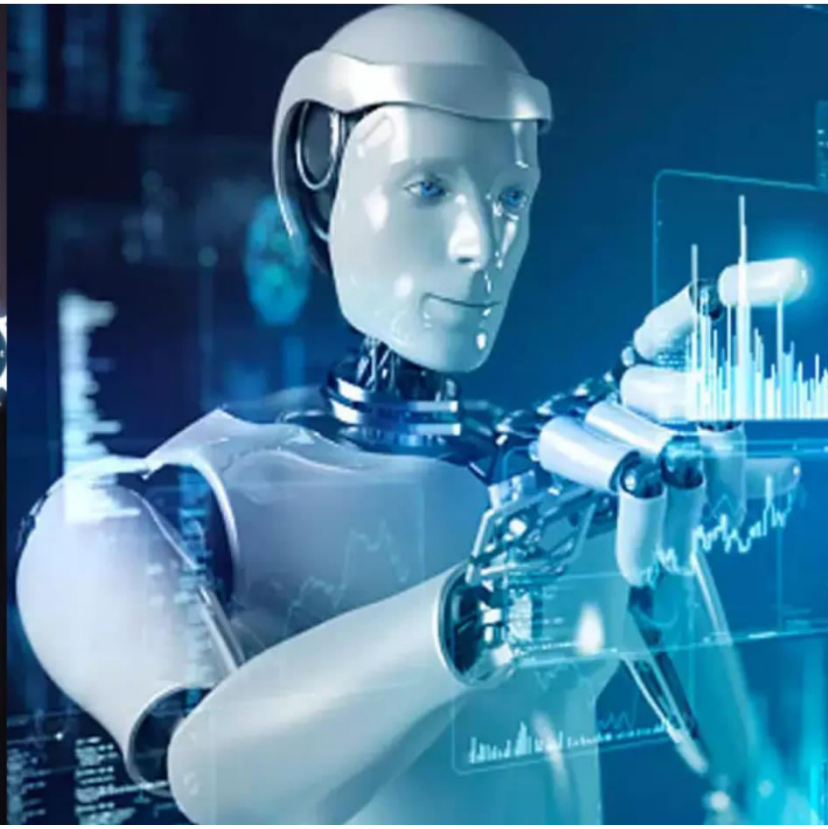
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404635|

Building Hybrid Search Engine: Combining Keyword and Semantic Search

Mr. Prabhakar Koturwar, Mr. Avinash Bhalerao, Mr. Pritam Patil, Ms. Anjali Patil, Mr. Anas Pathan, Prof. Pankaj Shinde

U.G. Student, Department of Computer Engineering, DY Patil University, Ambi, Pune Maharashtra India

U.G. Student, Department of Computer Engineering, DY Patil University, Ambi, Pune Maharashtra India

U.G. Student, Department of Computer Engineering, DY Patil University, Ambi, Pune Maharashtra India

U.G. Student, Department of Computer Engineering, DY Patil University, Ambi, Pune Maharashtra India

U.G. Student, Department of Computer Engineering, DY Patil University, Ambi, Pune Maharashtra India

Associate Professor, Department of Computer Engineering, DY Patil University, Ambi, Pune Maharashtra India

ABSTRACT: In the contemporary digital landscape, efficiently retrieving pertinent information from extensive data repositories is vital. Traditional keyword search methods often fall short by failing to grasp the context of user queries, resulting in irrelevant outcomes. This study delves into the creation of a hybrid search engine that amalgamates keyword and semantic search methodologies utilizing MySQL, ChromaDB, Python, Streamlit, and Sentence Transformers. The objective is to develop a hybrid search system capable of processing user inquiries through both keyword matching in a MySQL database and semantic comprehension using embeddings stored in ChromaDB. [1] By leveraging the Sentence Transformers library to generate embeddings for product descriptions, this system enhances search capabilities beyond mere keyword matching.

