1	Relation Completion Macro-F1	Cross-Modal Retrieval Macro-F1	Topic Mining Macro-F1	Average Macro-F1
Text-only	0.764	0.612	0.668	0.741	0.6963
Citation-GCN	0.692	0.724	0.695	0.681	0.6980
Late Fusion	0.806	0.803	0.761	0.782	0.7880
MKG-BERT	0.827	0.842	0.805	0.811	0.8213
Proposed MDF-KG	0.858	0.873	0.836	0.847	0.8535

In the entity alignment task, the Macro-F1 of Text only is 0.764, Citation GCN is 0.692, Late Fusion is 0.806, MKG-BERT is 0.827, and the proposed MDF-KG obtains 0.858. These results show that the standardization of authors, institutions, sources and keywords cannot be based only on the titles and abstracts. Similar texts in papers may use different author abbreviations or translations of institutions, and the citation network may not be able to identify the entity due to incomplete citation coverage. The fusion model integrates the text semantics, author institution neighbourhood and source topic information into the same representation space, resulting in an MRR of 0.898 and Hits@10 reaching 0.945, which facilitates the correct entity to appear among the top candidates.

In the relationship completion task, the Macro-F1 of Proposed MDF-KG is 0.873, and the MRR is 0.912, Hits@10 It is 0.956, which is 0.054 higher than MKG-BERT's MRR of 0.858. This task requires higher structural evidence, and simple text similarity may connect synonymous topic papers together, but cannot determine the specific type of relationship; The Macro-F1 of Citation GCN is 0.724, indicating that citation structure can improve relationship prediction, but weak related papers are easily included as candidates when text and topic constraints are lacking. The advantages of the fusion model come from relationship space scoring: citation relationships provide candidate edges, topic and source relationships limit edge types, text vectors supplement semantic similarity, and author institution information corrects local biases caused by the same research subject.

The cross modal retrieval task can better reflect the practical value of multimodal fusion, as shown in Figure 5.

