AI-Assisted Mathematical Modeling of Patents and Development of Automated Patent Search System

Emre Doruk Taşkın   Baran Kılıç   Ferayenur Bozkurt   İrfan Ulaş Geçin

Supervisor: Prof. Dr. Pınar Duygulu Şahin   **Industry Advisor**: Özcan Somuncu

Motivation

- **Time & Effort Reduction** Manual patent searches are time-consuming and labor-intensive—this system automates the process, saving engineers valuable time.
- **AI-Powered Search** Automating the search process using NLP and LLMs to improve accuracy.
- **Enhanced Query Understanding with LLMs** Users can enter natural-language queries, which are enriched through LLM-based query expansion, paraphrasing, and normalization to improve search precision.
- **Internal and Secure System** Unlike external tools like PatSeer or IPRally, this platform is fully internal to TUSAŞ, protecting sensitive and proprietary information. Additionally, LLMs are also ran locally using highly-optimized **Llama.cpp** instead of using external APIs.
- **Semantic Alignment Between Query and Patent Data** By using embeddings and vector databases, the system retrieves patents based on semantic similarity, not just keyword overlap — aligning much more closely with the user's true intent.

System Overview
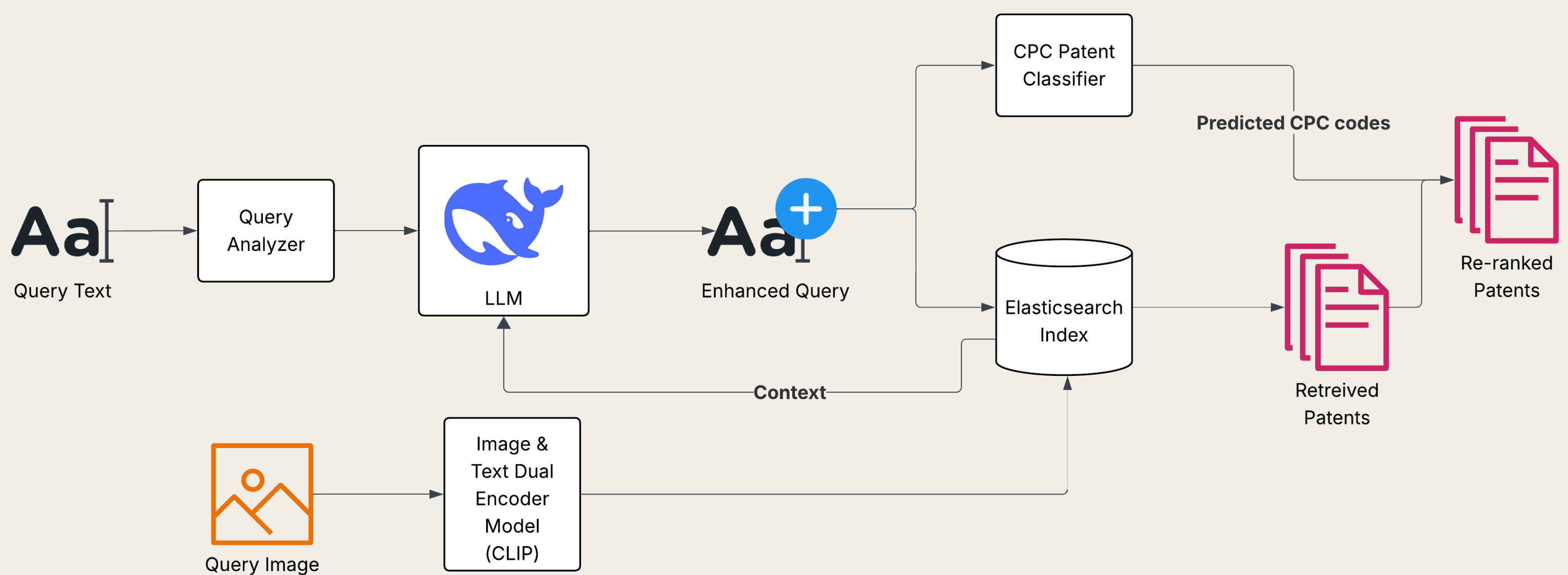
Methodology

- **Patents to Text Embeddings:** We generate sentence embeddings for patent texts using the **PatentSBERTa** model and index them in Elasticsearch to enable efficient vector-based similarity search.
- **Analyze User Query:** We analyze the user's search query to determine whether it is a conceptual query or a detailed query, where the user is looking for a specific patent. Based on this, we dynamically adjust search parameters, including weights of different patent sections in the search and the prompt template used for query expansion in the next step.
- **Query Expansion:** User queries often lack the necessary context and keywords that would be the most useful to find the patents the user is looking for effectively. To overcome this, we use quantized versions of state-of-the-art LLMs such as **DeepSeek-R1-8B** to expand the user query and increase its usefulness in the patent search. Prompting techniques like **Pseudo-Relevance Feedback (PRF)** are utilized to help constrain the LLM to only generate content aligned with our database, avoiding irrelevant additions to the query.
- **Image Search:** When the user provides an image alongside the text query, we apply a pre-trained **CLIP** model to map both the image and the query text into a shared embedding space. This enables **multimodal search** without needing to store image data for the patents.
- **Hybrid Patent Retrieval:** We embed the user query (expanded and optionally combined with image input) and retrieve relevant patents using a **hybrid approach**: K-Nearest Neighbor (KNN) search over vector embeddings for semantic similarity, and **BM25** to capture important keyword matches.
- **CPC-Based Re-ranking:** In the final step, we utilize the **Cooperative Patent Classification (CPC)** codes of patents to predict the potential CPC codes of the user query. This allows us to predict potential domains and subdomains of the topic in the user query and boost the patents with high overlap in the final ranking.

Web Application