Vol. 07, No. SI♣, pp. 19 –29 (2026)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Intelligent Multilingual Academic Journal Search: A Hybrid Semantic and Graph-Based Retrieval Approach

Shireen Fathi Malo[1,2,*], Adel AL-zebari[2]

[1]*Department of Information Technology, Duhok Polytechnic University, Iraq, shireen.fathi@auas.edu.krd*
[2]*Department of Information Technology, Akre University for Applied Sciences, Iraq, adel.hasan@auas.edu.krd*

**Correspondence: shireen.fathi@auas.edu.krd*

Abstract

In the rapidly expanding digital landscape of academic research, the sheer volume of scholarly publications presents significant challenges for researchers seeking relevant and high-quality information. Traditional keyword-based search engines often struggle with semantic nuances, interdisciplinary queries, and the growing multilingual nature of global scholarship. This paper introduces a novel, intelligent academic journal search engine designed to overcome these limitations. The system integrates a hybrid search architecture, combining Elasticsearch for lexical retrieval with Sentence-BERT (SBERT) for dense semantic embeddings, alongside Neo4j-powered knowledge graphs for relational exploration. It features robust bilingual support (Arabic-English) with domain-specific query correction and personalized recommendations. Empirical evaluation confirms the system's effectiveness, achieving a mean Precision@10 of 0.808 and an nDCG@10 of 0.889 on a curated bilingual dataset. Furthermore, a user study involving 40 academic participants revealed that 93% found the graph-based visualization essential for exploratory discovery, while load testing demonstrated stable performance under concurrent query loads. These results establish a new benchmark for inclusive, high-performance, and intelligent academic information retrieval.

Keywords: *Multilingual Information Retrieval, Semantic Search, Knowledge Graph-Based Recommendation, Academic Journal Discovery, Natural Language Processing*

Received: October 04th, 2025 /Revised: February 25th, 2026 / Accepted: April 12th, 2026 / Online: April 19th, 2026

I. INTRODUCTION

The digital revolution has fundamentally altered the paradigm of scholarly communication, leading to an exponential surge in academic output. Current estimates suggest that over 3 million articles are published annually across thousands of diverse journals, reflecting a complex and rapidly evolving knowledge landscape [1][2]. While this abundance of information offers unprecedented opportunities for scientific discovery, it simultaneously creates a significant "information overload" for researchers. Specifically, the process of identifying the most suitable and high-quality publishing venues or discovering journals that operate at the intersection of multiple disciplines has become increasingly cumbersome [3].

Traditional Academic Search Engines (ASEs), such as Scopus, Web of Science, and PubMed, serve as the foundational infrastructure for literature review and bibliometric analysis [4][5]. However, these legacy systems are predominantly built upon lexical retrieval models that rely on exact keyword matching and Boolean logic. Such approaches are inherently limited by the "vocabulary mismatch" problem, where the terminology used by a researcher may not align precisely with the indexed metadata of a journal [6]. Furthermore, as research becomes more interdisciplinary, traditional engines often fail to capture the semantic intent behind complex queries, leading to either overly broad result sets or the omission of highly relevant, semantically related journals [7].

In recent years, the field of Natural Language Processing (NLP) has introduced transformer-based architectures, most notably Bidirectional Encoder Representations from Transformers (BERT) and its variant, Sentence-BERT (SBERT), to bridge this gap[8][9][10]. These models represent text in high-dimensional vector spaces, enabling systems to understand the conceptual meaning of a query rather than just its surface string. While newer AI-driven tools like Elicit, ResearchRabbit, and Semantic Scholar have begun to integrate these technologies, they remain largely "article-centric." There is still a noticeable gap in tools specifically optimized for journal-level discovery, where the objective is to explore the

doi: 10.38094/jastt7031083

broader metadata, prestige, and relational context of a publication venue rather than an individual paper [11][12].

Another persistent challenge is the linguistic barrier in global scholarship. Despite the rise of multilingual models, most high-performance retrieval systems are optimized for English, leaving researchers who query in morphologically complex languages, such as Arabic, at a disadvantage[13][14] Existing cross-lingual solutions often struggle to maintain semantic fidelity during translation, especially within specialized academic domains [15][16]. Consequently, there is a practical need for systems that integrate robust translation pipelines with domain-specific correction to ensure equitable access to scholarly information.

To address these challenges, this paper presents a hybrid, intelligent academic journal search system. The proposed framework does not aim to replace established databases but rather to act as an intelligent abstraction layer that enhances the discovery process. By combining Elasticsearch for structured lexical filtering with SBERT for semantic retrieval, and utilizing Neo4j to model inter-journal relationships via knowledge graphs, the system provides a more nuanced approach to journal exploration. The integration of a bilingual processing pipeline further supports inclusive access for the Arabic-speaking research community.

The primary contributions of this work are summarized as follows:

- A Hybrid Retrieval Framework: Development of a dual-stream search engine that integrates lexical keyword matching with semantic embeddings, enabling balanced retrieval between exact term matching and conceptual similarity.
- Relational Knowledge Mapping: Implementation of a graph-based visualization module using Neo4j that allows researchers to explore relationships among journals based on shared subject domains and indexing characteristics.
- Bilingual Query Processing: Design of a tailored NLP pipeline supporting Arabic–English queries through translation and academic term normalization, improving accessibility for multilingual researchers.
- Empirical Performance Analysis: Comprehensive evaluation using standard retrieval metrics (Precision@10 and nDCG@10), system latency analysis, and a user study involving 40 academic participants.

Compared with existing AI-based research discovery tools such as Elicit and Semantic Scholar, which primarily focus on article-level search, the proposed system focuses on journal-level discovery while integrating multilingual processing and graph-based exploration capabilities.

The remainder of this article is organized as follows: Section 2 reviews the evolution of search engines and current AI advancements; Section 3 describes the system architecture and data integration; Section 4 presents the experimental results and comparative analysis; and Section 5 discusses limitations and potential future directions.

II. LITERATURE REVIEW

A. Academic Journal Search Engines

1) Traditional Journal Search Engines

Academic search engines have been the backbone of scholarly discovery. Academic search engines are the backbone of modern research, helping in access to peer-reviewed journal content and also serving the global research community's requirement for systematic literature review and bibliometric analysis [6]. Among these, Scopus, published by Elsevier since 2004, is a widely known database with multidisciplinary coverage and more than 24,000 active journals, conference papers, and patents indexed in it. Its powerful citation mapping measures provide users with the ability to estimate research quality, map the paths citations travel, and track cooperation [2][5]. The platform's advanced search utilizes structured keyword queries, extensive Boolean logic, and all within well-defined bibliographic fields. Though such an approach allows for fine-grained filtering and citation mapping, it can also present barriers to discovery especially when user queries do not exactly correspond to indexed terminology or are interdisciplinary or summarized [2] [4].