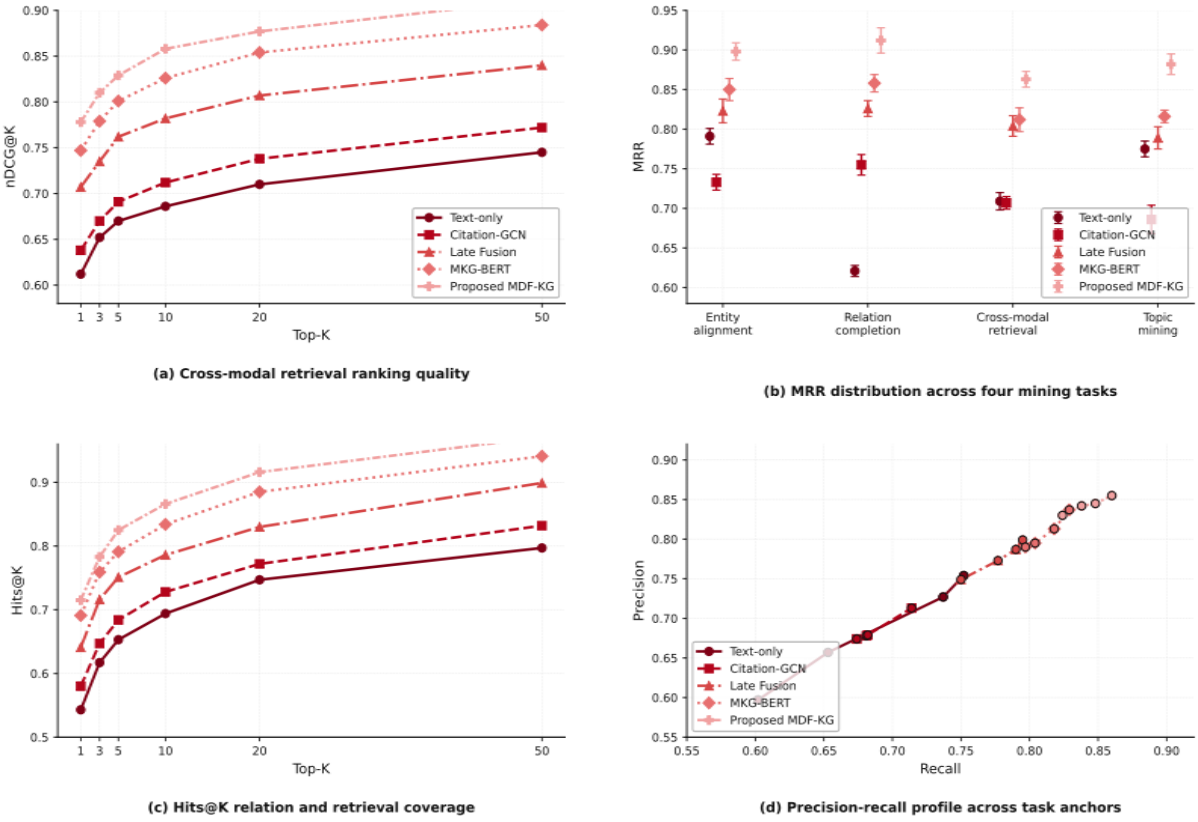

Figure 5: Comparison of Cross modal Retrieval and Relationship Completion Performance

Figure 5 (a) shows the proposed MDF-KG at K=10 nDCG@10 At 0.858, it is higher than MKG-BERT's 0.826, Late Fusion's 0.782, Citation GCN's 0.712, and Text only's 0.686. The MRR distribution in Figure 5 (b) shows that the MRR of Proposed MDF-KG in cross modal retrieval is 0.863, indicating that the matching results between the query object and the target object are more concentrated at the forefront. In Figure 5 (c), Proposed MDF-KG Hits@10 Reaching 0.866, higher than MKG-BERT's 0.834. Figure 5 (d) shows that after the recall rate is increased, the fusion model still maintains a high accuracy, which is suitable for the application requirement of "expanding the candidate range first and then maintaining the ranking quality" in scientific and technological intelligence retrieval.

In the theme mining task, the Macro-F1 of Proposed MDF-KG is 0.847 and the AUC is 0.900. The Macro-F1 of Text only on this task is 0.741, which is higher than the 0.681 of Citation GCN, indicating that topic recognition still relies on title, abstract, and keyword semantics. Late Fusion and MKG-BERT reached 0.782 and 0.811, respectively, indicating that displaying structural information can further enhance theme boundary recognition. The reason for the continued improvement of Proposed MDF-KG is that the temporal and source thematic modalities have a constraining effect on the thematic stage. The same keywords may correspond to different research questions in different years, and the same paper topic may also present different disciplinary orientations in journals from different sources. The fusion model reduces the misjudgment caused by this granularity difference through graph neighborhood.

3.3 Ablation, Robustness, Efficiency, and Error-Source Analysis

The ablation experiment is used to determine which modes the performance improvement comes from, as shown in Figure 6.

