Kathmandu University

Department of Artificial Intelligence

Panchkhal, Kavre

A Project Proposal

on

"Research Space"

[Code No.: AISP 301]

(For partial fulfillment of III/I Year/Semester in Artificial Intelligence)

Submitted by:

Sujal Bajracharya(Roll no. 4)

Ashim Shrestha(Roll no. 22)

Yajjyu Tuladhar(Roll no. 27)

Submitted to:

Subodh Acharya

Department of Artificial Intelligence

Submission Date: March 31, 2025

Abstract

In the rapidly evolving landscape of academic research, the sheer volume and velocity of publications pose significant challenges for researchers seeking relevant literature. Traditional keyword-based search methods often fail to capture the nuanced relationships between papers, leading to inefficiencies and information overload. This project proposes an innovative academic research recommendation system that leverages a hybrid approach integrating citation graphs, knowledge graphs, text embeddings, and Graph Retrieval-Augmented Generation (GraphRAG) to enhance literature discovery and comprehension. Using bulk dataset of research papers the methodology constructs interconnected graphs and text-content models, enabling citation-based and semantic search ranked by relevance and impact. A GraphRAG-powered ”chat with PDF” feature offers interactive paper exploration. The expected outcome is a scalable web application that not only retrieves and ranks highly relevant papers but also facilitates deeper engagement with academic content, significantly improving research efficiency and understanding for scholars and students alike.

Keywords:LLM, recommendation system, Graph Based Method, GraphRAG, Knowledge Graph, SciBERT, KeyBERT, citation graph, text-content graph

Table of Contents

List of Figures

Acronyms/Abbreviations

Chapter 1 Introduction

1.1 Background

In today's digital age, the abundance of academic literature can be overwhelming for researchers and upcoming students seeking relevant information. Finding relevant papers in specific domains is increasingly challenging due to the volume of publications and the lack of efficient tools for personalized recommendations. Also in fields like AI the velocity of new publications is so high that if you are not reading everyday you fall behind. To address this challenge, an academic research recommendation can streamline the process by utilizing citation networks and semantic search for promising discovery processes, while the LLMs can provide unprecedented capabilities in information retrieval and summarization.

1.2 Problem Statement

Current academic research platforms struggle to provide relevant paper recommendations due to limitations in traditional keyword-based search methods. These methods often fail to account for the detailed relationships between papers, leading to information overload and inefficiency in literature discovery. Also many papers don't explain the information that they learned from the documents they have cited. This makes research tedious as to understand one academic paper you will have to read multiple other papers too.

1.3 Objectives

The project aims to transform how researchers navigate and engage with academic literature by developing an LLM-powered system that integrates citation networks, knowledge graphs, and user-centric design. Our project aim to fulfill the following tasks:

- To retrieve papers relevant to user queries using a hybrid approach.
- Rank recommendations based on citation impact and semantic similarity.
- Explain papers using GraphRag.

1.4 Motivation and Significance

We ourselves have been reading or trying to read academic papers to understand AI and its current advancements but we have found that current academic research platforms like google scholar, arxiv, etc are not so easy to use if we want to read the cited papers as well. For that we found connected papers to be slightly better as it visualises the paper as a graph with its cited papers but understanding the paper completely is still a challenge and using LLMs like chatgpt, deepseek only gives partially correct or sometimes completely wrong answers.

Chapter 2 Related Works

Recommender Systems (RS) are one of the most popular and important applications of Artificial Intelligence (AI). They have been widely adopted to help the users of many popular content sharing and e-Commerce web sites to more easily find relevant content, products or services. Meanwhile, Graph Learning (GL), which relates to machine learning applied to graph structure data, is an emerging technique of AI which is rapidly developing and has shown its great capability in recent years (Wang et al. 2021).

Many researchers have explored graph-based recommendation methods. For instance, (Fan et al. 2019) employed a path-based approach for intent recommendation to automatically recommend user intent according to user historical behaviors without any input when users open the App, similarly (Song et al. 2022) also proposed a path based reinforcement learning method by formulating recommendation as a sequential decision process. While (Palumbo, Rizzo, and Troncy 2017) utilized an embedding-based method. Similarly, (F. Zhang et al. 2016) also used an embedding-based technique by exploiting the knowledge base, to design three components to extract items' semantic representations from structural content, textual content and visual content, respectively.