Web of Science (WoS), curated by Clarivate Analytics, is distinguished by its historical depth and selective curation, maintaining authoritative archives dating back to 1900. WoS is often considered a benchmark for research evaluation due to its rigorous journal selection and comprehensive citation tracking [4]. However, its coverage can be more limited in the social sciences, arts, and emerging interdisciplinary domains compared to Scopus, leading to content gaps and potential regional biases [6].

PubMed, operated by the US National Library of Medicine, provides an open-access platform focused on biomedical and health sciences literature, aggregating millions of citations through MEDLINE. While lauded for free and frequent updates, PubMed does not offer citation analysis and remains limited in disciplinary scope [5].

The Directory of Open Access Journals (DOAJ) plays a unique role in improving the discoverability of open-access, peer-reviewed journals worldwide [17] report that DOAJ's commitment to metadata standards and transparency supports the accessibility of open research. However, usability studies continue to identify areas for improvement, such as consistency in alternative text and keyboard navigation.

Despite their foundational importance, these traditional platforms share key limitations. Differences in indexing policies, subject coverage, and inclusion of document types (e.g., conference proceedings, regional or grey literature) result in both unique and overlapping content, with no single database offering exhaustive coverage[6] [5]. Scopus is often noted for its broader coverage, including conference proceedings and book chapters not found in WoS, while WoS's selectivity yields higher-quality but less comprehensive results, especially in document types like meeting abstracts and book reviews [18]. All rely primarily on keyword-based retrieval and Boolean logic, which, although precise, frequently fail to interpret user intent, synonymy, or paraphrased queries.

Given these challenges, cross-database integration has become increasingly important. Aggregating data from Scopus, WoS, PubMed, and DOAJ allows researchers and system designers to overcome individual database limitations, enhance recall, and build more robust, comprehensive academic discovery platforms[6] [19]. Such integration provides a richer foundation for advanced functionalities, including semantic search, hybrid retrieval, and knowledge graph-driven expansion, and is now recognized as a best practice in next-generation system design [20].

2) Advances in Academic Search Platforms

The past several years have marked a profound transformation in academic search, driven by advances in AI, machine learning, and large language models. While some established systems like Scopus and Web of Science have incorporated incremental AI enhancements, their core logic remains tied to keywords and citation metrics, with little adoption of advanced NLP or ontological reasoning[2]. These inherent limitations have fueled demand for next-generation platforms that offer true semantic understanding and more intuitive tools for exploration[2].

Tay [11] highlights that the emergence of transformer-based models like BERT and GPT has been the catalyst for this new era, enabling powerful capabilities in semantic search, retrieval-augmented generation (RAG), and intelligent information extraction.

Unlike conventional databases, recent AI-powered tools such as Elicit, ResearchRabbit, and Scite.ai utilize dense vector embeddings and neural ranking models to interpret queries in context, enabling semantic matching and producing synthesized answers with cited sources. These systems combine traditional keyword approaches (BM25, TF-IDF) with embedding-based semantic retrieval, frequently using hybrid strategies (cross-encoders, learning-to-rank) that outperform single-strategy systems, especially on broad or conceptually complex queries. Empirical evaluations such as the BEIR benchmark provide evidence for the superior performance and versatility of these hybrid models [11].

Parallel to these algorithmic advances, the importance of leveraging multiple reputable data sources has become a foundational principle in modern academic discovery[19] [6]. Scopus, Web of Science, and PubMed each contribute unique strengths and complementary coverage, but none offers complete coverage or perfect alignment with the needs of all disciplines or regions.

Studies of Idhris *et al.*[6] and Haron *et al.*[20]demonstrated that integrating data from several trusted sources yields more comprehensive bibliometric analyses, greater result accuracy, and higher user satisfaction. further describe the “Innovative Journal Finder System,” which integrates WoS and Scopus APIs and utilizes advanced search algorithms, citation metrics, and filtering mechanisms, including predatory journal detection, to provide credible and relevant journal recommendations.

Despite these significant advancements, substantial gaps persist. Most modern platforms remain focused on article-level retrieval, with limited support for Arabic and non-English languages, and a lack of advanced hybrid, knowledge graph-driven, and interactive visual features tailored for journal-level exploration.

B. Semantic Search and NLP in Academic Discovery

1) Keyword Search vs. Semantic Search

The development of scholarly search techniques shows a trend from traditional keyword queries to semantically enhanced search methods. In the face of an ever-increasing volume and heterogeneity of the scholarly digital content, researchers increasingly want search engines that not only give them exact results, but also ones that are capable of understanding the implicit intention and context of a query. Due to its growing demand, Natural Language Processing (NLP) technologies are now being incorporated into the academic search engines, thus resulting in a new way of interpreting the user query and retrieving related information.

Ahluwalia *et al.* [12] noted that traditional keyword-based retrieval, a staple of information systems, was based on the surface string matching of those terms, typically using models like BM25 for retrieving the matching document with those terms. Although sufficient for simple searching, it was recognized that such an approach was fundamentally limited in its ability to handle synonyms, paraphrasing, and the fact that natural language is context-rich (resulting in 'insufficient' or 'unhelpful' responses). They also discussed that these restrictions were remarkable, especially between different disciplines with variations in terminology and semantic ambiguity. Instead, they argued that semantic search was "a huge leap forward" based on recent breakthroughs in natural language processing and deep learning, allowing systems to understand the query and even the semantics and intent behind the user's search rather than just the individual words in the search query. By embracing architectures such as Transformers, semantic search systems, they contended, went beyond word matching, then deriving deeper meanings from queries and documents, to finally improve result relevance and user satisfaction.

Setlur *et al.* [21] explained that though keyword search continues to be effective for direct retrieval, it is inherently limited when users attempt to retrieve answers for structured, analytical, or open-ended questions. Semantic search approaches, enabled by NLP tools like entity recognition and relation extraction, are more likely to enable such applications by returning results that respond to not only the literal query but also its intent and context within larger databases. Hybrid approaches that blend keyword and semantic strategies are discovered to be particularly potent, allowing systems to satisfy a wider range of user requirements.

Wiklund & Maranan Hansson [22] highlighted that keyword-based systems are prone to missing relevant content due to their dependence on exact matches, whereas semantic models, using dense vector representations and embeddings, capture the broader context of queries, overcoming issues such as misspelling and paraphrasing.