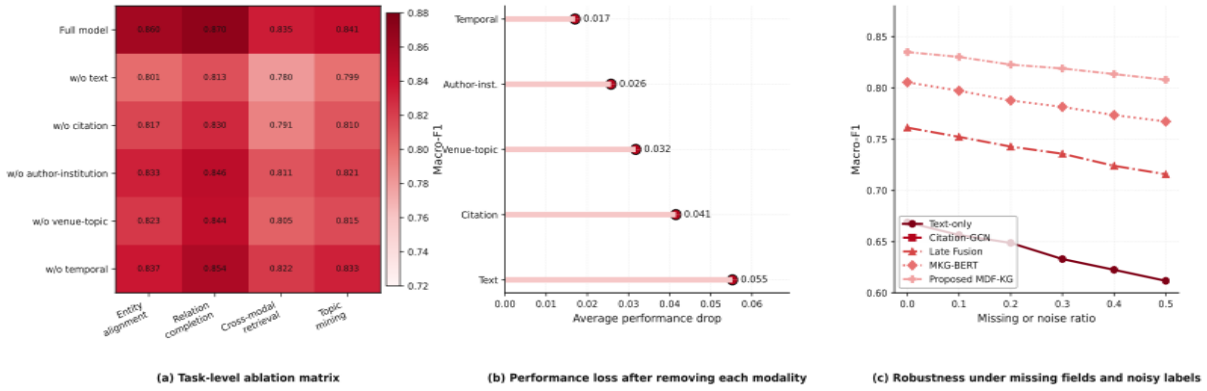

Figure 6: Multi modal module ablation experiment

Figure 6 (a) shows that the complete model maintains the highest or near highest metrics on all four types of tasks. The average performance loss after deleting the text modality is 0.0553, which is the highest among all modalities; The loss after deleting the citation mode is 0.0415; The losses after deleting the source theme, author institution, and time modality are 0.0318, 0.0258, and 0.0170, respectively. This result indicates that the text modality remains the core evidence for paper topic recognition and entity normalization, while the citation modality contributes more to relationship completion and cross modal retrieval. The source topic and author institution modality mainly provide boundary constraints and alias correction, while the direct contribution of the time modality is relatively small, but it plays a complementary role in topic evolution and response surface stability.

The robustness results showed that when the field missing or noise ratio reached 0.5, the average Macro-F1 of Text only decreased to 0.6118, Late Fusion was 0.7158, MKG-BERT was 0.7673, and Proposed MDF-KG remained at 0.8080. The average performance of Proposed MDF-KG decreased to 0.0280, lower than MKG-BERT's 0.0378, Late Fusion's 0.0453, and Text only's 0.0563. This difference indicates that the fusion model is insensitive to missing single fields. When the abstract is missing, the model can still utilize keywords, source topics, and citation neighborhoods; When citations are missing, the relationship between the text and the author's institution can still provide candidate restrictions; When institutions are missing, source themes and time windows can supplement semantic constraints. The decrease in the fusion model curve in Figure 6 (c) is relatively small, reflecting the anti-interference ability of multimodal redundant evidence.

The parameter response surface indicates that there is an appropriate interval for the fusion weight and graph propagation depth, as shown in Figure 7.

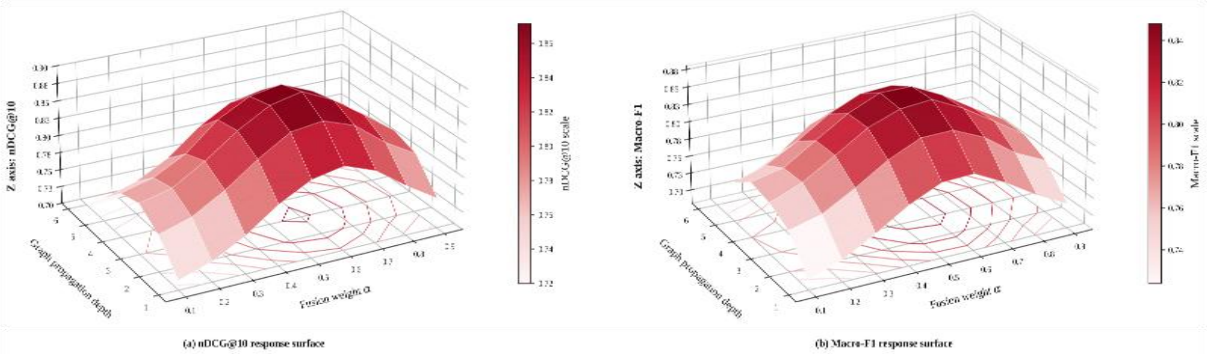

Figure 7: Three dimensional response surface integrating weights and graph propagation depth

In Figure 7 (a), nDCG@10 When the fusion weight α=0.6 and the graph propagation depth is 3, the highest value of 0.881 is reached; In Figure 7 (b), Macro-F1 reaches 0.850 around the same parameters. When alpha is too low, there is insufficient textual and local semantic evidence, and the model relies more on structural neighborhoods, making it susceptible to weak citations and cross topic citations; When alpha is too high, the structural information is weakened, making it difficult for topic and source boundaries to participate in sorting. After the depth of graph propagation exceeds 4, the response surface shows a decline, indicating that excessive propagation will push adjacent topics and weakly related citations into the same representation space, resulting in semantic smoothing. This result provides parameter intervals for subsequent deployment, rather than just providing the optimal value for a single point.

Efficiency analysis shows that Proposed MDF-KG maintains a good balance between performance and computational cost, as shown in Figure 8.