(F. Zhang et al. 2016) used the TransR which was proposed by (Z. Zhang et al. 2021) which is a modification to the original TransE proposed by (Bordes et al. 2013). TransE and TransR are both methods to represent knowledge graphs (like "Paris is capitalof France") as numerical vectors for AI. TransE treats relationships as simple translations in one shared space (e.g., "Paris + capitalof = France"). It's fast but struggles with complex relationships (e.g., one person writing many books). TransR fixes this by first projecting entities (like "Paris") into a custom space for each relationship before translating, making it better for tricky cases but slower. Think of TransE as a one-size-fits-all map, while TransR uses different maps for different roads.

Graph based methods have also been researched for academic paper recommendation. (Liu et al. 2025) proposed a hybrid recommendation framework for scientific article recommendation by constructing a large-scale citation network comprising 190,381 articles from 70 journals spanning statistics, econometrics, and computer science from 1981 to 2022. The framework integrates network-based citation patterns with content-based semantic similarities. To enhance content-based recommendations, OpenAI's text-embedding-3-small model was employed to generate embedding vectors for article abstracts, ensuring computational efficiency and embedding stability for handling dynamic academic databases.

Another study, (Church et al. 2024) explored the complementary nature of content-based filtering (CBF) and graph-based methods (GB) in academic search recommendations. They described how CBF analyzes abstracts to infer authors' positions, while GB examines citations to capture audience responses. Their study identified nine key differences between CBF and GB, highlighting synergistic opportunities for hybrid approaches. Notably, they utilized two embedding techniques: Specter (Cohan et al. 2020), a BERT-based model for encoding abstracts, and ProNE (J. Zhang et al. 2019), a spectral clustering-based method applied to over 200M papers and 2B citations from Semantic Scholar. As research in recommender systems advances, Retrieval-Augmented Generation (RAG) presents a promising approach to enhancing the explainability and interpretability of academic papers and recommendation methods. RAG combines retrieval mechanisms with generative models to dynamically generate responses based on external knowledge sources, making it an effective tool for summarizing and interpreting complex research articles. RAG fails on global questions directed at an entire text corpus, such as "What are the main themes in the dataset?", since this is inherently a query-focused summarization (QFS) task, rather than an explicit retrieval task (Edge et al. 2025).

(Edge et al. 2025) proposed GraphRAG, a graph-based approach to question answering over private text corpora that scales with both the generality of user questions and the quantity of source text.

Chapter 3 Design and Implementation

3.1 Data Collection

Datasets will be collected from reliable sources like arXiv using their public API, Semantic Scholar's S2ORC. After the dataset has been collected most of it will already be in structured format except the PDFs of the papers. For the pdfs we will parse and extract relevant textual information from it.

3.2 Graph Construction and Storage

3.2.1 Citation Graph

The citation graph represents papers as nodes and citation links as directed edges, where an edge from Paper A to Paper B indicates that A cites B. This graph will be implemented using NetworkX for rapid prototyping and small-scale analysis or Neo4j for handling larger datasets with efficient querying capabilities.

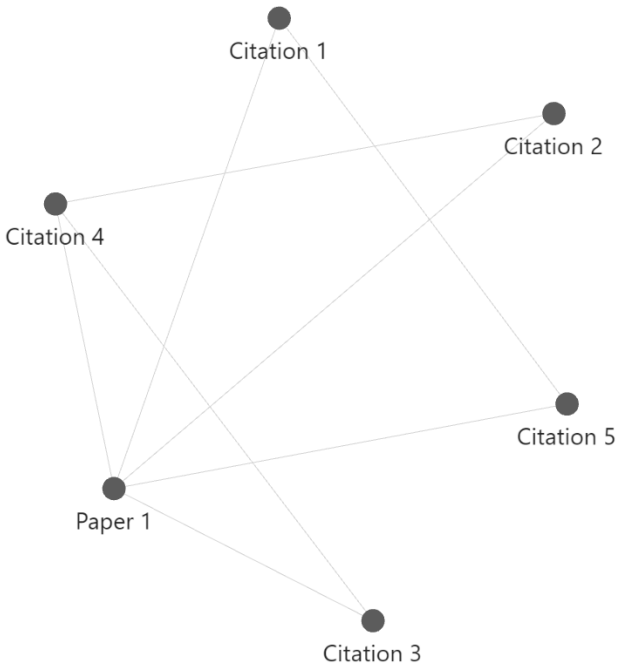

Figure 1: Example citation graph with dummy nodes

3.2.2 Knowledge Graph

Knowledge Graph will extend the citation graph by incorporating additional information: author networks (edges between co-authors), subject networks(edges to papers of the same subject), and topic embeddings (vector representations of paper content derived from paper). This multi-layered graph provides a holistic view of academic connections, enabling more nuanced retrieval by linking papers through authors, and topics beyond direct citations.