KEYWORDS: Semantic Search, Hybrid Search Engine, Keyword Matching, Search Optimization, MySQL, ChromaDB, Python, Streamlit, Sentence Transformers

I. INTRODUCTION

In today's digital landscape, the ability to efficiently retrieve pertinent information from extensive datasets is not just a convenience but a necessity. Traditional keyword search methods, while foundational, often stumble in grasping the nuanced context of user queries, leading to less relevant or even misleading search results. Recognizing this limitation, the development of hybrid search engines[5]

that integrate both keyword and semantic search techniques marks a significant leap forward. This project embarks on creating such a hybrid search engine, seamlessly blending keyword and semantic search methodologies using a robust technological stack that includes MySQL, ChromaDB, Python, Streamlit, and Sentence Transformers. The primary aim is to build a search system that enhances the user experience by accurately interpreting and processing queries.At the heart of this system lies a twofold approach: leveraging traditional keyword matching within a MySQL database for straightforward query handling, and employing semantic search capabilities through embeddings stored in ChromaDB for deeper contextual understanding. The power of the Sentence Transformers

1.Hybrid Search Approach:

Traditional search engines only look for exact matches of keywords in a database. If the user doesn't type the exact product name or words, they might not find what they're looking for.[3]

•Example: If a user searches for "cheap smartphones," the system might miss products with descriptions like "affordable phones."

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404635|

- **Context Understanding:**

Traditional searches struggle to understand the meaning behind the user's query. They can't recognize synonyms or related concepts.

Example: Searching for "running shoes" may not show results for "sports sneakers" even though they mean the samething.

- **Inefficiency in Handling Descriptions:**

Keyword searches don't handle long product descriptions well. They miss out on context, leading to irrelevant or incomplete results.

Example: Searching for "comfortable office chair" might not bring up results with long descriptions that talk about ergonomic design and comfort.

Hybrid Search Solves:
1. Combines Keyword and Semantic Search:
•The hybrid system allows users to search using both traditional keyword matching (MySQL) and a more advanced semantic understanding (ChromaDB). This means users can find relevant products even if they don't know the exact words used in the product description.[2]
•Solution: Improves accuracy and relevance by considering both exact matches and the meaning of the words.

2.Understands the Meaning of Queries:
•By using sentence embeddings (semantic search), the system understands the context and meaning behind a user's query.
•Solution: Finds products based on the overall description, synonyms, and similar concepts, rather than just keywords.

3.Handles Descriptions More Effectively:
•The system can search through and understand lengthy product descriptions, pulling out relevant results even if certain keywords aren't present.
•Solution: Users get more accurate results for detailed or complex product descriptions.

4. Vector Database
• A vector database is a type of database designed to store and search data based on numerical vectors (numbers arranged in a specific order). In this case, each piece of text, like a product description, is turned into a vector, which is a series of numbers that represent the meaning or context of the text.

•How it works: Instead of just storing text like traditional databases (e.g., MySQL), vector databases store the meaning of the text as vectors.

•Why it's useful: By storing the meaning as vectors, these databases can perform semantic searches, which means they find similar or related items, even if they don't contain the exact same words. For example, if you search for "comfortable chair," it can return results that talk about "ergonomic seating" because they mean similar things.

In our project, ChromaDB is a vector database that stores embeddings (meaningful vectors) of product descriptions. When a user searches by description, the database returns results with similar meanings, not just exact keyword matches.

II. RELATED WORK

1.Paper Name: Hybrid Search: Effectively Combining Keywords and Semantic Searches
Author: Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfranchi and Daniela Petrelli

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404635|

Description : This paper focuses on leveraging the benefits of semantic technologies to enhance information retrieval and knowledge management. The authors investigate how semantic web technologies can facilitate more accurate and context-aware search experiences. The study delves into methods for improving query formulation and information retrieval in specialized domains like academic research. The goal is to refine search mechanisms to improve efficiency and relevance in retrieval tasks, making information more accessible to users.

2. **Paper Name:** Sem Search - Refining Semantic Search
Author: Victoria Uren, Yuangui Lei, and Enrico Motta

Description: This paper introduces SemSearch, a search engine developed for the semantic web, and presents its architecture and functionalities. The authors aim to improve the user's search experience by combining semantic web data with advanced search techniques to deliver more accurate results.

3.**Paper Name:** How Useful are Natural Language Interfaces to the Semantic Web for Casual End-users?
Author: Esther Kaufmann and Abraham Bernstein

Description: Kaufmann and Bernstein examine the effectiveness of natural language interfaces (NLI) in enabling casual users to interact with the semantic web. The study evaluates the practical application of NLI tools in enhancing user experience and access to web-based data. The authors analyze the usability for non-technical users, identifying limitations such as understanding complex queries and the lack of seamless integration with the user's natural way of speaking. Their research provides insights into the gap between technical potential and real-world usability.