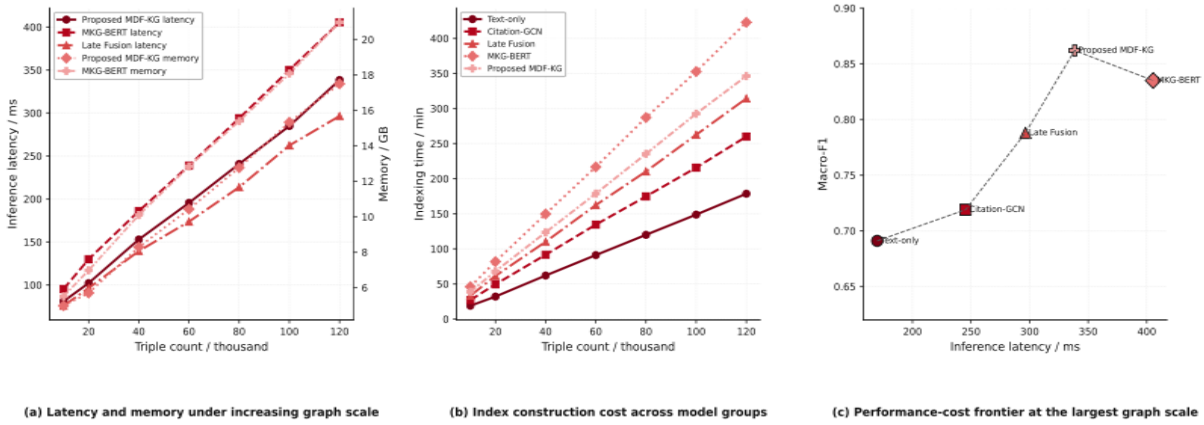

Figure 8: Efficiency and Performance Boundaries under Graph Scale Expansion

When the number of triples is expanded to 120000, the indexing time of Proposed MDF-KG is 346.30 min, the training time is 559.19 min, the inference delay is 338.43 ms, the memory usage is 17.50 GB, and Macro-F1 is 0.862, nDCG@10 It is 0.876. The inference delay of MKG-BERT at the same scale is 405.50 ms, the memory usage is 20.97 GB, and Macro-F1 is 0.835. Text only has the lowest latency, only 169.85 ms, but Macro-F1 has a performance of 0.691, which is not sufficient to support complex relational tasks. The performance cost boundary in Figure 8 (c) shows that the Proposed MDF-KG is located in the high-performance and cost lower range than MKG-BERT, suitable for medium-sized intelligence monitoring and digital library semantic retrieval scenarios.

Error source analysis can point out the unsolved problems in the model. Among the 600 error cases, there are 109 OCR label noises and entity alias conflicts, 101 abstract truncations, 100 topic granularity mismatches, 92 venue normalizations, and 89 weak citation contexts. In the repair plan, the repair ratio for venue normalization is 0.7609, abstract expansion is 0.7228, keyword denoising is 0.7156, topic re-clustering is 0.7000, citation expansion is 0.6629, and alias merging is 0.6055. The repair ratio of entity alias conflicts is the smallest, but the average score after repair has improved from 0.5448 to 0.7106, suggesting that author and institution standardization is still the main factor influencing the accuracy of entity matching. The weak citation context repair ratio is 0.6629, which indicates that merely expanding the citation neighborhood cannot fully resolve the problem of ambiguous citation meanings, and finer reference context is required.