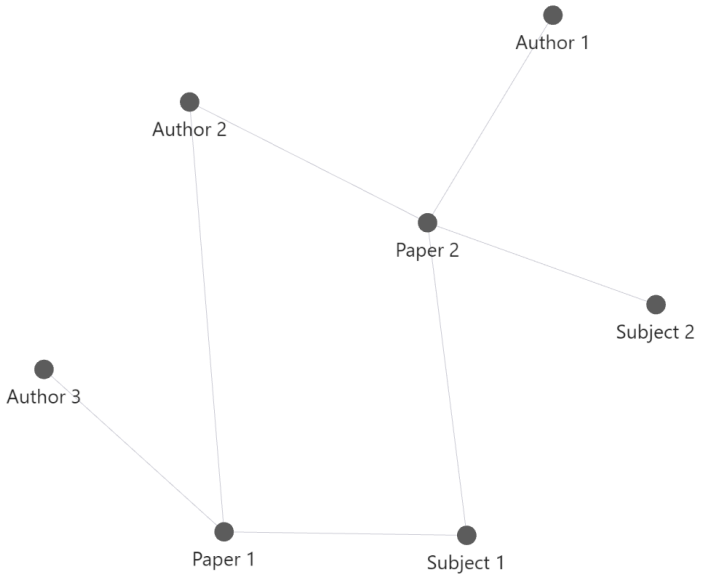

Figure 2: Example knowledge graph with dummy nodes

3.2.3 Text-Content Graph

For each paper, a Text-Content Graph will be constructed from its full text (title, abstract, and body) to capture intra-paper semantic relationships. Key entities and concepts (e.g., ”vision transformer,” ”healthcare,” ”deep learning”) will be extracted as nodes using SciBERT-based Named Entity Recognition (NER) (Beltagy, Lo, and Cohan 2019) and KeyBERT (Issa et al. 2023) for keyphrase identification. Edges will be defined based on relationships identified through multiple methods: co-occurrence (entities appearing in the same sentence or paragraph, weighted by fre-

quency), dependency parsing (syntactic relationships like ”vision transformer improves healthcare,” labeled accordingly), and semantic similarity (cosine similarity between SciBERT embeddings of entities exceeding a threshold, e.g., 0.8). These graphs will be built using NetworkX for in-memory processing, with each node representing an entity/concept and each edge reflecting a contextual or syntactic link within the paper.

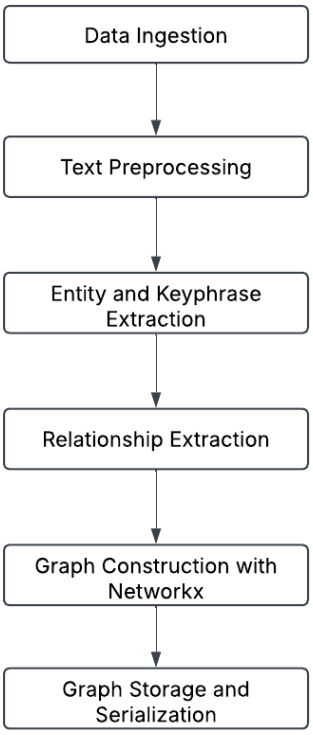

Figure 3: Text-Content Graph Flowchart

3.2.4 Semantic Embeddings

Semantic representations will be generated for the entire text of each paper (title, abstract, and body) to fully capture its content and context. SciBERT, a transformer model fine-tuned on scientific texts, will process the full text in segments (e.g., 512-token chunks) due to its input constraints, producing contextual embeddings (768-dimensional vectors) for each segment. These will be averaged across segments to create a single content-based embedding per paper, reflecting its scientific meaning. Simultaneously, TransR will embed each paper based on its position in the knowl-

edge graph, generating structural embeddings (e.g., 100-dimensional vectors) that encode relationships like citations and co-authorships. For each paper, the SciBERT (Beltagy, Lo, and Cohan 2019) and TransR (Z. Zhang et al. 2021) embeddings will be combined (e.g., concatenated or weighted) into a unified vector representation. These vectors will be stored in a vector database for efficient similarity search. The full corpus approach ensures richer representations, capturing details from methods, results and discussions. The database will be updated incrementally as new papers are added, recomputing embeddings only for new documents to maintain scalability.