4.**Paper Name:** Contexts for the Semantic Web
Author: R.Guha, R.McCool and R.Fikes

Description: Guha, McCool, and Miller's paper explores semantic search technologies, focusing on how they differ from traditional search engines by using the relationships between concepts and entities. The study investigates the process of understanding user intent more accurately and retrieving data that is semantically connected rather than keyword-based. The authors discuss

5.Paper Name: The Usability of Semantic Search Tools
Author: Uren et al.

Description: This review discusses the usability of semantic search tools, highlighting the challenges of adapting semantic web technologies to meet the needs of end-users. The authors evaluate existing tools for their ease of use, identifying barriers such as complex interfaces and the steep learning curve associated with these systems. They emphasize the need for more intuitive tools that cater to a broader audience, potentially improving the adoption rate of semantic search engines by non-experts.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

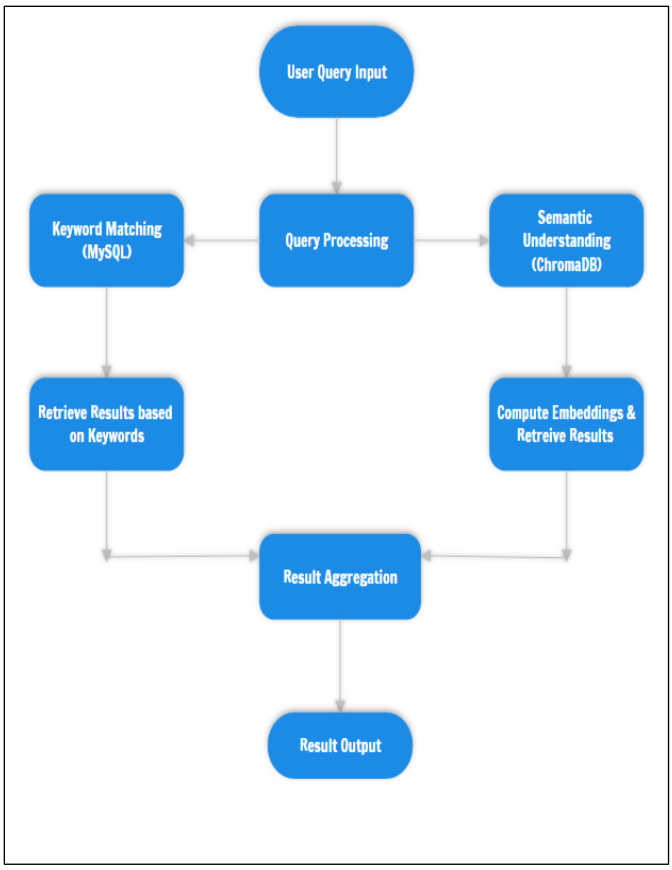
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404635|

III. SYSTEM ARCHITECTURE



The flowchart represents a hybrid search approach that combines keyword-based and semantic search to process user queries effectively. The flowchart outlines a hybrid search system that efficiently processes user queries by combining keyword-based and semantic search techniques. The process begins when a user submits a query, which is then analyzed in the Query Processing stage. At this stage, the system determines whether to retrieve results based on keyword matching, semantic understanding, or a combination of both. [2] For keyword-based searches, the system utilizes MySQL to find exact matches for the query terms. It scans the database and retrieves relevant results based on keyword occurrence. Simultaneously, for more contextual and meaning-based searches, the system employs ChromaDB, a vector database that uses embeddings to understand the semantic meaning of the query. The query is converted into numerical vectors, which are then compared with stored embeddings to find semantically relevant results.[4]After retrieving results from both methods, the Result Aggregation stage combines and ranks them based on their relevance. Finally, the Result Output stage presents the most accurate and meaningful results to the user. By leveraging both structured keyword searches and advanced semantic understanding, this system enhances information retrieval, ensuring a more effective and comprehensive search experience.

IV. APPLICATIONS

E-Commerce Website

Helps users find products based on both names and detailed descriptions, even if their search terms don't exactly match product titles.

Example: Users can search for "budget-friendly smartphones" and still find phones described as "affordable."