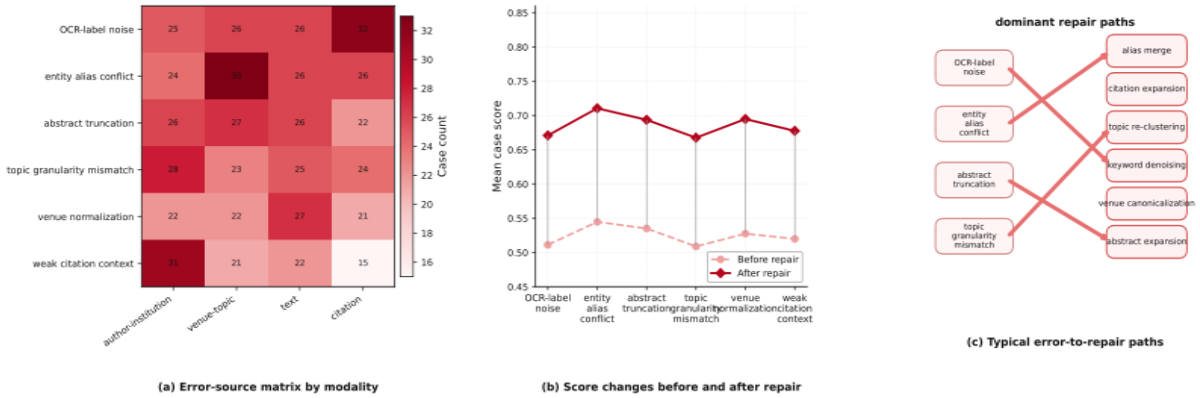

Figure 9: Error Source and Repair Effect

The error type matrix in Figure 9 (a) shows that different sources of errors are not evenly distributed. OCR label noise has a greater impact on the source topic and text fields, entity alias conflict focuses on the author institution modality, and weak citation context is related to both citation and author institution relationships. Figure 9 (b) shows that the scores of various types of errors have improved after repair, with venue normalization increasing by 0.1673, entity alias conflict increasing by 0.1658, and OCR label noise increasing by 0.1597. Figure 9 (c) further illustrates that error repair is not a single post-processing action, but rather follows paths such as entity normalization, topic reclassification, citation expansion, and keyword denoising. For actual intelligence systems, model deployment cannot only focus on average performance, but also needs to retain error source records for continuous validation when new data enters.

Table 4: Performance loss in ablation experiments

Ablation Setting	Removed Modality	Entity Alignment	Relation Completion	Cross-Modal Retrieval	Topic Mining	Average Performance Loss
w/o text	text	0.057	0.060	0.056	0.048	0.0552
w/o citation	citation	0.041	0.043	0.045	0.037	0.0415
w/o author-institution	author-institution	0.025	0.027	0.025	0.026	0.0258
w/o venue-topic	venue-topic	0.035	0.029	0.031	0.032	0.0318
w/o temporal	temporal	0.021	0.019	0.014	0.014	0.0170

4 Conclusion

A multimodal knowledge graph empirical sample package for open academic metadata field structure has been established to solve the problems of multi-source heterogeneity, missing fields and inconsistent semantic granularity in information science. The fusion modeling effect has been evaluated in four tasks: entity alignment, relationship completion, cross modal retrieval and topic mining. The sample includes 5200 papers, 13172 entities and 105148 relationships from 2014 to 2024, offering a consistent experimental method for knowledge organization and data mining in information science.

(1) At the data organization level, several entity structures have been established, such as papers, authors, institutions, sources, keywords, themes and years, and text, citations, author institutions, source themes and temporal information have been integrated into the same graph

space. The ratio of the largest connected components in the graph has risen from 0.569 in 2014 to 0.942 in 2024, suggesting that multi-modal relationships can improve the overall connectivity of the samples.

(2) At the level of methods and results, the proposed MDF-KG showed an average Macro-F1 of 0.8535 on four types of tasks, which is 0.0323 higher than MKG-BERT; Cross-modal retrieval nDCG@10 Reached 0.858, relationship completion MRR reached 0.912. The ablation results indicate that text and citations are the main contributing modalities, while source themes, author institutions, and temporal modalities mainly improve boundary constraints, robustness, and error interpretation.

(3) The research is still affected by the coverage of sample fields and the quality of open source data. Author aliases, institutional abbreviations, inconsistent topic granularity, and weak citation context can still cause mismatches. In the future, more complete online metadata, ORCID standardized information, open full-text paragraphs, and finer grained citation contexts can be accessed to further test the deployment performance of the model in real intelligence monitoring and digital library retrieval systems.

About the Author

Li Ming was born in Changzhi, Shanxi Province in 1978. He obtained his bachelor's and master's degrees from Hohai University and his doctorate from Nanjing University of Science and Technology. Currently, he is an associate professor and master's supervisor at the Business School of Hohai University. His main research areas include project management, information system development and management, information resource development and management, and digitalization and intelligence. lm@hhu.edu.cn

Xu Yongjia was born in Taiyuan, Shanxi Province in 2001. She obtained her bachelor's degree from Anhui University of Science and Technology. Currently, she is studying at the Business School of Hohai University. Her main research interests lie in information fusion and data mining. xuyongjia0427@163.com

References

[1] Priem J; Piwowar H; Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv, 2022: arXiv:2205.01833. DOI: 10.48550/arXiv.2205.01833.