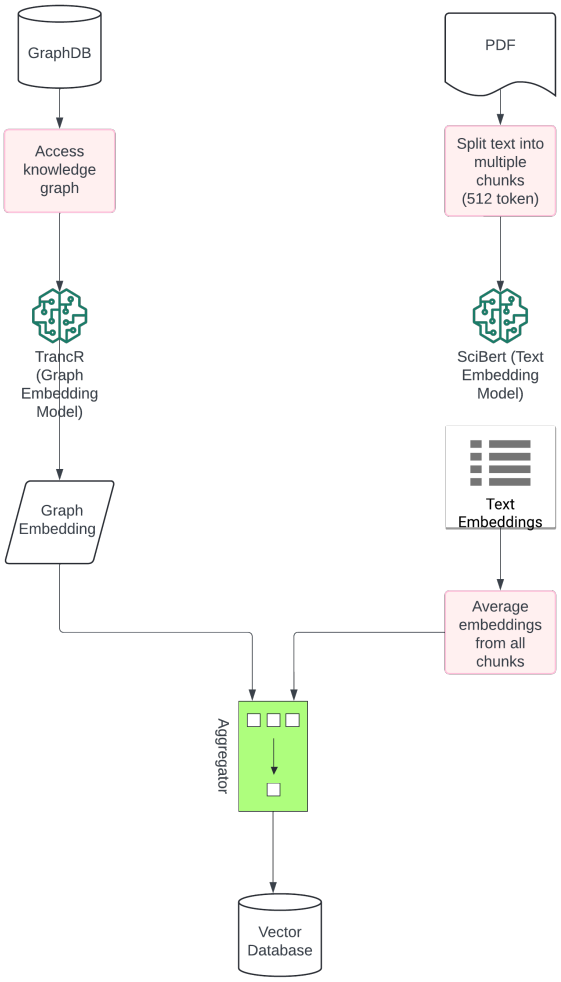

Figure 4: Embedding generation pipeline

3.3 Generating Recommendation

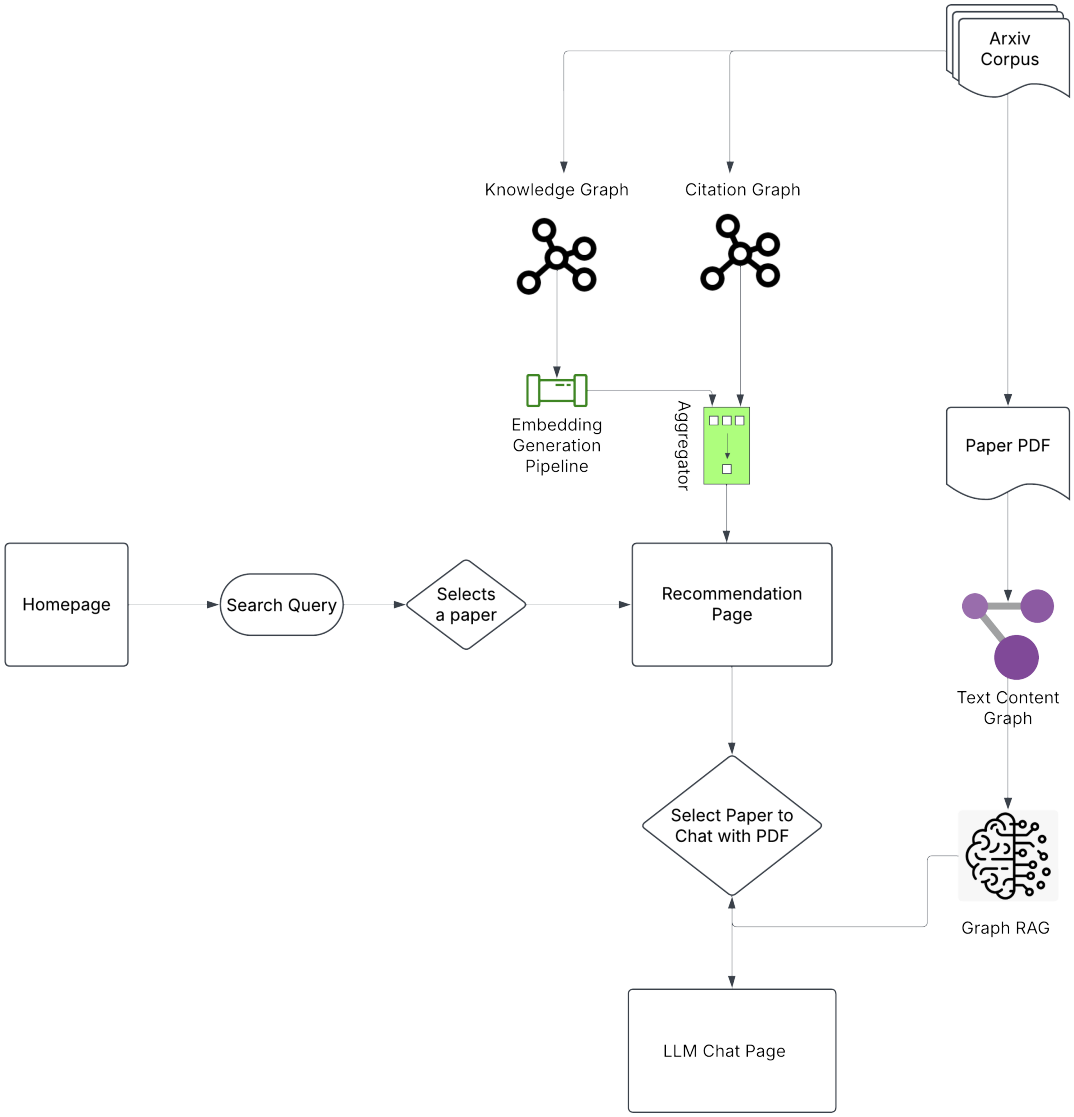

Figure 5: High-level System Architecture Overview

3.3.1 Citation-Based Search

This approach of generating recommendation is a path based approach where we will simulate random walks on the citation graph and the papers with the most visits will be recommended to the user. This approach ensures we suggest he most relevant paper.

3.3.2 Semantic Search

Semantic Search queries the vector database using a user input, embedding it with SciBERT, and retrieving the top K papers based on cosine similarity of content embeddings, supplemented by TransR embeddings for rational alignment. Using these embedding will compute similarity measures and recommend the most relevant papers.

3.3.3 Ranking

Each candidate is scored using a weighted combination including Citation strength, SciBERT similarity and TransR similarity. The top K papers are selected according to need.

3.4 GraphRAG Integration

GraphRAG leverages the Text-Content Graph to enable the LLM to chat with the content of a single PDF paper, providing an interactive question-answering feature.The text-content graph constructed from the paper's full text captures intra-paper semantic relationships. The graph comprises nodes as key entities and concepts and edges as relationships. Alongside this, the paper's SciBERT embeddings, which encode the full-text content into 768-dimensional vectors, are included to provide detailed semantic context. The LLM processes this hybrid input within the Graph RAG framework, reasoning over the text-content graph's structure and the embeddings' semantic depth. When a user