Semantic Similarity Analysis of Scientific Papers in Scholarly Knowledge Graphs

Thu Huong Nguyen[*,†], Cédric Pruski[*,†] and Marcos Da Silveira[*,†]

Luxembourg Institute of Science and Technology, 5 avenue des hauts-fourneaux, L-4362 Esch sur Alzette, Luxembourg

Abstract

Scholarly Knowledge Graphs (SKGs) structure academic information, enabling research discovery and knowledge synthesis. However, detecting their structural and semantic evolution remains challenging due to contextual variations and implicit knowledge shifts. In this paper, we introduce MOKA, a framework that integrates bibliometric data, advanced Natural Language Processing, and Agentic Retrieval-Augmented Generation to characterize SKG evolution. As part of this process, we propose a method to measure semantic similarity between papers by refining the representation of documents and applying four pretrained language models. Our approach assesses semantic similarity across different textual granularity levels, including full papers, abstracts, and the ORKG metadata. Our experimental results demonstrates that the similarity scores highly depend on the combination of language model and the evoked dimensions of the papers to evaluate.

Keywords

Scholarly Knowledge Graphs, Semantic Similarity, Knowledge Graph Evolution.

1. Introduction

In recent years, numerous Scholarly Knowledge Graphs (SKGs) have emerged, employing diverse approaches to structuring research and scientific data. Bibliographic and citation-centric SKGs [1, 2, 3, 4] primarily align publications through citations and metadata, such as authors, institutions, and domains. In contrast, content-oriented SKGs, such as CS-KG[1] [5] and the ORKG[2] (Open Research Knowledge Graph [6]), focus on capturing scientific knowledge by detailing research problems, methods, and findings.

As scientific output expands and new disciplines emerge, SKGs must evolve both *structurally* and *semantically* to keep pace with academic research. *Structural evolution* refers to changes in the SKG's network over time, observed through the continuous addition of nodes (representing newly published articles) and edges (capturing relationships like citations, co-authorship, and venues). This evolution is particularly prominent in bibliographic and citation-focused SKGs, reflecting the ever-growing body of scholarly work. Tracking these changes offers valuable insights into emerging research trends, influential studies, academic collaborations, and funding patterns, all of which shape the trajectory of scientific progress.

Semantic evolution, on the other hand, pertains to shifts in academic ideas within SKGs. In content-focused SKGs like presented in the ORKG, we can obtain comparison tables showing studies that introduce methods, findings, and concepts that redefine existing knowledge. Unlike structural changes, detecting semantic evolution is more complex, as it requires not only identifying modifications in relationships but also interpreting the evolving meanings of concepts within a dynamic research landscape. This demands sophisticated Natural Language Processing (NLP) techniques capable of capturing nuances in context, terminology, and evolving discourse.

2nd International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2025), co-located with ESWC 2025, June 01–02, 2025, Portorož, Slovenia

[*] Corresponding author.

[†] These authors contributed equally.

thuhuong.nguyen@list.lu (T. H. Nguyen); cedric.pruski@list.lu (C. Pruski); marcos.dasilveira@list.lu (M. D. Silveira)
0009-0001-3300-8581 (T. H. Nguyen); 0000-0002-2103-0431 (C. Pruski); 0000-0002-2604-3645 (M. D. Silveira)

[1] http://w3id.org/cskg

[2] https://ORKG.org/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Recent advancements in Artificial Intelligence (AI), particularly in generative AI (GenAI) and the development of Large Language Models (LLMs), have substantially improved the ability of computer systems to analyze and interpret semantic evolution such as approaches, methodologies, or goals. Trained on vast amounts of textual data, these models enable more precise detection of related (but with some variations) works, offering deeper insights into knowledge evolution within domains over a period of time. However, detecting semantic evolution in SKGs remains particularly difficult due to challenges such as contextual understanding, concept disambiguation, and knowledge synthesis. Effective identification of approaches evolution requires not only tracking explicit changes in terminology but also capturing implicit shifts in meaning driven by new discoveries, interdisciplinary interactions, and emerging research trends.

Addressing these challenges necessitates the integration of advanced AI-driven techniques with domain-specific expertise to enhance the accuracy and reliability of semantic change detection in SKGs. By refining methods for analyzing knowledge evolution, researchers can benefit from effective tools for tracking scientific progress, fostering interdisciplinary collaboration, and supporting data-driven decision-making in academia. In this paper, we deal with the challenge of Scholarly Knowledge Graph (SKG) evolution, a complex problem that requires the integration of multiple AI-driven techniques. Managing this evolution effectively demands a comprehensive approach that goes beyond structural updates, incorporating advanced methods for detecting semantic evolution in scholarly content.

