

Course Title: Research Methodology in Computer Science

Course Code: AI800

Lecturer: Dr. Mohammed Anbar

Prepared by: Mahmoud Mohammed Alra'e

***Literature Review of Semantic Search in AI for
Automated Matching***

Abstract

This literature review explores the application of semantic search within artificial intelligence (AI) to solve the problem of matching item descriptions in customer purchase orders to corresponding items in a database. Semantic search interprets the intent and context of text, rather than merely matching keywords, and thus promises to improve the accuracy and efficiency of product matching. Six research papers were selected based on their relevance to semantic search, retrieval-augmented generation, knowledge graph integration, multilingual retrieval, and the wider technological context of AI-driven search systems. The review provides concise summaries, critical analysis, connections, conflicts, and a synthesis matrix summarizing the key aspects of each study.

Summary research paper's:

Paper 1 – Semantic approaches for query expansion: taxonomy, challenges, and future research directions (Allahim et al., 2025)

Allahim et al. (2025) present a comprehensive survey of semantic query expansion (QE) techniques developed between 2015 and 2024. The study categorizes existing approaches into three main groups: linguistic-based methods, which use lexical resources like dictionaries and thesauri; ontology-based methods which rely on structured domain knowledge; and hybrid methods that combine linguistic tools with modern artificial intelligence models such as word embedding and transformers. The authors analyze each category in terms of flexibility, scalability, and contextual understanding, noting that linguistic methods are broadly adaptable, ontology methods excel in domain-specific applications, and hybrid approaches offer the most effective semantic depth by capturing conceptual relationships. The paper identifies key challenges facing semantic QE, including the limited availability of multilingual ontologies, difficulties in real-time data processing, and the need to extend semantic expansion to non-textual data such as images and videos. Allahim et al. emphasize the importance of automatic ontology construction, context-aware retrieval, and efficient management of query drift to enhance future systems. Overall, the survey provides a structured framework for understanding current semantic QE techniques and highlights directions for future research toward more adaptive, context-sensitive, and multimodal semantic search systems.

Paper 2 – Secure semantic search using deep learning in a block chain-assisted multi-user setting (Khan et al., 2024)

Khan et al. (2024) present SDBMS, a secure semantic search framework designed for multi-owner, multi-user cloud environments. The system employs deep-learning-based document embedding to capture semantic meaning and combines them with attribute based encryption to ensure data confidentiality. A block chain layer manages decentralized key distribution and access control by removing the need for a single trusted authority and enabling transparent activity logging. Search relevance is determined through a truncated cosine similarity that balances semantic and exact term matching. The model is tested on the 20 Newsgroups dataset, achieving an average precision of roughly 88 % for the top 50 results. Experimental results show effective retrieval accuracy and strong privacy protection. The approach introduces higher computational overhead and remains untested in large scale or cross domain contexts limiting immediate real world deployment potential.

Paper 3 – Evaluation of Semantic Search and its Role in Retrieval-Augmented Generation (RAG) for Arabic Language (Mahboub et al., 2024)

The authors construct a dataset of 2030 Arabic customer-support call summaries and generate 406 search queries with relevance labels using GPT-4 【453049840448257†L153-L183】. They evaluate five multilingual embedding models using $nDCG@3$, $MRR@3$ and $mAP@3$; the Paraphrase-Multilingual-MPNet model achieves the highest scores ($nDCG@3 \approx 0.879$, $MRR@3 \approx 0.911$, $mAP@3 \approx 0.888$) 【453049840448257†L403-L422】. For retrieval-augmented generation, they pair each query with semantically similar FAQs and use GPT-3.5 to generate answers evaluated by GPT-4; the best model yields $\approx 63\%$ top-3 accuracy 【453049840448257†L449-L463】. The authors note that retrieval and generation performance do not always align, especially when comparing symmetric (question–question) and asymmetric (query–document) search 【453049840448257†L472-L487】

Paper 4 – Semantic Search and Retrieval-Augmented Generation for Academic Literature Analysis Using ColBERT (Yadav & Gopinathan, 2025)

Yadav and Gopinathan (2025) propose a retrieval-augmented generation (RAG) pipeline designed to enhance academic literature analysis through semantic search. Their system automatically extracts metadata—such as titles, abstracts, and categories—from large research datasets like arXiv, and represents this information using dense embedding generated by a Hugging Face transformer model. These embedding are stored in Chroma DB, an open-source vector database optimized for similarity search. When users submit a query the system retrieves the most relevant results and applies ColBERT (Contextualized Late Interaction over BERT) for re-ranking that improves precision by considering contextual token level similarity rather than global embedding alone.