poses a question, the LLM traverses the graph to identify relevant relationships and uses the embeddings to extract precise content, generating a natural language response. This integration empowers users to explore a paper's content interactively, offering a conversational interface.

Chapter 4 System Requirement Specification

4.1 Software Specifications

4.1.1 Front-End Techologies

- React.js (for building the user interface)
- Redux (for state management)
- Plotly (for interactive visualizations)

4.1.2 Back-End Techologies

- Python (primary programming language)
- FastAPI (for building scalable and high-performance APIs)

4.1.3 Database Systems

- Qdrant Vector Database (for storing and retrieving high-dimensional embeddings of academic papers)
- Neo4j Graph Database (for managing citation networks and knowledge graphs)

4.1.4 Machine Learning and AI Frameworks

- LangChain (for retrieval-augmented generation and LLM-based interactions)
- PyTorch (for training and fine-tuning deep learning models)

4.2 Hardware Specifications

4.2.1 Minimum Hardware Requirements

- Processor: Intel Core i7 (or equivalent) with at least 8 cores
- RAM: 16GB (recommended: 32GB for handling large embeddings)
- Storage: SSD with at least 500GB (recommended: NVMe SSD for fast retrieval)
- GPU: NVIDIA RTX 3060 (or higher) for model inference and training

4.2.2 Recommended Server Setup

- Cloud-based deployment on AWS, GCP, or Azure with GPU-enabled instances
- Dedicated database instances for Qdrant and Neo4j to optimize retrieval speed

This specification ensures that the system functions optimally for processing academic literature, generating knowledge graphs, and facilitating efficient search and recommendation.