To tackle this issue, we introduce MOKA, our view of a mid-term framework designed to capture both structural and semantic evolution in SKGs. MOKA integrates three complementary techniques: *Natural Language Processing* (NLP) technique for in-depth content analysis, with *Agentic Retrieval-Augmented Generation* (Agentic-RAG) technique [7] to dynamically formulate queries and plans, extract and complete missing data, detect novelties, and *Knowledge graphs* (KGs) to represent the evolution of domain knowledge. MOKA enables a more efficient and automated interrogation, analysis, and synthesis of diverse information sources. In this paper, we mainly focus on the *Semantic Similarity Analysis* of MOKA framework, providing a detailed description on how to identify if two papers are 'related'. Our approach involves harmonizing document representations and employing language models [8] to assess their content similarity [9]. To validate our approach, we conduct an extensive experimental evaluation, comparing semantic similarity measurements across different levels of textual granularity, including full papers, titles and abstracts, and the ORKG metadata. A key component of our approach is its integration with the comparison table feature of the ORKG infrastructure [10]. This feature provides structured semantic descriptions of research articles, systematically organizing key attributes such as objectives, methodologies, and findings. By leveraging this structured representation, MOKA enhances the discovery of related work and facilitates meaningful comparisons between research contributions. Our mid-term goal is to automate the identification of 'evolutionary relations' between papers, a process that currently requires a laborious manual effort.

The remainder of the paper is structured as follows: Section 2 presents related work of the field semantic similarity between scientific papers. Section 3 presents our general approach for identifying evolution relationships between scientific articles. More details of the Semantic Similarity Analysis approach is given in Section 4. Section 5 proposes an experimental assessment of our method. Section 6 wraps up with concluding remarks and outlines future work.

2. Related Work

Several recent studies have focused on automatically detecting relationship between scientific publications. One prominent approach is to use natural language generation techniques to produce citation texts that summarize and contextualize these relationships. Luu et al. [11] operationalize this task by using citing sentences as a proxy for scientific relationships and train a large language model (LLM) to generate relationship-explaining text. Their approach involves pre-training a large language model as a foundation for autoregressive approaches and explores different perspectives on the documents, using dense representations extracted with scientific information extraction systems. Li and Ouyang [12, 13] extend this concept, by leveraging feature-based, LLM-prompting approaches to generate richer citation-texts that capture complex interconnection among multiple papers. Their method extracts features from local citation network, incorporate them into a prompt, and generate citation paragraphs enriched with transition sentences, making citation texts more cohesive and interpretable. Similarly, Xing et al. [14] propose a multi-source pointer-generator network with cross-attention to automatically generate citation texts, demonstrating that such models can synthesize meaningful scholarly relationships. In addition, researchers have explored graph-based methods for analyzing the evolution of scientific knowledge. Dalle Lucca Tosi & Dos Reis [15] introduce a concept-based evolution tracking approach, using knowledge graphs to identify and compare scientific subfields over time. Similarly, Rossanez et al. [16] employ temporal knowledge graphs to model and analyze knowledge evolution within unstructured scientific corpora. Aparicio et al. [17] leverage dynamic knowledge graphs to extract emerging research trends by analyzing evolving knowledge communities, while [18] applies knowledge graph embedding techniques to enhance citation recommendation systems, addressing limitations in static models.

Other works focus on forecasting the impact of scientific research. Gu & Krenn [19] construct an evolving knowledge graph of over 21 millions scientific papers, integrating semantic and citation networks to predict the future impact of emerging research ideas. Salatino et al. [20] take a bibliometric-driven approach, using ontology-based research topic modeling to analyze and forecast trends in research dynamics. Meanwhile, [21] introduces the AIDA knowledge graph, which characterizes research topics and industrial sectors, providing detailed insights into academia-industry knowledge transfer.

Beyond citation text generation, various approaches have been explored to measure semantic similarity between scientific papers. These methods range from traditional vector-space models to deep learning-based techniques, each offering distinct advantages and limitations. Early approaches to measuring document similarity rely on vector-space models such as TF-IDF [22], Latent Semantic Analysis (LSA) [23]. TF-IDF ranks documents based on term frequency while down-weighting common words, making them effective for lexical matching but inadequate for capturing deeper semantic relationships. LSA, on the other hand, applies singular value decomposition (SVD) to uncover latent relationships between words and documents, improving synonym recognition but suffering from interpretability issues and high computational cost. These models work well for keyword-based retrieval but struggle with polysemy, paraphrasing, and contextual understanding, limiting their applicability in semantic similarity tasks.