Knowledge Databases

Useful for companies or organizations that have large amounts of information (like articles, product guides, or manuals) to help users quickly find the most relevant document.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404635|

Example: A user searching for "how to fix a car engine" might find an article titled "engine repair guide."
The project detailed in the article "Building Hybrid Search Engine: Combining Keyword and Semantic Search"

V. RESULT AND DISCUSSION

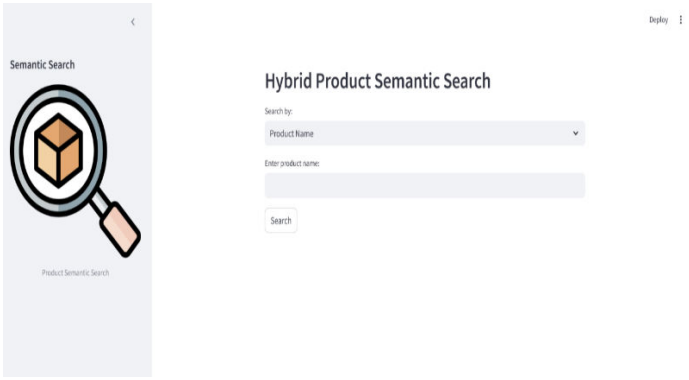


Fig. 1 A hybrid product search interface combining keyword and semantic search

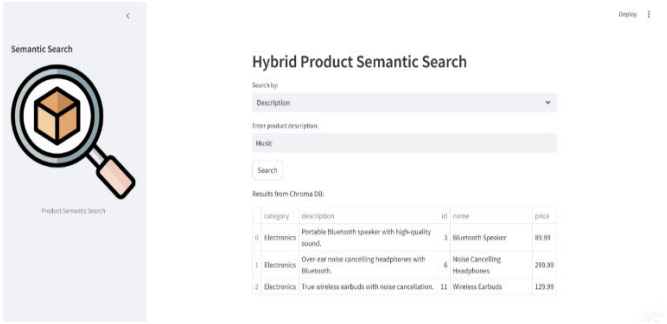


Fig. 2 The system performs a **semantic search** using ChromaDB to retrieve products based on description.

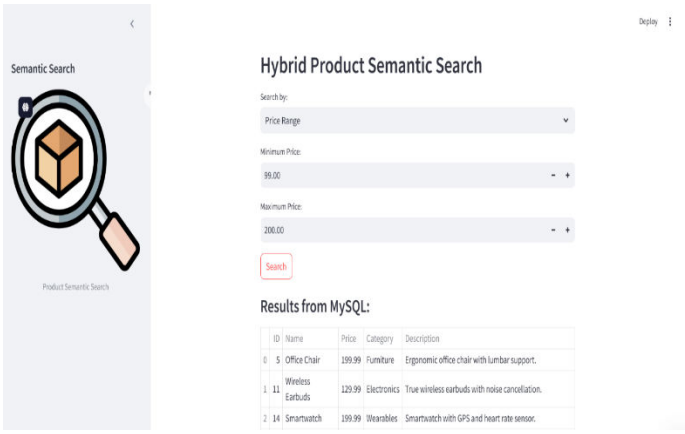


Fig. 3 The system uses **keyword-based filtering** with MySQL to find products within a specified price range

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404635|

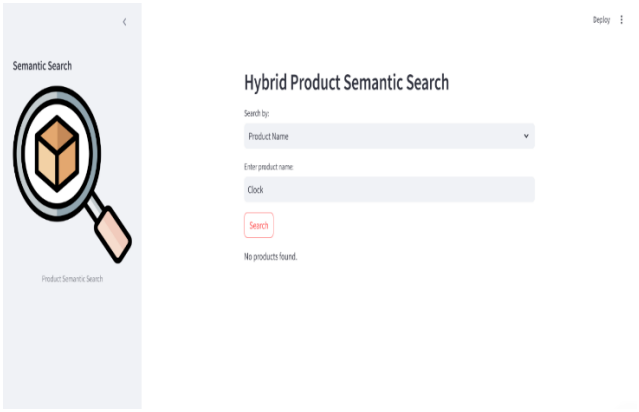


Fig. 4 No products were found for the entered product name "Clock" using keyword-based search.

The Hybrid Product Semantic Search system combines **se**mantic (ChromaDB) and keyword-based (MySQL) search for better product retrieval. While semantic search finds context-based results, keyword search relies on exact matches. The system works well but may return no results if an exact match is missing, emphasizing the need for a well-structured database.