The platform then uses a generation component to produce concise context aware summaries of the retrieved papers providing researchers with curated collections on specific topics such as knowledge graphs or diffusion models. Evaluation results indicate high user satisfaction scores (4.6–4.8/5) and relevance ratings between 0.65 and 0.76 outperforming baseline retrieval methods by 15–20% in both precision and recall. Despite these promising results, the study’s dataset details and evaluation depth remain limited reducing reproducibility and generalizability. Nevertheless the approach demonstrates the strong potential of combining ColBERT based semantic retrieval with RAG for academic research discovery.

Paper 5 – An Overview of Semantic Search Engines (Roy et al., 2019)

Roy et al. (2019) present a comprehensive overview of semantic search engines outlining how they differ from traditional keyword-based systems by focusing on contextual meaning and user intent. The paper traces the development of semantic search within the broader evolution of the Semantic Web emphasizing how understanding relationships between concepts allows for more intelligent information retrieval. The authors review several major semantic search engines such as Hakia, Kngine, Kosmix, DuckDuckGo Powerset and Sensebot and compare their architectures search mechanisms result formats and multilingual capabilities. Unlike conventional search engines that return simple lists of links semantic engines often provide aggregated summaries or answer oriented pages that synthesize information from multiple sources. The paper highlights how ontologies natural language processing (NLP) and knowledge representation frameworks enable these systems to interpret the deeper semantics of user queries. Roy et al. identify several ongoing research challenges including the automatic translation of natural language queries into ontology based representations and the dynamic construction and updating of ontologies. While the review offers valuable foundational insights it predates the widespread use of neural and transformer based approaches limiting its applicability to modern AI-driven semantic retrieval systems.

Paper 6 – Semantic Technology Trends in 2024 (Foote, 2023)

Foote (2023) explores the emerging trends shaping semantic technologies and their growing intersection with artificial intelligence (AI). The article argues that the widespread adoption of large language models (LLMs) such as GPT-4 and BARD has accelerated global interest in semantic search and context aware AI systems. Foote highlights how these models when integrated with knowledge graphs can enhance machine understanding by linking structured and unstructured data thereby enabling more intelligent and adaptive responses. The study also discusses the evolution of Web 3.0 emphasizing its aim to merge semantic web principles with decentralization and user control.

According to Foote these converging technologies will redefine how digital systems process meaning leading to smarter virtual assistants improved information retrieval and more personalized user experiences. The article predicts that ChatGPT based assistants will surpass traditional systems in customer service and daily task automation. It foresees generative AI becoming ubiquitous across industries driving innovation in communication knowledge management and automation. Foote's work is primarily conceptual and predictive lacking empirical validation or case studies. Despite this limitation it provides a valuable forward looking perspective on how semantic technologies and AI are shaping the future of intelligent information systems.

Critical Analysis

The selected studies collectively illustrate the evolution of semantic search and its applications, yet they reveal several gaps relative to our problem of matching purchase-order descriptions with database items. Allahim et al. provide a comprehensive taxonomy of query expansion techniques highlighting linguistic and ontological approaches but offering no empirical evidence or domain specific guidance.

Khan et al. demonstrate a secure semantic search system that integrates deep learning with block chain. The evaluation uses a small homogeneous dataset and does not explore cross language or product matching scenarios. Mahboub et al. design an Arabic dataset and show that embedding models such as MPNet produce high search quality but the synthetic nature of the data and narrow domain constrain the findings. Yadav & Gopinathan present a retrieval augmented pipeline using ColBERT while user satisfaction is high the lack of dataset details limits reproducibility and generalizability. Roy et al. survey semantic search engines and identify research challenges but do not address newer neural methods or domain specific applications like product matching. Foote offers an industry perspective on emergent technologies such as large language models and knowledge graphs although informative the article lacks rigorous evaluation and may not translate directly to the purchase-order matching domain.

Connections between the Findings

Across the papers several common themes emerge. all acknowledge that semantic search improves retrieval by considering context and intent rather than mere keyword matching. Allahim et al. and Roy et al. emphasize the importance of linguistic resources ontologies and hybrid approaches while Khan et al. Mahboub et al. and Yadav & Gopinathan leverage embedding models and neural architectures to capture semantic similarity. Second, the integration of semantic search with other technologies encryption and blockchain (Khan et al.) generative AI and virtual assistants (Foote) or retrieval augmented generation (Mahboub et al., Yadav & Gopinathan) demonstrates the multi disciplinary nature of the field. Third, the challenges identified by Allahim et al. (e.g., lack of multilingual resources handling real-time data) are echoed by Mahboub et al. and Yadav & Gopinathan who confront limitations in data size synthetic queries and evaluation metrics. The papers indicate that accurate semantic matching requires robust embedding efficient retrieval and often domain specific knowledge.