To address the limitations of traditional models, word embeddings such as Doc2Vec [24], fastText [25] or GloVe [26] have been introduced. These techniques represent words as dense, continuous vectors in a high-dimensional space, capturing semantic relationships through contextual co-occurrence patterns. While effective in preserving word-level semantics, these models lack sentence-level understanding and struggle with out-of-vocabulary words, particularly in the ever-evolving landscape of scientific terminology. Recent advances in deep learning have led to the development of sentence embeddings, which extend word embeddings to entire sentences or documents. One of the most prominent models in this category is SBERT (Sentence-BERT) [27], which fine-tunes BERT-based models using Siamese and triplet network architectures to generate meaningful sentence-level embeddings. SBERT significantly outperforms traditional word embeddings by capturing contextual meaning and sentence semantics, making it well-suited for scientific text similarity tasks. However, while SBERT excels at capturing semantic nuances, it requires computationally expensive fine-tuning and may struggle with domain-specific jargon if not properly trained on scholarly corpora.

An emerging trend in document similarity assessment is the integration of generative AI, including retrieval-augmented generation (RAG) [28] and large language models (LLMs) [29], to enhance scholarly knowledge extraction and relationship modeling. RAG combines information retrieval with generative capabilities, retrieving contextually relevant scientific text from large corpora and incorporating it into the generation or analysis of scholarly relationships. This approach helps bridge gaps in missing contextual information, particularly for sparsely cited or newly published papers. However, these techniques are computationally intensive, prone to hallucination, and rely on well-curated external knowledge sources to ensure accuracy and reliability. Despite these challenges, generative AI-driven approaches hold great promise in advancing automated literature review, citation recommendation, and knowledge graph enrichment, making them valuable tools for scientific research discovery.

3. The MOKA framework

We propose the MOKA framework, illustrated in Fig. 1, based on an Agentic-RAG architecture [30, 7] to dynamically identify and explicitly represent the evolutionary relationships between scientific publications described in the ORKG. This approach leverages LLM advancements, incorporating sequential processes such as temporal analysis, citation tracking, and semantic analysis of abstracts and full texts. These steps provide context augmentation, improving the discovery of evolutionary links between research contributions within the ORKG.

In this framework, Agentic-RAG acts as an orchestrator, planning complex information flows and enabling flexible task execution across multiple data sources, including the ORKG, citation databases, and abstract/full-text repositories. The model capitalizes on the reasoning capabilities of ReAct Agents [31], embedded within the Agentic-RAG framework, to execute retrieval and analysis tasks efficiently.

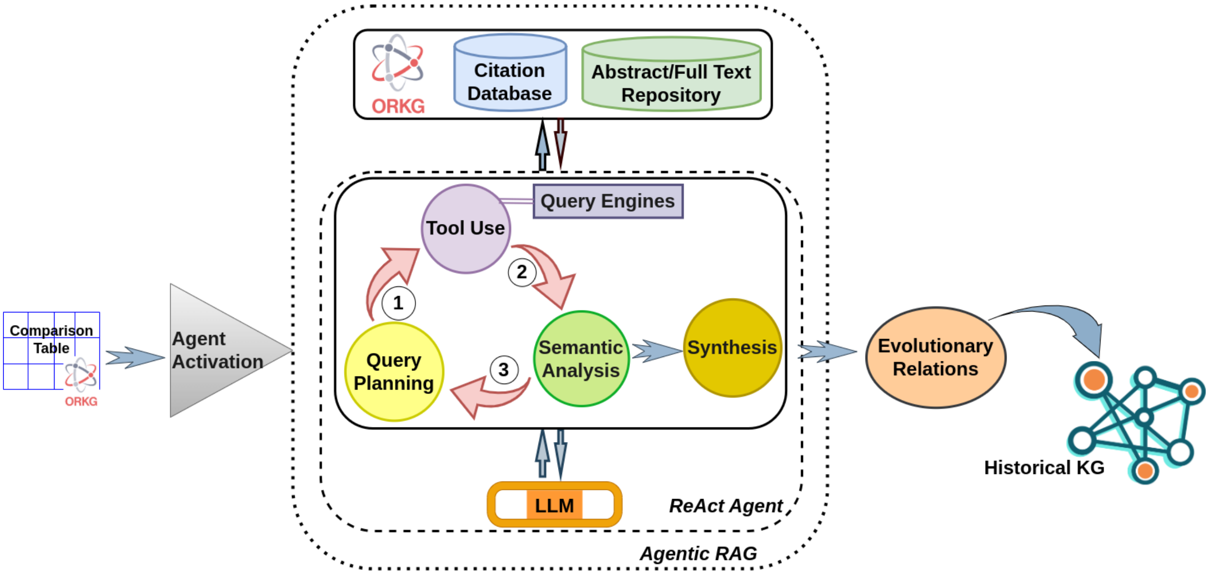

Figure 1: MOKA framework for discovering semantic evolutionary relationships in the ORKG

The process is initiated whenever a new comparison table related to a specific research topic is added or updated in the ORKG. This triggers an agent-based reasoning process, structured as follows:

- *Query Planning (QP)*: The agent, assisted by a Large Language Model (LLM), formulates an exploration strategy using the existing comparison table as input. The exploration task is decomposed into specific retrieval objectives, which are assigned to Query Engines (QE) within the Tool Use framework for efficient data acquisition.
- *Citation Tracking & Data Retrieval*: The Citation Tracking QE connects to citation databases, identifying citation relationships between research papers. This step enriches the dataset with explicit citation links, uncovering potential evolutionary relationships and preparing for semantic analysis.
- *Semantic Analysis*: The agent dynamically adapts its reasoning to infer semantic relationships, independent of direct citations. The Semantic Analysis QE processes targeted text segments, including: Contextual snippets surrounding citations in full texts, abstracts of related studies, and the ORKG metadata, capturing both explicit and inferred connections. The agent performs paired analyses between the new contribution and each related study using:

  - *Semantic Similarity Analysis* [32, 33] to quantify content alignment, and
  - *Textual Entailment Analysis* [34] to detect conceptual dependencies and knowledge transfer.

  This phase operates iteratively, refining retrieval strategies within QP through a continuous feedback loop. The agent adjusts retrieval parameters to enhance relevance, ensuring that only the most contextually aligned studies are retained for synthesis.
- Synthesis & Evolution Mapping. In the final phase, the agent, aided by the LLM, aggregates and synthesizes findings from the semantic analysis phase to construct a coherent view of research evolution. Studies are arranged chronologically, allowing the LLM to map the progression of research over time. This results in a refined set of evolutionary relationships, outlining how the new contribution aligns with, diverges from, or builds upon prior studies in the ORKG.

By automating the identification of research evolution and semantic transformation, this framework will enhances the discovery of implicit relations within scholarly literature. These connections can either be updated directly in the ORKG or stored in a separate historical knowledge graph (HKG) [35] to effectively describe evolutionary relationships while distinguishing them from factual research data, enabling in-depth analysis, preventing overload on the ORKG, and allowing for independent refinement of complex temporal data within the HKG.

4. Semantic Similarity Analysis

In this section, we will delve into semantic similarity analysis, one of the key component of semantic analysis step mentioned in Section 3. By evaluating the semantic relatedness between publications, a comprehensive and detailed view of the interrelationships among documents can be revealed which can serve as a foundational basis for subsequent semantic analyses in exploring the evolutionary relations between them. Specifically, we developed a system to measure semantic similarity between scientific publications across three key dimensions:

- the properties extracted from the papers, which consist of metadata available on the ORKG comparison tables
- the combined title and abstract across papers
- the main content of the paper (excluding abstract)

In order to quantify similarity across these dimensions, we employ several prominent sentence embedding models SBERT- *all-MiniLM-L6-v2, all-MiniLM-L12-v2*, known for its efficiency and robust general language representations that capture broad semantic features, and *allenai-specter, allenai/specter2_classification*, which, being specifically trained on scientific literature, provides deeper domain specific insights. Cosine similarity is then used as the metric to compute the similarity between the resulting sentence embeddings.

Analyzing Properties Similarity:

In this dimension, the system focuses on evaluating technical and contextual metadata associated publications. These properties provided by the ORKG users capture details about the research aim, employed methods, outcome, and other relevant information (e.g., author, publisher,...). Initially, a set of heuristic rules is applied to groups these properties into four main categories: *goal*, *method*, *result* and *other*. For example, properties containing phrases like “aim” or “objective” are grouped under “goal” while those mentioning “technique” or “algorithm” are classified as “method”. In the case when the heuristic rules cannot assign a property to a category, a zero-short classification model is used by default to assign it to the most appropriate category. This systematic grouping refers to standardizing the diverse metadata provided by various paper templates which enables fine-grained comparison of their core characteristics. The text corresponding to each category is converted into vector embedding and pairwise cosine similarity is computed between the corresponding groups of different papers to quantify their semantic alignment.

Analyzing Title and Abstract Similarity:

The semantic relatedness of each paper's core identifying information is evaluated by merging its title and abstract in a single composite text. This combined text is then transformed into a high dimensional vector representation using advanced embedding models of SBERT, describing the key themes and contributions of the paper. The cosine similarity between these vectors is computed to measure how closely the core ideas of different papers align.

Analyzing Main Content Similarity:

The main content structure similarity is assessed by leveraging the extension of IMRaD standard (with the addition of a *Related Work* section) [36, 37] which organizes scientific publications into five main sections: *Introduction*, *Methodology*, *Discussion*, *Results*, and *Related Work*. Specifically, the similarity between each pair of papers is determined by comparing the corresponding sections in each document, thereby capturing the semantic relatedness across the publications. This content similarity analysis begins with the extraction of main content of the paper. Input documents in PDF format are preprocessed using GROBID [38, 39], a robust tool for extracting the core content of scientific publications as an XML file. Following this, a comprehensive cleaning procedure is applied to the raw XML text to remove extraneous header blocks (e.g., *Title*, *DOI*, *Abstract*), non-essential formatting and to convert table and figure elements into plain text with section titles normalized to title case. This step ensures that only the core content, standardized content remains for accurate section classification and semantic similarity analysis. Next, the classification of content sections is performed by applying IMRaD framework. For each publication, if explicit section headers are present, the corresponding content is directly extracted, ensuring the original structure of the document is preserved. In the cases when the paper lacks clear section markers or does not adhere to the IMRaD format, a fine-tuned classification model built on the available dataset of [37] is employed to analyze each paragraph and automatically assign it to the most appropriate IMRaD category, thereby standardizing the content regardless of its original format.