Esteva *et al.* [23] provided practical evidence from biomedical information retrieval, showing that systems which combine NLP-based semantic search with traditional keyword methods, alongside features such as summarization and entity recognition, are able to answer user queries more precisely and

robustly even when the information sought is complex or nuanced.

The increasing volume and complexity of scientific and web content have underscored the need for search systems that can move beyond traditional keyword matching to truly understand user intent and context. As Khan and Malik [24] explained that traditional information retrieval methods had relied on syntactic matching, where query terms were directly compared with the contents of documents. They observed that this approach often resulted in large and imprecise result sets, as it failed to adequately handle ambiguity, redundancy, and the context-dependent nature of natural language. According to Khan and Malik, even minor variations in phrasing or the use of synonyms could significantly alter retrieval outcomes, highlighting the inherent limitations of keyword-based search in capturing user intent or the semantic relationships between concepts.

Cujar-Rosero *et al.* [25] provided a practical illustration of these principles in their development of the FENIX semantic search engine. Their work demonstrated that semantic search systems, built on ontologies and machine learning models trained with NLP algorithms (such as SpaCy, NLTK, Word2Vec, and Doc2Vec), can unify queries and results by focusing on meaning rather than surface word matching.

Further, Latard *et al.* [26] also noted that, with scientific publications growing exponentially, keyword search alone is no longer sufficient. Their paper emphasizes that linking keywords through semantic relations such as synonyms, categories, or domains, through resources like BabelNet, enables the retrieval of articles based on conceptual similarity rather than the literal presence of words. Not only does this enhance precision, it also introduces contextual filtering that is especially critical in scientific research, where the same keyword can mean different things across disciplines.

Collectively, these studies indicate that NLP conforms to semantic technologies such as knowledge graphs, semantic parsing, deep learning language models, etc. The integration allows current search engines to better understand, disambiguate, and answer complex and subtle search tasks that are not collapsed to the matching of terms in queries and documents. Traditional keyword search formed the basis of academic discovery tools over the years, but NLP-driven semantic search is an evolution required to match the needs of today's researchers.

2) Word embeddings and document embeddings

The accelerated growth of scientific literature, combined with the increasing sophistication of research topics, has rendered traditional keyword-based search systems ever more insufficient for traversing academic knowledge. Researchers are frequently confronted with large, multidisciplinary document sets and need tools that fetch information based on meaning and context, not superficial keyword overlap. This has generated an intense need for retrieval systems that can interpret both the literal terms and the underlying semantics. Consequently, information retrieval research has moved beyond straightforward lexical matching to representation learning, where words, sentences, and documents are mapped into dense vector spaces that embed semantic relationships and contextual subtleties. This shift has been well-traced by Khan & Malik [24].

One key milestone in this change was brought about by Mikolov *et al.*[27] with Word2Vec, a model that learns distributed word representations by examining co-occurrence patterns within large corpora. Shortly after. Pennington et al [28] took the step further with GloVe, which added global co-occurrence statistics to generate even more refined embeddings. These models saw extensive usage in tasks ranging from text classification and sentiment analysis to information retrieval. However, as observed by Reimers & Gurevych [10] both Word2Vec and GloVe generate static embeddings; each term is assigned one vector, independent of context, thus proving less useful in dealing with polysemy or in capturing domain-specific differences, especially within scientific texts.

To overcome these limitations, Devlin *et al.*[9] created BERT, a transformer-based model that produces contextual embeddings by taking into account bidirectional sentence context. BERT brought huge gains across tasks such as question answering, named entity recognition, and reading comprehension. However, its architecture was not ideal for the creation of fixed-size sentence embeddings required for efficient similarity calculations. In reaction, Reimers and Gurevych [10] brought forth SBERT, which uses a Siamese network architecture to produce semantically meaningful sentence embeddings, making semantic search and clustering faster and more precise.

Following this, Song *et al.*[29] came up with MPNet, which extended masked language modeling with permuted language modeling to better capture contextual dependencies. This was followed by the creation of all-mpnet-base-v2 in the SentenceTransformers library, which merged MPNet with SBERT's Siamese setup. The outcome was a state-of-the-art model that surpassed numerous predecessors in semantic similarity benchmarks and was found to be effective in applications like document retrieval, paraphrase detection, and question answering.

With these findings in mind, our framework leverages all-mpnet-base-v2 to generate contextual embeddings for both user queries and journal metadata. This enables the evaluation of semantic similarity through the inclusion of domain-specific terms and context variations. The framework thus presents a flexible, efficient, and semantically meaningful solution to keyword-based searches and static embeddings, improving the accuracy and relevance of academic journal retrieval.

III. METHODOLOGY

A. System Architecture and Design

This academic journal search engine uses a modular design for scalability and intelligent features. The backend, built with the high-performance Python framework FastAPI, handles core logic and coordinates with other system components. A standard web front end (HTML, CSS, JavaScript) provides a user-friendly interface for search and interactive visualizations.

For storage and search, it uses a hybrid approach: Elasticsearch for semantic and full-text search, and Neo4j, a

graph database, to map relationships between journals for recommendations. A key component is the NLP and Semantic Embedding Pipeline. It processes queries and uses the SBERT model to represent journals and queries as vectors for semantic similarity searches.

The Semantic Search and Recommendation Engine is a multi-stage process that starts with a title search and can expand to a more complex semantic search using NLP and vector similarity. Results are ranked and fused, with recommendations generated from various similarity signals.

To visualize the interactions between components, the system architecture can be summarized in Figure 1.

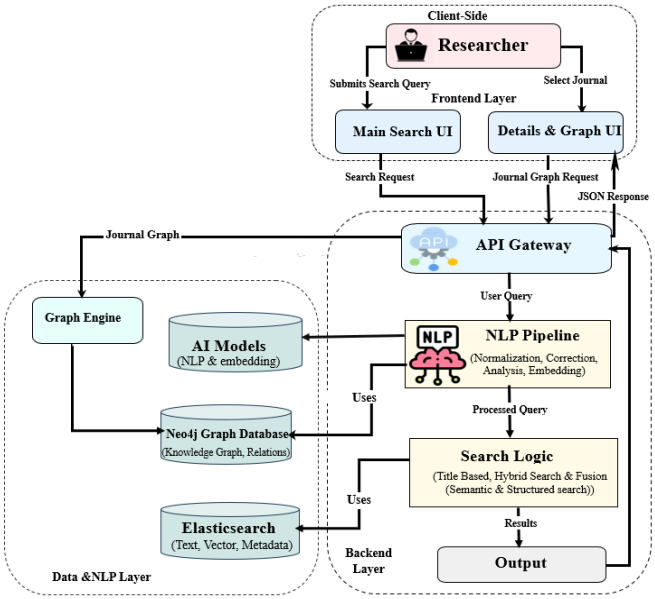

Fig. 1. High-Level Architectural Diagram of the Proposed Method

B. *Data Collection, Integration, And Preparation*

The dataset was constructed from journal metadata obtained from well-established academic databases, including Scopus, DOAJ, and PubMed, chosen for their broad coverage and credibility. Data acquisition involved official APIs (e.g., DOAJ), public CSV exports from Scopus, and limited web scraping for missing attributes. All collected data was then merged into a centralized raw dataset.