Chapter 5 Project Planning and Scheduling

We are looking forward to completing the project in the time frame of 16 weeks by distributing the tasks to each member of the group . We are planning to work on our own on the weekdays and gather on the project day to discuss and compile the codes collectively. The following Gantt chart shows the time allocation for different aspects of our project:

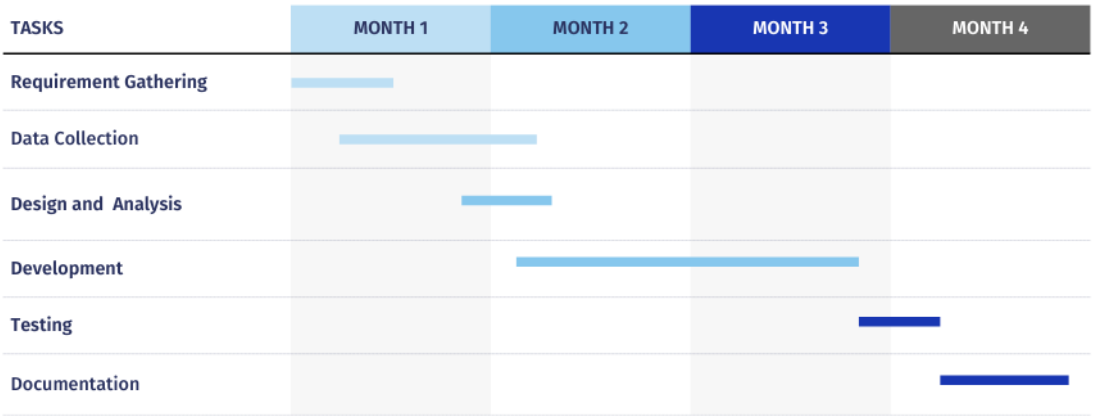

Figure 6: Project Gantt Chart

Chapter 6 Expected Outcome

The expected outcome of this project is a web application designed to revolutionize academic literature discovery. The system will employ a hybrid approach that combines citation networks and knowledge graphs to retrieve papers relevant to user queries. Recommendations will be ranked based on citation impact and semantic similarity, ensuring both influential and contextually relevant papers are prioritized.

Additionally, the integration of GraphRAG will enable users to interactively explore and understand the content of papers, enhancing their research experience. By addressing the limitations of current academic search platforms, this project aims to significantly improve the efficiency and effectiveness of literature discovery for researchers and students alike.

References

Wang, Shoujin et al. (2021). *Graph Learning based Recommender Systems: A Review*. arXiv: 2105.06339 [cs.IR].

Fan, Shaohua et al. (2019). "Metapath-guided Heterogeneous Graph Neural Network for Intent Recommendation". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, pp. 2478–2486. ISBN: 9781450362016.

Song, Weiping et al. (2022). *Ekar: An Explainable Method for Knowledge Aware Recommendation*. arXiv: 1906.09506 [cs.IR].

Palumbo, Enrico, Giuseppe Rizzo, and Raphaël Troncy (2017). "entity2rec: Learning User-Item Relatedness from Knowledge Graphs for Top-N Item Recommendation". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys '17. Como, Italy: Association for Computing Machinery, pp. 32–36. ISBN: 9781450346528.

Zhang, Fuzheng et al. (2016). "Collaborative Knowledge Base Embedding for Recommender Systems". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 353–362. ISBN: 9781450342322.

Zhang, Zhenghang et al. (Jan. 2021). "TransR*: Representation learning model by flexible translation and relation matrix projection". In: *J. Intell. Fuzzy Syst.* 40.5, pp. 10251–10259. ISSN: 1064-1246.

Bordes, Antoine et al. (2013). "Translating Embeddings for Modeling Multi-relational Data". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc.

Liu, Kun et al. (2025). *Academic Literature Recommendation in Large-scale Citation Networks Enhanced by Large Language Models*. arXiv: 2503.01189 [stat.AP].

Church, Kenneth et al. (2024). *Academic Article Recommendation Using Multiple Perspectives*. arXiv: 2407.05836 [cs.IR].

Cohan, Arman et al. (July 2020). "SPECTER: Document-level Representation Learning using Citation-informed Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 2270–2282.

Zhang, Jie et al. (2019). "Prone: Fast and scalable network representation learning." In: *IJCAI*. Vol. 19, pp. 4278–4284.

Edge, Darren et al. (2025). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. arXiv: 2404.16130 [cs.CL].

Beltagy, Iz, Kyle Lo, and Arman Cohan (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. arXiv: 1903.10676 [cs.CL].

Issa, Bayan et al. (2023). "A Comparative Study on Embedding Models for Keyword Extraction Using KeyBERT Method". In: *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, pp. 40–45.