After the sections of each paper are identified and grouped, each section is processed to compute semantic similarity. For each section, if the content is extensive, it is segmented into smaller chunks based on a fixed number of sentences. Each chunk is then converted into a vector representation using an SBERT model. The resulting chunk vectors are averaged to generate a single vector that represents the entire section. Finally, the cosine similarity between the averaged vectors of corresponding sections from two papers is calculated to assess their semantic relatedness.

5. Experiments

In this section we describe the experimental assessment of our method to measure semantic similarity between papers of a SKG. We start with the introduction of the experimental protocol and material and then we present the results and discuss the limitations of the approach.

5.1. Experimental Protocol

Our objective is to evaluate the approach we have described in Section 4. To do so, we evaluated the agreement between our algorithm and reference datasets as depicted in Fig. 2. Our reference datasets consist of two sets of scientific papers selected from two different comparison tables from the ORKG.

The **Paper selection and preprocessing** task consists in preparing the papers that will serve in our experiments. In *Papers group 1*, we have a set of 66 pairs of *related* papers. We obtained this group by considering the 12 papers contained in the *ORKG CT1* comparison table[3]. As the papers came from the same comparison table, we can assume that these papers are identified as related by the user who has created the table. *Papers group 2* contains 72 pairs of papers (p, q) where $p \in$ *ORKG CT1* and $q \in$ *ORKG*

[3] https://ORKG.org/comparison/R764428

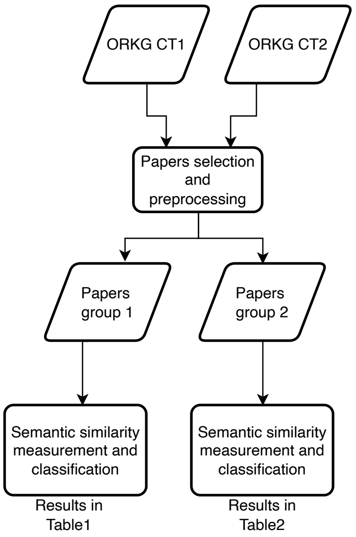
ORKG CT1
ORKG CT2
Papers selection and preprocessing
Papers group 1
Papers group 2
Semantic similarity measurement and classification
Semantic similarity measurement and classification
Results in Table1
Results in Table2

Figure 2: Experimental protocol

CT2. *ORKG CT2* is another comparison table[4] which thematic is different from the one of *ORKG CT1* and contains 6 papers. The pair (p, q) therefore form a pair of unrelated papers.

The **Semantic similarity measurement and classification** task consists in taking the pairs $(p_1, q_1) \in$ *Papers group 1* and the pairs $(p_2, q_2) \in$ *Papers group 2* and computing the semantic similarity between p_1 and q_1 on one hand and between p_2 and q_2 on the other hand. To do so, we consider the three cases described in Section 4 and generated embeddings using a dedicated model (in our experiments we have tested *all-MiniLM-L6-v2*, *allenai-specter*, *all-MiniLM-L12-v2* and *allenai/specter2_classification*). Once we have the required embeddings, we use the well-known cosine distance to measure the distance between the embeddings.

The classification is then done based on the interpretation of the consolidated value of the cosine distance we obtained. This consolidation is done as follows:

- Regarding **the ORKG properties**: We compute the distance between the *Goal* property of each paper, the *Method* property of each paper and the *Result* property of each paper. The resulting distance is the average value obtained from these 3 values.
- Regarding **title and abstract**: The distance is obtained by considering the full abstract and title.
- Regarding **main content**: After harmonizing the structure of each papers according to the IMRaD structure (see Section 4), we compute the distance between the *Introduction* section of each paper, the *Method* section of each paper and the *Result* section of each paper. The resulting distance is the average value obtained from these 3 values.

Then if the obtained consolidated distance value is below 0.4 we consider the papers as “Not related”, if the value is between 0.4 and 0.7 the relationship is “Unclear” and if the value is greater than 0.7 we consider the papers as “Related”.