A multi-step cleaning process addressed inconsistencies, missing fields, and duplicate records. This included removing null fields, deduplicating based on unique identifiers (ISSN/EISSN), and correcting typographical errors using domain-specific dictionaries. This ensured a reliable and clean dataset. Data normalization was followed to enhance consistency for filtering and retrieval. Subject categories were aligned with a unified taxonomy, and publisher names, indexing databases, and access policies were harmonized. Numerical indicators like the impact factor were standardized. This significantly improved interoperability and compatibility with the system's filter extraction engine.

The raw metadata was then enriched with critical attributes such as quartile rankings (Q1–Q4) and impact metrics (CiteScore, H-index) from external sources like Scopus and SCImago Journal Rank. Subject classifications were harmonized, and indexing information (Scopus, WoS, DOAJ inclusion) was explicitly annotated. This enrichment phase enhanced the system's ability to support meaningful filtering, ranking, and recommendations.

Each journal record was semantically encoded using Sentence-BERT (SBERT) to produce high-dimensional dense vectors from journal descriptions and editorial aims. These embeddings were stored alongside traditional metadata in Elasticsearch, facilitating hybrid search and serving as primary features for the recommendation component.

Finally, the prepared dataset was indexed into Elasticsearch using custom mappings for multi-field retrieval, vector similarity queries, and structured filtering. Relational aspects were also inserted into Neo4j for graph-based querying. This comprehensive workflow created a robust and semantically aware foundation for the system's natural language understanding, retrieval, and recommendation modules.

Figure 2 visualizes the entire staging process, from raw data acquisition to final indexing, unifying the data preparation workflow.

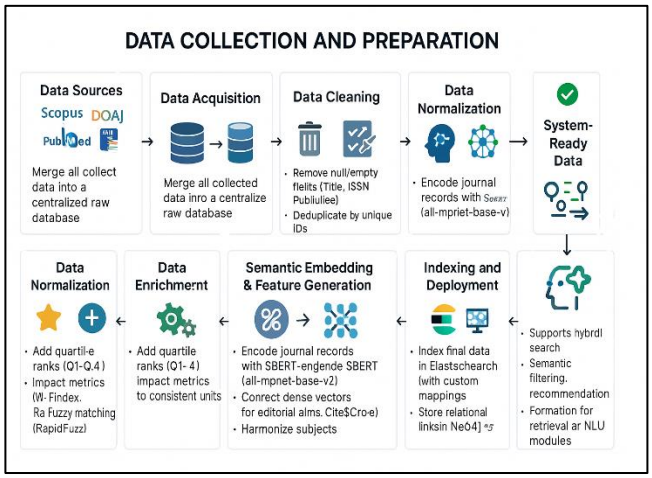

Fig. 2. The End-to-End Data Collection and Preparation Pipeline

C. *Natural Language Processing and Semantic Embedding*

For accurate interpretation of user intent and retrieval of conceptually related journals, the system incorporates a comprehensive natural language processing (NLP) and semantic embedding pipeline. This pipeline converts both users' queries and journal descriptions into a form suitable for semantic similarity and structured filtering. The architecture enhances robustness and multilingual capabilities, beginning with text preprocessing and normalization. This stage standardizes input by removing excess whitespace, harmonizing punctuation, unifying character encodings, and lowercasing text for consistent downstream processing. For Arabic queries, a custom multilingual sub-pipeline detects the language, applies spelling correction using SymSpell with domain-specific dictionaries, and translates the query into English via Google Translate to

ensure compatibility with the English-based embedding model. This enables Arabic-speaking users to query naturally while benefiting from English-trained semantic retrieval. The refined query then undergoes advanced query correction and intent refinement within a semantic alignment module. This component improves precision by handling ambiguous input through three hierarchical layers: manual academic corrections, contextual enrichment (e.g., expanding acronyms), and controlled spell correction targeting non-technical terms. This layered approach ensures linguistic precision and semantic fidelity, crucial for scientific literature.

The core of the semantic retrieval engine is the generation of high-dimensional embeddings for user queries and journal descriptions using the pre-trained Sentence-BERT (SBERT) "all-mpnet-base-v2" model, highly effective for semantic similarity in scientific data. Journal summaries are encoded offline, while user queries are encoded online at search time. These dense vector representations capture both syntactic and semantic content, indexed and searched through Elasticsearch's K-Nearest Neighbors (KNN) engine for approximate nearest neighbor (ANN) search. This embedding strategy allows for the retrieval of semantically similar journals even without keyword matches, enabling concept-based and interdisciplinary exploration.

A Query Analysis module is also integrated to support fine-grained attribute filtering alongside semantic search. This module transforms unstructured user queries into structured filters for bibliometric constraints, indexers, disciplines, and qualitative indicators. The process includes Discipline Identification via a Neo4j knowledge base, matching both root and sub-fields based on token overlap. Indexing Database Detection uses direct and fuzzy substring matching to identify platforms like Scopus or DOAJ. Numerical Filter Extraction employs advanced regular expressions to parse constraints on Impact Factor, CiteScore, and H-index, supporting flexible syntax. Quartile Recognition detects and normalizes mentions of quartiles (e.g., "Q1"). Finally, Context-Aware Qualitative Phrase Mapping interprets subjective terms like "prestigious" dynamically. This involves offline statistical profiling to compute percentile values for metrics within each academic discipline, followed by dynamic threshold resolution during real-time query analysis. The system identifies the discipline and resolves the qualitative phrase to a dynamic threshold based on that discipline's statistical profile, ensuring contextually appropriate interpretation.

This tightly integrated NLP–retrieval pipeline ensures that the system supports a wide spectrum of search behaviors, from precise filtering to exploratory browsing, across languages and academic domains. The entire query processing pipeline, spanning language detection, multilingual preprocessing, semantic embedding, and structured filter extraction, is visualized in Figure 3.