Conflicts between the Findings

Despite these connections, several tensions arise. The survey by Roy et al. concludes that semantic search engines provide aggregated summaries, while retrieval-augmented systems like those of Yadav & Gopinathan and Mahboub et al. aim to combine retrieval with generation, potentially shifting focus away from summaries toward novel content generation. Khan block chain-based secure search emphasizes privacy and multi user control, whereas Foote’s trends article assumes wide adoption of generative models, raising concerns about content veracity and computational cost. The effectiveness of multilingual models in Mahboub et al. contrasts with the largely English-focused datasets of Yadav & Gopinathan. Finally, the papers differ in evaluating performance: some rely on user satisfaction (Yadav & Gopinathan), others on precision and recall (Khan et al., Mahboub et al.), making direct comparison difficult.

Synthesis Matrix for Literature Review

Study	Authors	Year	Research Design	Key Findings / Results	Methodology	Limitations
Semantic approaches for query expansion	Allahim et al.	2025	Systematic survey	Categorizes methods into linguistic, ontology- based and hybrid approaches; highlights challenges (multilingual resources, real- time data, multimedia search)	Surveys 34 frameworks published between 2015– 2024; organizes them into a taxonomy	No empirical evaluation; limited guidance for domain- specific applications
Secure semantic search using deep learning	Khan et al.	2024	Experimental prototype	Demonstrates secure semantic search with decentralized key management; achieves ~88 % precision on 20 Newsgroups	Integrates doc2vec embeddings with ABE and blockchain; simulation- based evaluation	Small dataset and simulation; high computational overhead; not tested in real- world or product contexts
Evaluation	Mahboub et al.	2024	Empirical	MPNet	Constructs	Synthetic data

Study	Authors	Year	Research Design	Key Findings / Results	Methodology	Limitations
of semantic search for Arabic RAG			evaluation	embeddings perform best; RAG improves answer quality, but retrieval and generation performance may differ	synthetic dataset (2030 call summaries, 406 queries); evaluates five multilingual encoders; uses GPT- 3.5/4 for RAG	may not reflect real queries; limited domain and size; high computational cost for embeddings
Semantic search and RAG using ColBERT	Yadav & Gopinathan	2025	User study and prototype	Presents a RAG pipeline with ColBERT; users report high satisfaction (4.6–4.8/5) and relevance scores (0.65–0.76)	Embeds metadata with HuggingFace; stores vectors in ChromaDB; uses ColBERT for post-ranking	Full text unavailable; dataset size and diversity unspecified; relies on subjective user evaluations
An overview of semantic search engines	Roy et al.	2019	Descriptive survey	Provides taxonomy of semantic search engines; discusses features, results types and multilingual support	Compares multiple engines (Hakia, Kngine, Kosmix, DuckDuckGo, Powerset, Sensebot)	No empirical evaluation; outdated relative to modern neural methods; not tailored to product matching
Semantic technology trends in 2024	Foote	2023	Opinion article	Identifies trends (LLMs, knowledge graphs, Web 3.0, generative AI) influencing semantic search	Summarizes industry developments and predictions	Not peer- reviewed; lacks empirical evidence; focuses on high-level trends rather than domain-specific applications

References

1. Allahim, Cherif & Imine (2025). Semantic approaches for query expansion: taxonomy, challenges, and future research directions. [Semantic approaches for query expansion: taxonomy, challenges, and future research directions \[PeerJ\]](#)
2. Khan, Abbas & Binsawad (2024). Secure semantic search using deep learning in a blockchain-assisted multi-user setting. [Secure semantic search using deep learning in a blockchain-assisted multi-user setting | Journal of Cloud Computing | Full Text](#)
3. Mahboub, Za'ter, Al-Rfooh, Estaitia, Jaljuli & Hakouz (2024). Evaluation of semantic search and its role in retrieval-augmented generation for Arabic. [\[2403.18350\] Evaluation of Semantic Search and its Role in Retrieved-Augmented-Generation \(RAG\) for Arabic Language](#)
4. Yadav & Gopinathan (2025). Semantic search and retrieval-augmented generation for academic literature analysis using ColBERT. [Semantic Search and Retrieval-Augmented Generation for Academic Literature Analysis Using ColBERT | Request PDF](#)
5. Roy, Modak, Barik & Goon (2019). An overview of semantic search engines. https://www.ijrrjournal.com/IJRR_Vol.6_Issue.10_Oct2019/IJRR0012.pdf[IJRR0012.pdf](#)
6. Foote (2023). Semantic technology trends in 2024. [Semantic Technology Trends in 2024 - Dataversity](#)