[4] https://ORKG.org/comparison/R36099

5.2. Results

In this section, we present the results obtained from executing our experimental protocol on two groups of data. Table 1 and Table 2 provide a comparative analysis of the models, with columns representing the models and sub-columns distinguishing between the types of information used in the experiments: *title & abstract*, *full harmonized text*, and *structured data from the ORKG*. The rows in these tables show the percentage of agreement between the model predictions and the manual classification of papers performed by the ORKG experts, which serves as the baseline for evaluation.

In Table 1, the row "Related" indicates the true positive rate, while the row "Not Related" represents the false positive rate. Conversely, in Table 2, these definitions are reversed: the "Not Related" row corresponds to the true positive rate, and the "Related" row corresponds to the false positive rate. In both tables, the "Unclear" row captures cases where the algorithm was unable to make a definitive classification (i.e., when the similarity score falls between 0.4 and 0.7).

Table 1
with the papers within a group (%)

	all-MiniLM-L6-v2			**all-MiniLM-L12-v2**		
	Title & Abstract	**Main Content**	**the ORKG metadata**	**Title & Abstract**	**Main Content**	**the ORKG metadata**
Related	12.1	24.2	19.7	22.7	9.1	34.8
Not Related	1.5	0.0	0.0	0.0	0.0	0.0
Unclear	86.4	75.8	80.3	77.3	90.9	65.2

	allenai-specter			**allenai/specter2_classification**		
	Title & Abstract	**Main Content**	**the ORKG metadata**	**Title & Abstract**	**Main Content**	**the ORKG metadata**
Related	100.0	100.0	100.0	100.0	100.0	100.0
Not Related	0.0	0.0	0.0	0.0	0.0	0.0
Unclear	0.0	0.0	0.0	0.0	0.0	0.0

Table 2
with the papers from different groups (%)

	all-MiniLM-L6-v2			**all-MiniLM-L12-v2**		
	Title & Abstract	**Main Content**	**the ORKG metadata**	**Title & Abstract**	**Main Content**	**the ORKG metadata**
Related	0.0	0.0	0.0	0.0	0.0	0.0
Not Related	0.0	84.7	100.0	100.0	97.2	100.0
Unclear	100.0	15.3	0.0	0.0	2.8	0.0

	allenai-specter			**allenai/specter2_classification**		
	Title & Abstract	**Main Content**	**the ORKG metadata**	**Title & Abstract**	**Main Content**	**the ORKG metadata**
Related	0.0	83.3	5.6	100.0	59.7	100.0
Not Related	0.0	0.0	0.0	0.0	0.0	0.0
Unclear	100.0	16.7	94.4	0.0	40.3	0.0

The results from Table 1 highlight the exceptional performance of *allenai-specter* and *allenai-specter2* model in detecting similar papers. Even when using only title and abstract, these models achieved 100% precision. This outcome aligns with the fact that *allenai-specter* and *allenai-specter2* are specifically trained to detect co-citation similarity, meaning they excel at identifying pairs of papers that cite the same cluster of references. However, while highly effective at classifying related papers, Table 2 reveals that both models are less efficient in identifying non-related pairs. A possible explanation is that the models primarily focus on signals that establish connections between papers but lacks mechanisms to detect dissimilarities effectively. Another possibility is that the threshold used to make decisions should be different from the adopted one.

A different approach is taken by *all-MiniLM-L6-v2* and *all-MiniLM-L12-v2*, both models are based on SBERT and trained to assess sentence-level similarity. As observed in Table 1, for *all-MiniLM-L6-v2*, using only title & abstract does not provide sufficient context to precisely determine whether two papers are related, often leading to misclassifications. However, structured the ORKG metadata appears to capture essential similarities in related papers, while the *main content* provides even more reliable results though it requires preprocessing PDF files. The model *all-MiniLM-L12-v2* shows a significant improvement in the results for Title & Abstracts, as well as for the ORKG, but the main content seams to be less important for decisions. Table 2 also shows the improvement on the new version of *MiniLM* for all three data sources. We would like to highlight that the ORKG presents a balanced compromise between these data sources. It provides high-quality, structured information while maintaining low execution time, as the data in comparison tables is both concise and highly relevant. In contrast, processing full-text papers requires significantly more preprocessing effort, generates a large volume of data proportional to the document size, and results in longer execution times.

A noteworthy observation is that false negatives are minimal for title & abstract corpora and completely absent when using full text or the ORKG metadata. However, the rate of unclear classifications remains relatively high. A deeper analysis of these cases reveals that in over 70% of unclear instances, the similarity score falls between 0.6 and 0.7.

In future works, we plan to expand our experiments to all comparison tables of the ORKG. We will also further refine our classification thresholds. In particular, we will investigate how to systematically define the threshold for each model (independently) for determining relatedness between papers. Additionally, we aim to analyze the impact of different IMRaD (Introduction, Methods, Results, and Discussion) categories on classification accuracy. This will allow us to explore whether assigning different weights to these categories (e.g., calculate weighted average of categories) or focusing solely on specific categories could enhance decision-making in scholarly knowledge graph evolution.