D. Search And Retrieval Mechanisms

The search and retrieval pipeline in the system embodies a carefully orchestrated integration of traditional information retrieval (IR) techniques, semantic embedding-based reasoning, and structured field-level filtering. This architecture is designed to robustly accommodate diverse user intents, whether expressed through precise journal titles, abstract thematic phrases, or multi-criteria filtering conditions. The primary technologies that facilitate this hybrid approach include Elasticsearch for vector and keyword indexing, and Neo4j for topic expansion and identifier mapping. The endpoint /search in the FastAPI backend encapsulates this process within a unified search interface.

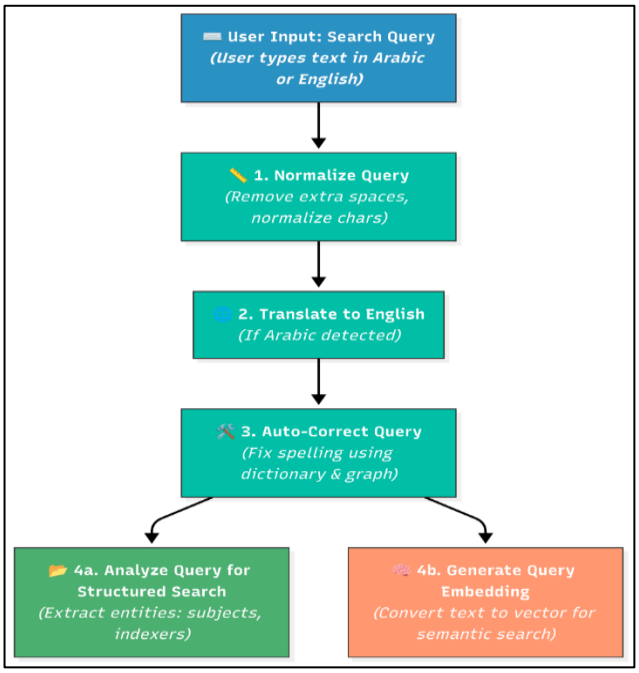

Fig. 3. Architectural Flowchart of the Query Processing Pipeline

Following preprocessing, the retrieval engine executes conditional and parallel strategies. Initial handling involves a cache lookup; if a cached response exists for the query, it is returned immediately. Otherwise, the system performs a direct title-based search on Elasticsearch using a hybrid strategy combining strict term-level matching and general phrase-level matching. This dual-layered approach robustly detects both exact and approximate journal title matches. If this search yields a single matching journal, the system assumes high confidence and enriches the result set with semantic recommendations by using the matched journal's embedding vector for a secondary semantic search. This expands results with thematically similar journals, improving recall and supporting exploration, with duplicate journals filtered out.

In cases where the initial title search is insufficient or the query is abstract, thematic, or exploratory, the system transitions into a dual retrieval mode, executing both semantic and structured searches in parallel. Semantic search leverages a fine-tuned Sentence-BERT (SBERT) model to embed the user query into a high-dimensional vector. This vector is then passed to Elasticsearch's vector search engine, which performs a k-Nearest Neighbor (kNN) search on pre-computed journal description embeddings. The system finds the 'k' closest candidates with the least cosine distance, followed by post-filtering to ensure semantic matching, particularly useful for

interdisciplinary or emerging topics. Concurrently, structured search uses metadata constraints derived from query analysis, such as subject category, indexing databases, quality indicators (e.g., Impact Factor), or quartile placement. These structured queries are formulated using Elasticsearch's Boolean query syntax with term, range, and filter clauses. This provides precise filtering for journals meeting explicit user requirements, such as "open-access Q1 journals from DOAJ, indexed in Scopus with an Impact Factor over 3." The parallel execution of semantic and structured retrieval offers both interpretive depth and precise filtering; semantic search generalizes the query scope, while structured search narrows it with exact matching.

The system uses an advanced fusion and ranking strategy to combine results from parallel semantic and structured searches into a single, unified list via the smart_fusion_results function. This adaptive scoring approach combines relevance scores, result list position, and query context. The process maps each retrieved journal to its unique serial_number, calculating a semantic_score (cosine similarity) and a structured_score (BM25 with domain-specific boosting). Crucially, each initial score is modulated by the reciprocal of its rank to give higher weight to top-ranked results. Journals present in both semantic and structured results receive a multiplicative confidence bonus (1.4x). The fusion logic employs a dynamic weighting scheme, adapting to the query type. If structured constraints are present, a higher weight (e.g., 90%) is assigned to the structured_score to prioritize precision. For general, open-ended queries without filters, a dominant weight (e.g., 70%) is given to the semantic_score for thematic discovery. The final score is a weighted sum of rank-adjusted component scores, multiplied by the confidence bonus. The consolidated list is then sorted by this final fused score, truncated to the top-k results. This adaptive fusion balances the broad generalization of semantic search with the precise filtering of structured search. Following retrieval and ranking, a formatting module standardizes each journal entry for uniform output, including essential metadata like publication identifiers, subject classifications, and bibliometrics. A caching mechanism stores these results for recurring queries, and an analytics layer logs transaction data to support continuous system improvements. The /search endpoint end-to-end pipeline is depicted in Figure 4. It starts with cache validation and continues through a priority chain: a title matching, a semantic and structured retrieval, and fusion, up to results presentation. Its behavior is dynamically adjusted according to the specificity of the query and the availability of results, to guarantee that a balanced trade-off between efficiency and retrieval performance is maintained.

E. User Interface Illustrations

The search interface forms the primary entry point into the system's interactive environment, designed to combine semantic input processing with an intuitive, user-friendly layout. As depicted in Figure 5, the interface is characterized by a centered search field accompanied by minimal but effective filter buttons, namely Scopus Indexed, Open Access, and High Impact Factor. These components are carefully arranged to reduce cognitive load while supporting structured query formulation.