VI. DATA AI REVOLUTION

It focuses on integrating traditional keyword-based search with semantic search capabilities to enhance information retrieval accuracy. This integration leverages technologies such as MySQL, ChromaDB, Python, Streamlit, and Sentence Transformers to address the limitations of conventional keyword searches by incorporating contextual understanding. By exploring these avenues, the hybrid search engine can evolve to offer more accurate, personalized, and secure search experiences, addressing the dynamic needs of users in an ever-changing digital landscape.

Future Scope and Potential Enhancements:
Incorporation of User Behavior Analytics: By analyzing user interaction data, such as click-through rates and dwell time, the search engine can learn and adapt to user preferences, thereby refining the relevance of search results over time. Integration of Multimodal Search Capabilities: Expanding the search engine to process and understand various data types—such as text, images, and videos—would allow users to retrieve information across different media formats, catering to diverse query inputs. Implementation of Personalized Search Experiences: Utilizing user profiles and historical search data can enable the search engine to deliver personalized results, enhancing user satisfaction by aligning outcomes with individual interests and search histories. Enhancement of Real-Time Data Processing: Developing mechanisms to index and retrieve information from rapidly updating data sources, like news feeds and social media, would ensure that users receive the most current information available. Development of Advanced Natural Language Understanding: Incorporating more sophisticated natural language processing techniques can improve the search engine's ability to comprehend complex queries, manage ambiguities, and interpret nuanced language, leading to more accurate results.

Expansion to Multilingual and Cross-Lingual Search: Implementing multilingual support would enable the search engine to process queries and retrieve documents in multiple languages, breaking language barriers and providing a more inclusive user experience. Adoption of Scalable and Distributed Architectures: Transitioning to scalable architectures can enhance the search engine's performance, allowing it to handle larger datasets and higher query volumes efficiently, which is essential for widespread adoption. Integration of Voice Search and Conversational AI: Incorporating voice recognition and conversational AI capabilities would enable users to interact with the search engine through natural language speech, aligning with the growing trend of voice-activated assistants. Implementation of Robust Privacy and Security Measures: Ensuring user data privacy and implementing security protocols is crucial for maintaining user trust, especially when personal data is utilized for personalization and analytics. Evaluation and Enhancement of Search Result Explainability: Developing features that provide users with insights into why certain results were presented can increase transparency and trust in the search engine's operations.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404635|

VII. CONCLUSION

This hybrid search engine project exemplifies a cutting- edge approach to information retrieval, blending the precision of keyword-based search with the contextual depth of semantic search techniques. In the vast and rapidly growing digital landscape, users often struggle to find relevant information due to the limitations of traditional keyword search methods, which can fail to capture the true intent behind user queries. This project addresses that gap by integrating MySQL for structured data management and ChromaDB for advanced semantic understanding, ultimately enhancing the user experience. In summary, this hybrid search engine project showcases an advanced and user-centric approach to information retrieval. By combining the strengths of keyword and semantic search techniques, it offers a more accurate, relevant, and satisfying search experience. The integration of MySQL and ChromaDB, supported by the powerful Sentence Transformers library, paves the way for future innovations in search technology, providing a robust framework for addressing the challenges of modern digital data retrieval. This project not only improves user satisfaction but also sets a new standard for search engine performance in the digital age

REFERENCES

[1] Ravish Bhagdev , Sam Chapman , Fabio Ciravegna , Vitaveska Lanfranchi and Daniela Petrelli Department of Information Studies University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP Sheffield, United Kingdom
[2] Uren, V., Lei, Y., Lopez, V., Liu, H., Motta, E.and Giordanino, M.: The usability of semantic search tools: a review, Knowledge Engineering Review, in press.
[3] Kaufmann, E. and Bernstein, A.: How Useful are Natural Language Interfaces to the Semantic Web for Casual End-users? Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, Busan, Korea, November 2007
[4] Lei, Y., Uren, V. and Motta, E. SemSearch: A Search Engine for the Semantic Web. in 15th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks (EKAW 2006). 2006. Podebrady
[5]Guha, R., McCool, R. Miller, E. Semantic Search. in 12th International Conference on World Wide Web. 2003
[6] Chen, H., et al. (2021). Hybrid Search: Integrating Retrieval, Extraction, and Generation for Effective Information Access. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing
[7] Liu, Y., et al. (2020). Knowledge Graph Embedding with Iterative Guidance from Soft Rules. In Proceedings of the 34th AAAI Conference on Artificial Intelligence

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details