6. Conclusion

In this paper, we present the mid-term vision of our agentic-RAG framework, with a particular focus on one key feature: semantic similarity analysis. We describe our approach to select and structure information from scientific papers and how these data sources are utilized within our MOKA framework. Additionally, we detail the integration of various semantic similarity tools to assess whether pairs of papers are related. Our extensive evaluation demonstrates that the choice of similarity model significantly impacts precision, with results ranging from 12% to 100%. These findings highlight the advantages of agentic approaches, which dynamically adjust the framework's configuration based on the specific objectives and intentions of the end-user.

The preliminary results shows a promising path for MOKA. This framework for analyzing how new research contributions align, diverge from, or build upon prior studies, can potentially facilitate a deeper understanding of the evolving scholarly landscape. By leveraging structured knowledge graphs (SKGs) such as the ORKG, MOKA will enhances the discovery of semantic evolutionary relationships and offers a novel approach to tracking research progression. However, several limitations must be acknowledged, highlighting key areas for future improvements.

One major limitation lies in the restricted exploration scope within the ORKG. While the ORKG serves as an effective prototype SKG, the model is designed to generalize across different scholarly knowledge graphs. Adapting the framework to other SKGs with unique data structures and ontologies would enhance its applicability across various research domains. Future work should explore fine-tuning the model for diverse SKGs to maximize its impact on a broader scholarly landscape.

Additionally, the model currently lacks a dedicated mechanism to represent discovered evolutionary relationships. Directly updating the ORKG with inferred connections may lead to data overload and ambiguity. A promising alternative is to store these relationships in a separate structure, such as a historical knowledge graph (HKG) [35], allowing for more refined temporal analysis while preserving the integrity of factual research data.

The access restrictions to full-text publications pose another significant limitation. Many scholarly works remain locked behind paywalls, restricting the model's ability to perform comprehensive text analysis. A practical mitigation strategy is to rely on open-access repositories (e.g., Arxiv[5], HAL[6]), which provide unrestricted access to valuable research content.

Finally, processing lengthy scholarly texts with large language models (LLMs) remains a challenge, particularly in maintaining contextual accuracy across multi-page documents. However, with rapid advancements in LLM architectures, this limitation is expected to diminish over time. In the short term, a strategic approach is to focus on key sections of publications such as abstracts, methods, or discussions, and literature reviews where crucial research contributions and evolutionary relationships are most explicitly stated.

Despite these challenges, MOKA framework represents a significant step toward automating the identification of research evolution in scholarly literature. Future work should focus on refining data retrieval strategies, expanding applicability to multiple SKGs, and integrating advanced validation frameworks to enhance reliability and scalability.

Acknowledgments

This project is funded by the European Union's Horizon 2022 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 101102337.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. Wang, Y. Yan, J. Wang, Y. Jia, Y. Zhang, W. Zhang, X. Wang, Acekg: A large-scale knowledge graph for academic data mining, in: CIKM, ACM, 2018, pp. 1487–1490.

[2] M. Färber, The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data, in: ISWC (2), volume 11779 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 113–129.

[3] C. D. Giambattista, I. Heibi, S. Peroni, D. M. Shotton, Opencitations: an open e-infrastructure to foster maximum reuse of citation data, Int. J. Digit. Curation 17 (2022) 5.

[4] M. Färber, D. Lamprecht, J. Krause, L. Aung, P. Haase, Semopenalex: The scientific landscape in 26 billion RDF triples, in: ISWC, volume 14266 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 94–112.

[5] D. Dessì, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, CS-KG: A large-scale knowledge graph of research entities and claims in computer science, in: ISWC, volume 13489 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 678–696.

[6] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, Bibliothek Forschung und Praxis 44 (2020) 516–529. URL: https://doi.org/10.1515/bfp-2020-2042. doi:doi:10.1515/bfp-2020-2042.

[7] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, Agentic retrieval-augmented generation: A survey on agentic rag, ArXiv abs/2501.09136 (2025). URL: https://api.semanticscholar.org/CorpusID:275570331.

[5] https://arxiv.org/

[6] https://hal.science/

[8] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: EMNLP/IJCNLP (1), Association for Computational Linguistics, 2019, pp. 3980–3990.

[9] J. Gómez, P.-P. Vázquez, An empirical evaluation of document embeddings and similarity metrics for scientific articles, Applied Sciences 12 (2022) 5664.

[10] A. Oelen, M. Y. Jaradeh, K. E. Farfar, M. Stocker, S. Auer, Comparing research contributions in a scholarly knowledge graph, in: SciKnow@K-CAP, volume 2526 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 21–26.

[11] K. Luu, X. Wu, R. Koncel-Kedziorski, K. Lo, I. Cachola, N. A. Smith, Explaining relationships between scientific documents, in: ACL/IJCNLP (1), Association for Computational Linguistics, 2021, pp. 2130–2144.