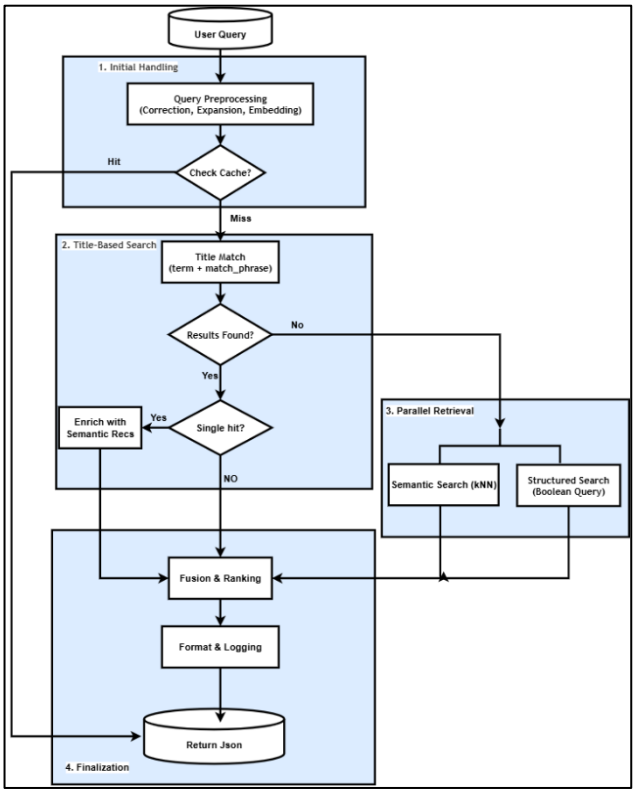

Fig. 4. Architectural Overview of the Search and Retrieval System.

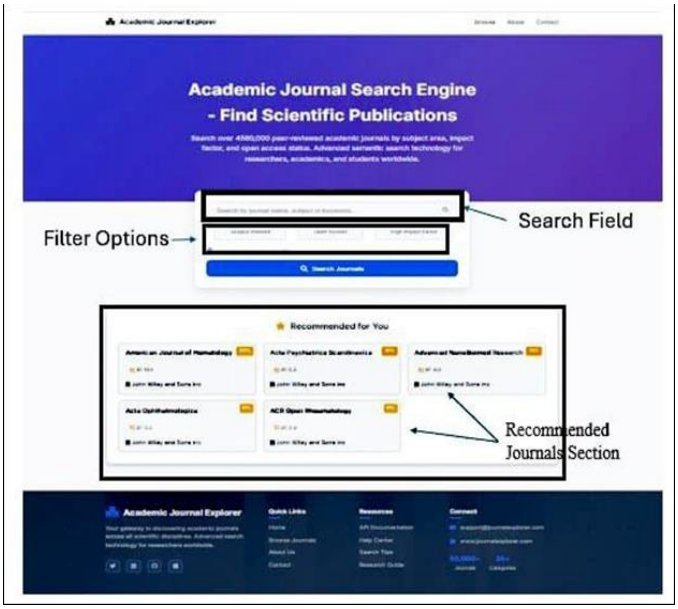

Fig. 5. The Main Search Interface of The System.

When a user enters a search phrase such as journal names, academic domains, or general keywords, the system applies the multi-path Retrieval Strategies that were discussed previously and retrieves the most relevant journal records from the backend database.

Upon submission of a query, the system presents a visually structured list of journal entries directly below the search area.

Each journal card includes (Title, Impact Factor (IF) and Citation Score (CS), Quartile Ranking (Q1, Q2, etc.), Publisher name, Key subject tags, Indexing status (e.g., Scopus, DOAJ, PubMed), Action buttons for Details and Visit Site), as shown in Figure 6.

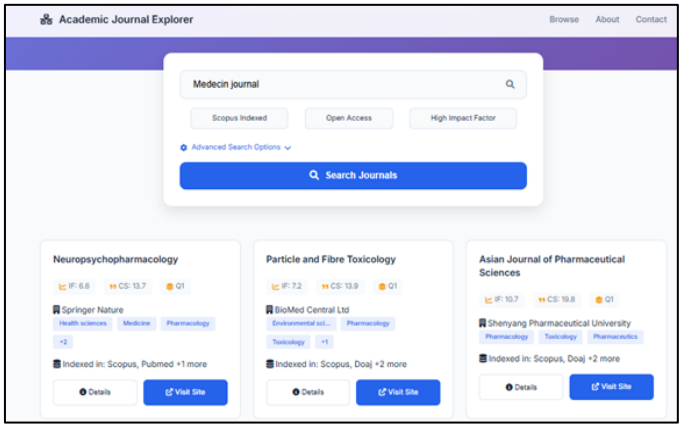

Fig. 6. Search Results Interface

The modular, card-based layout with rich, color-coded, tag-based metadata allows users to quickly scan, compare, and visually parse journal attributes. This presentation model facilitates rapid evaluation of academic prestige, thematic relevance, and other critical publishing information at a glance. It supports informed decision-making by combining structured filters, intelligent suggestions, and visual clarity to enhance the academic discovery experience.

Upon selecting a journal, the system displays a detailed, interactive profile, as shown in Figure 7. This view comprises two interconnected parts: a metadata panel and a graph visualization. The left panel provides comprehensive metadata (publisher, impact factor, quartile, indexing, etc.), enabling users to efficiently evaluate a journal's academic standing and indexing.

On the right, an interactive network visualization places the selected journal at the center of an auto-generated graph. Nodes representing other journals are connected based on shared metadata like subjects, indexing databases, or publishers, explicitly modeling scholarly ecosystem relations.

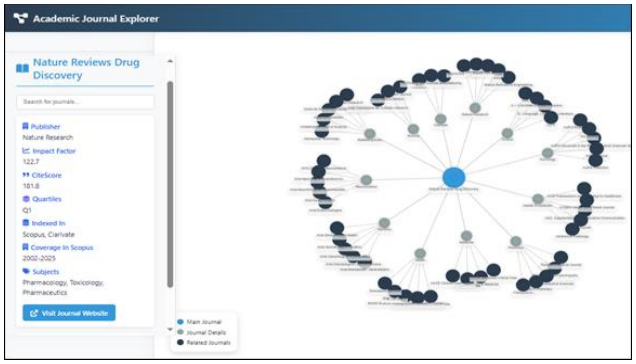

Fig. 7. Details View with Related Journals Graph

IV. Results And Discussion

A. Quantitative Retrieval Performance

The retrieval effectiveness of the proposed hybrid engine was evaluated using two primary rank-aware metrics: Precision at 10 (P@10) and Normalized Discounted Cumulative Gain (nDCG@10). The evaluation was conducted on a curated dataset of real-world academic queries to ensure practical relevance.

As shown in Table I, the system achieved a mean P@10 of 0.808 and an nDCG@10 of 0.889. These scores indicate that the system not only retrieves relevant journals but consistently ranks the most pertinent results at the top of the list. The high nDCG score specifically validates the effectiveness of the SBERT-based semantic component in capturing user intent beyond mere keyword matching.