[2] Hendricks G; Tkaczyk D; Lin J; et al. Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies, 2020, 1(1): 414-427. DOI: 10.1162/qss_a_00022.

[3] Lo K; Wang L L; Neumann M; et al. S2ORC: The Semantic Scholar Open Research Corpus. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 4969-4983. DOI: 10.18653/v1/2020.acl-main.447.

[4] Färber M; Lamprecht D; Krause J; et al. SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples. The Semantic Web: ISWC 2023, Lecture Notes in Computer Science, 2023, 14265: 94-112. DOI: 10.1007/978-3-031-47243-5_6.

[5] Jaradeh M Y; Oelen A; Farfar K E; et al. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. Proceedings of the 10th

International Conference on Knowledge Capture, 2019: 243-246. DOI: 10.1145/3360901.3364435.

[6] Verma S; Bhatia R; Harit S; et al. Scholarly knowledge graphs through structuring scholarly communication: A review. Complex & Intelligent Systems, 2023, 9(1): 1059-1095. DOI: 10.1007/s40747-022-00806-6.

[7] Manghi P; Mannocci A; Bardi A; et al. Challenges in building scholarly knowledge graphs for research assessment in open science. Quantitative Science Studies, 2024, 5(4): 991-1021. DOI: 10.1162/qss_a_00322.

[8] Zhong L; Wu J; Li Q; et al. A comprehensive survey on automatic knowledge graph construction. ACM Computing Surveys, 2024, 56(4): Article 94. DOI: 10.1145/3618295.

[9] Hogan A; Blomqvist E; Cochez M; et al. Knowledge Graphs. ACM Computing Surveys, 2021, 54(4): Article 71. DOI: 10.1145/3447772.

[10] Peng C; Xia F; Naseriparsa M; et al. Knowledge Graphs: Opportunities and Challenges. Artificial Intelligence Review, 2023, 56(11): 13071-13102. DOI: 10.1007/s10462-023-10465-9.

[11] Shi C; Li Y; Zhang J; et al. A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 17-37. DOI: 10.1109/TKDE.2016.2598561.

[12] Wang Q; Mao Z; Wang B; et al. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743. DOI: 10.1109/TKDE.2017.2754499.

[13] Hamilton W L; Ying R; Leskovec J. Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 2017, 30: 1024-1034.

[14] Schlichtkrull M; Kipf T N; Bloem P; et al. Modeling relational data with graph convolutional networks. The Semantic Web: ESWC 2018, Lecture Notes in Computer Science, 2018, 10843: 593-607. DOI: 10.1007/978-3-319-93417-4_38.

[15] Veličković P; Cucurull G; Casanova A; et al. Graph Attention Networks. International Conference on Learning Representations, 2018. arXiv: 1710.10903.

[16] Devlin J; Chang M W; Lee K; et al. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 2019: 4171-4186. DOI: 10.18653/v1/N19-1423.

[17] Reimers N; Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, 2019: 3982-3992. DOI: 10.18653/v1/D19-1410.

[18] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Neural Computing and Applications, 2022, 34: 22073-22094. DOI: 10.1007/s00521-022-07533-4.

[19] Blei D M; Ng A Y; Jordan M I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3: 993-1022.

[20] van Eck N J; Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 2010, 84(2): 523-538. DOI: 10.1007/s11192-009-0146-3.

[21] van Eck N J; Waltman L. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. Journal of Informetrics, 2014, 8(4): 802-823. DOI: 10.1016/j.joi.2014.07.006.

[22] Newman M E J. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 2006, 103(23): 8577-8582. DOI: 10.1073/pnas. 0601602103.

[23] Page L; Brin S; Motwani R; et al. The PageRank citation ranking: Bringing order to the Web. Stanford InfoLab Technical Report, 1999: 1999-66.

[24] Järvelin K; Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 2002, 20(4): 422-446. DOI: 10.1145/582415. 582418.

[25] Sun Z; Deng Z H; Nie J Y; et al. RotatE: Knowledge graph embedding by relational rotation in complex space. International Conference on Learning Representations, 2019. arXiv: 1902.10197.

[26] He X; Deng K; Wang X; et al. LightGCN: Simplifying and powering graph convolution network for recommendation. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020: 639-648. DOI: 10.1145/3397271.3401063.