[12] X. Li, J. Ouyang, Explaining relationships among research papers, CoRR abs/2402.13426 (2024).

[13] X. Li, J. Ouyang, Explaining relationships among research papers, in: COLING, Association for Computational Linguistics, 2025, pp. 1080–1105.

[14] X. Xing, X. Fan, X. Wan, Automatic generation of citation texts in scholarly papers: A pilot study, in: Annual Meeting of the Association for Computational Linguistics, 2020. URL: https://api.semanticscholar.org/CorpusID:220045125.

[15] M. Dalle Lucca Tosi, J. C. dos Reis, Understanding the evolution of a scientific field by clustering and visualizing knowledge graphs, Journal of Information Science 48 (2022) 71–89.

[16] A. Rossanez, J. C. dos Reis, R. da Silva Torres, Representing scientific literature evolution via temporal knowledge graphs, in: MEPDaW@ISWC, 2020. URL: https://api.semanticscholar.org/CorpusID:233433129.

[17] J. T. Aparicio, E. Arsenio, F. Santos, R. Henriques, Using dynamic knowledge graphs to detect emerging communities of knowledge, Knowledge-Based Systems 294 (2024) 111671.

[18] J.-C. Liu, C.-T. Chen, C. Lee, S.-H. Huang, Evolving knowledge graph representation learning with multiple attention strategies for citation recommendation system, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–26.

[19] X. Gu, M. Krenn, Impact4cast: Forecasting high-impact research topics via machine learning on evolving knowledge graphs, in: ICML 2024 AI for Science Workshop, 2024.

[20] A. A. Salatino, A. Mannocci, F. Osborne, Detection, Analysis, and Prediction of Research Topics with Scientific Knowledge Graphs, Springer International Publishing, Cham, 2021, pp. 225–252. doi:10.1007/978-3-030-86668-6_11.

[21] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, Aida: A knowledge graph about research dynamics in academia and industry, Quantitative Science Studies 2 (2021) 1356–1398.

[22] A. Widianto, E. Pebriyanto, F. Fitriyanti, M. Marna, Document similarity using term frequency-inverse document frequency representation and cosine similarity, Journal of Dinda: Data Science, Information Technology, and Data Analytics 4 (2024) 149–153.

[23] H. Xu, W. Zeng, J. Gui, P. Qu, X. Zhu, L. Wang, Exploring similarity between academic paper and patent based on latent semantic analysis and vector space model, in: 2015 12th international conference on fuzzy systems and knowledge discovery (FSKD), IEEE, 2015, pp. 801–805.

[24] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[25] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[26] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[27] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[28] J. Jeong, D. Jin, Automatic classification of scientific and technical papers using large language models and retrieval-augmented generation (2024).

[29] G. Mitrov, B. Stanoev, S. Gievska, G. Mirceva, E. Zdravevski, Combining semantic matching, word embeddings, transformers, and llms for enhanced document ranking: Application in systematic

reviews, Big Data and Cognitive Computing 8 (2024) 110.

[30] C. Ravuru, S. S. Srinivas, V. Runkana, Agentic retrieval-augmented generation for time series analysis, CoRR abs/2408.14484 (2024).

[31] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=WE_vluYUL-X.

[32] S. Xu, Z. Wu, H. Zhao, P. Shu, Z. Liu, W. Liao, S. Li, A. Sikora, T. Liu, X. Li, Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis, CoRR abs/2402.11398 (2024).

[33] Y. Feng, Semantic textual similarity analysis of clinical text in the era of llm, 2024 IEEE Conference on Artificial Intelligence (CAI) (2024) 1284–1289. URL: https://api.semanticscholar.org/CorpusID:271040424.

[34] I. Dagan, W. B. Dolan, B. Magnini, D. Roth, Recognizing textual entailment: Rational, evaluation and approaches – erratum, Natural Language Engineering 16 (2010) 105 – 105. URL: https://api.semanticscholar.org/CorpusID:8336653.

[35] S. D. Cardoso, M. D. Silveira, C. Pruski, Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies, Knowl. Based Syst. 194 (2020) 105508.

[36] I. Ahmed, M. T. Afzal, A systematic approach to map the research articles' sections to imrad, IEEE Access 8 (2020) 129359–129371. URL: https://api.semanticscholar.org/CorpusID:220733920.

[37] T. Saier, J. Krause, M. Färber, unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network, in: Proceedings of the 23rd ACM/IEEE Joint Conference on Digital Libraries, JCDL '23, 2023.

[38] Grobid, https://github.com/kermitt2/grobid, 2008–2025.

[39] P. Lopez, L. Romary, Grobid - information extraction from scientific publications, ERCIM News 2015 (2015). URL: https://api.semanticscholar.org/CorpusID:36526770.