TABLE I. Overall Retrieval Performance Metrics

Metric	Mean Value	Standard Deviation
Precision@10	0.808	0.234
nDCG@10	0.889	0.133

To isolate the contribution of the hybrid retrieval strategy, we compared the proposed configuration against two baseline methods: (1) a traditional keyword-based BM25 retrieval implemented using Elasticsearch, and (2) a pure semantic search pipeline relying solely on Sentence-BERT (SBERT) embeddings without incorporating structural journal metadata. This comparison enables a clearer assessment of whether combining lexical, semantic, and structured signals provides measurable advantages over simpler alternatives.

Table II summarizes the results. While both baselines demonstrate reasonable effectiveness, the hybrid approach consistently outperforms them across all metrics. The improvement over the keyword-only baseline highlights the limitations of exact term matching for capturing conceptual similarity, whereas the gains over the semantic-only setup demonstrate the additional value of integrating structured journal attributes such as quartile ranking, indexing status, and subject filters. These findings validate the design choice of combining heterogeneous retrieval signals within a unified framework.

TABLE II. Retrieval Performance Comparison Against Baseline Methods

Method	P@10	nDCG@10
BM25 (keyword-based)	0.64	0.71
SBERT only (semantic)	0.75	0.82
Hybrid (proposed)	0.808	0.889

A comparative linguistic analysis (Table III) reveals that while English queries perform slightly better (P@10 = 0.818), the Arabic retrieval performance remains robust at 0.797. The marginal performance gap (approx. 2.1%) is attributed to the inherent complexities of machine translation in handling highly

specialized academic terminology, a challenge that could be further mitigated by integrating native Arabic semantic models in future iterations.

TABLE III. RETRIEVAL PERFORMANCE BY QUERY LANGUAGE

Language	Precision@10	nDCG@10
Arabic	0.797	0.875
English	0.818	0.903

These results are competitive with recent AI-assisted academic discovery tools such as Elicit and ResearchRabbit, which primarily focus on article-level retrieval. In contrast, the proposed system is specifically designed for journal-level discovery while maintaining comparable retrieval accuracy.

B. System Efficiency and Scalability

To assess the system's viability for production-level deployment, load testing was conducted by simulating 120 concurrent retrieval requests. Across multiple test runs, the architecture leveraging FastAPI, Elasticsearch, and Neo4j demonstrated exceptional reliability with a 100% success rate. The average response time remained between 3.08 and 3.37 seconds, which falls within the acceptable threshold for interactive search systems, even under moderate concurrent loads.

C. User-Centric Qualitative Evaluation

A user study involving 40 academic participants (PhD candidates, researchers, and faculty) was conducted to measure perceived utility. The feedback was overwhelmingly positive:

- Usability: 88% of participants rated the interface as "Easy" or "Very Easy" to navigate.
- Visual Exploration: A significant 93% of users highlighted the interactive graph-based visualization as a transformative feature for exploratory discovery.
- Preference: 79% of users ranked the proposed system as superior to their current search tools due to its integrated, multi-constraint handling capabilities.

D. Scenario-Based Functional Comparison

To further position the proposed approach within the current ecosystem of academic discovery tools, a scenario-based functional comparison was conducted against representative platforms that support literature exploration and scholarly search. The objective of this analysis is not to compare retrieval accuracy, but to evaluate how effectively each system supports the practical task of journal selection and venue recommendation.

1) Semantic Scholar (Article-Oriented Retrieval)

Semantic Scholar is primarily optimized for discovering individual research papers and citation relationships. When issuing venue-oriented queries such as “Q1 open access journals in life sciences”, the system predominantly returns articles rather than journals. As a result, users must manually inspect sources and infer suitable venues, introducing additional steps and cognitive overhead. In contrast, the proposed system treats journals as first-class entities and directly returns ranked publication venues enriched with structural metadata (quartile, impact metrics, indexing, and access type), thereby aligning more closely with researchers’ submission needs.

2) Connected Papers (Paper-Level Graph Exploration)

Connected Papers provides interactive citation graphs to explore relationships among papers. While effective for understanding the evolution of research topics, it depends on a seed paper and remains inherently article-centric. The platform does not expose journal-level attributes or recommendation mechanisms. Conversely, the proposed system constructs a journal-centric knowledge graph that models relationships among venues based on subject similarity, publishers, and bibliometric indicators, enabling direct exploration of the publication landscape without requiring an initial document.

3) AI-Assisted Literature Discovery Tools

AI-based tools such as Elicit, ResearchRabbit, and Scite introduce advanced features, including automated summarization, citation network visualization, and citation sentiment analysis. However, their unit of analysis remains the individual paper. They assist researchers in reviewing literature but provide no support for filtering or recommending publication venues. Accordingly, these systems complement the literature review process but do not address the distinct problem targeted in this work: intelligent journal discovery and recommendation under semantic and structural constraints.

E. Comparative Summary

The presented results in Table IV show that existing platforms predominantly focus on paper-level discovery and analysis, whereas the proposed framework uniquely delivers venue-level intelligence by integrating semantic retrieval, structured filtering, and journal-centric graph analysis within a unified workflow. This specialization directly supports researchers in identifying appropriate publication outlets, thereby addressing a gap not covered by current literature exploration tools.

V. LIMITATIONS AND FUTURE WORK

Despite the promising results, this study acknowledges several limitations that provide opportunities for future research:

- Translation Precision: The system currently relies on general-purpose translation APIs, including tools similar to Google Translate, to process Arabic queries. While effective for general academic language, such systems may occasionally misinterpret highly specialized technical terminology. Future work could integrate domain-adapted multilingual models such as AraBERT or mBERT to enable native semantic processing of Arabic academic queries without intermediate translation.
- Dataset Scale and Coverage: Although the current prototype indexes journals from major databases such as Scopus, DOAJ, and PubMed, the repository is not yet exhaustive. Scaling the system to support millions of journals and additional regional or domain-specific repositories will require distributed indexing strategies and optimized vector search infrastructure.

- Cold-Start Problem: The recommendation module relies primarily on existing journal metadata, textual descriptions, and semantic embeddings. Newly indexed journals with limited metadata, citation history, or descriptive information may experience reduced visibility in the graph-based recommendation network.
- User Study Demographics: While the user evaluation involving 40 academic participants provided valuable usability insights, the participant pool was primarily drawn from a limited set of academic environments. Future evaluations should involve a larger and more geographically diverse research community to improve the generalizability of the findings.

TABLE I. FUNCTIONAL COMPARISON OF EXISTING DISCOVERY TOOLS AND THE PROPOSED JOURNAL RECOMMENDATION SYSTEM

Capability	Semantic Scholar	Connected Papers	Elicit	ResearchRabbit	Scite	Proposed System
Paper discovery	✓	✓	✓	✓	✓	Partial
Journal discovery	Partial	✗	✗	✗	✗	✓
Natural language semantic search	Partial	✗	✓	Partial	Partial	✓
Structural filtering (quartile, IF, indexing)	✗	✗	✗	✗	✗	✓
Graph visualization	✗	✓ (papers)	✗	✓ (papers)	✗	✓ (journals)
Direct journal recommendation	✗	✗	✗	✗	✗	✓
Arabic query support	✗	✗	✗	✗	✗	✓

VI. CONCLUSION

This study introduced an intelligent academic journal search system that integrates lexical search, semantic embeddings, and knowledge graph exploration within a unified hybrid architecture. The system supports bilingual Arabic–English queries and enables researchers to discover journals based on both conceptual similarity and bibliometric attributes.

Experimental evaluation demonstrated strong retrieval performance, achieving a Precision@10 of 0.808 and an nDCG@10 of 0.889. The user study further confirmed that the graph-based visualization significantly enhances exploratory discovery.

These findings suggest that hybrid semantic retrieval combined with structured metadata filtering can significantly improve journal-level discovery. The proposed framework provides a scalable foundation for next-generation academic search platforms.

REFERENCES

[1] C. Kacperski *et al*., "Audit of academic search engines 2 Examining bias perpetuation in academic search engines: an algorithm audit of Google and Semantic Scholar," 2023.

[2] R. Verma and S. Sharma, "Scopus: a comprehensive literature review," *Int J Prof Dev*, vol. 11, no. 2, pp. 107–110, 2022.

[3] M. Gusenbauer and N. R. Haddaway, "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources," *Res. Synth. Methods*, vol. 11, no. 2, pp. 181–217, Mar. 2020, doi: 10.1002/jrsm.1378.

[4] R. Pranckutė, "Web of Science (WoS) and Scopus: the titans of bibliographic information in today's academic world," Mar. 01, 2021, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/publications9010012.

[5] R. Hosseiniara, "General comparison of scientific databases of Scopus, PubMed, and Web of Science," *www.nclinmed.com Novelty in Clinical Medicine*, vol. 2, no. 3, pp. 168–169, 2023, doi: 10.22034/NCM.2023.380213.1059.

[6] M. Idhris, M. Peter, M. B. Ali, and A. Pandiyarajan, "Library Knowledge Management (LKM) assessment comparison between Scopus and web of Science: A Bibliometric view," *Library Philosophy and Practice*, 2021.

[7] S. Khalid, S. Almutairi, A. Namoun, J. Khan, H. Ali Khattak, and H. Shah, "Comprehensive review of academic search systems: evolution, analysis, and future research directions," Dec. 01, 2025, *Springer*. doi: 10.1007/s13278-025-01476-1.

[8] V. Ogunrinde, A. Clarke, N. Hughes, and P. Banerjee, "Semantic Search and Its Role in Knowledge Discovery." [Online]. Available: https://www.researchgate.net/publication/392014461

[9] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.10084

[11] A. Tay, "Academic search and discovery tools in the age of AI and large language models: An overview of the space," 2024.

[12] A. Ahluwalia, B. Sutradhar, K. Ghosh, I. Yadav, A. Sheetal, and P. Patil, "Hybrid Semantic Search: Unveiling User Intent Beyond Keywords," 2024.

[13] J. Yang, F. Jiang, and T. Baldwin, "Language Bias in Multilingual Information Retrieval: The Nature of the Beast and Mitigation Methods," in *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, J. Sälevä and A. Owodunni, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 280–292. doi: 10.18653/v1/2024.mrl-1.23.

[14] H. Alotaibi, "The research gap in the introductions of Arabic research articles," Sep. 2016.

[15] D. Lawrie, E. Yang, D. W. Oard, and J. Mayfield, "Neural Approaches to Multilingual Information Retrieval," in *Advances in Information Retrieval*, J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo, Eds., Cham: Springer Nature Switzerland, 2023, pp. 521–536.

[16] S. Saleh and P. Pecina, "Document translation vs. query translation for cross-lingual information retrieval in the medical domain," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6849–6860.

[17] B. Marino and K. F. Mason, "Exploring Accessibility in DOAJ: A Case Study," *Serials Review*, vol. 46, no. 2, pp. 82–90, Apr. 2020, doi: 10.1080/00987913.2020.1782632.

[18] M. Visser, N. J. van Eck, and L. Waltman, "Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic," *Quantitative Science Studies*, vol. 2, no. 1, pp. 20–41, Apr. 2021, doi: 10.1162/qss_a_00112.

[19] P. Khurana, G. Ganesan, G. Kumar, and K. Sharma, "A Comparative Analysis of Unified Informetrics with Scopus and Web of Science,"

Journal of Scientometric Research, vol. 11, no. 2, pp. 146–154, May 2022, doi: 10.5530/jscires.11.2.16.

[20] N. H. Haron, N. H. Abd Samad, R. Mahmood, F. B. Hamzah, W. A. Wan Muhamad Tahir, and M. A. Mohd Yusof, "Innovative Journal Finder System: Enhancing Research Outcomes in Higher Learning Institutions through WOS and Scopus Integration," *Journal of Advanced Research in Technology and Innovation Management*, vol. 14, no. 1, pp. 1–9, Mar. 2025, doi: 10.37934/jartim.14.1.19.

[21] V. Setlur, A. Kanyuka, and A. Srinivasan, "Olio: A Semantic Search Interface for Data Repositories," in *UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, Association for Computing Machinery, Inc, Oct. 2023. doi: 10.1145/3586183.3606806.

[22] E. Wiklund and I. K. Maranan Hansson, "Semantic search in historical documentation," 2024.

[23] A. Esteva *et al.*, "COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization," *NPJ Digit. Med.*, vol. 4, no. 1, p. 68, 2021.

[24] A. A. Khan and S. K. Malik, "Semantic search revisited," in *2018 8th international conference on cloud computing, data science & engineering* (confluence), IEEE, 2018, pp. 14–15.

[25] F. Cujar-Rosero, D. S. P. Ortiz, S. R. T. Pereira, and J. M. G. Restrepo, "Fenix: A Semantic Search Engine Based on an Ontology and a Model Trained with Machine Learning to Support Research," in *CS & IT Conference Proceedings*, CS & IT Conference Proceedings, 2021.

[26] B. Latard, J. Weber, G. Forestier, and M. Hassenforder, "Towards a semantic search engine for scientific articles," in *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings 21*, Springer, 2017, pp. 608–611.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781

[28] J. Pennington, R. Socher, and 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.

[29] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 16857–16867, 2020.