Optimizing Information Retrieval for Healthcare Literature with Semantic Embeddings and Vector-based Indexing

1st Dávid Miklo
Dept. of Computer Science
Universitas Gadjah Mada
Yogyakarta, Indonesia

2nd Jari Roossien
Dept. of Computer Science
Universitas Gadjah Mada
Yogyakarta, Indonesia

***Abstract*—The exponential growth of scientific literature presents significant challenges in efficiently retrieving relevant information, particularly in healthcare, where timely access to research is critical. Traditional information retrieval (IR) systems, often relying on keyword matching, fall short in addressing the semantic complexities of textual data, leading to suboptimal outcomes for ambiguous or short queries. This paper proposes a vector-driven approach to IR using semantic embeddings to represent documents and queries as high-dimensional vectors, enabling precise similarity matching based on contextual meaning.**

We integrate advanced Sentence Transformer models for vectorization and FAISS (Facebook AI Similarity Search) for scalable indexing and similarity search. The proposed system is evaluated against two baselines—Boolean retrieval and TF-IDF models using a dataset of 200,000 healthcare-related research abstracts. Experimental results are assessed on precision and recall using a ground truth set of relevant abstracts. Surprisingly, the findings reveal that the experimental approach, while leveraging advanced techniques, fails to outperform the simpler baseline models, particularly in terms of recall. This highlights the need for further optimization and refinement of vector-driven retrieval systems in the context of healthcare literature.

***Index Terms*—Information Retrieval, Health care, Vector-space Model, FAISS, TF-IDF**

I. Introduction

The exponential growth in published research papers and scientific literature has created immense opportunities and significant challenges. The sheer volume of new data has made it increasingly difficult for researchers and healthcare professionals to access and utilize relevant information efficiently [src1]. This problem is particularly critical in healthcare, where staying updated on the latest findings is crucial for patient care and research advancement.

Traditional Information Retrieval systems, which often rely on keyword matching or statistical approaches, struggle to meet the demands of this growing dataset. These methods frequently fail to account for the semantic relationships within textual data, resulting in less effective retrieval outcomes, especially when queries are ambiguous or short.

To address these limitations, this paper explores a vector-driven approach to information retrieval for healthcare literature. By representing documents and queries as high-dimensional vectors, we leverage advanced semantic embedding techniques to capture the contextual meaning of textual data. These embeddings enable precise matching between queries and relevant literature based on their semantic similarity.

To manage and search through large-scale vector representations efficiently, we utilize FAISS (Facebook AI Similarity Search), a state-of-the-art library for fast vector indexing and similarity search. This combination of semantic embeddings and FAISS indexing provides a scalable and accurate solution for retrieving relevant healthcare information from vast datasets.

II. Literature Review

Research indicates that user queries, particularly in healthcare, are often short and imprecise, leading to suboptimal results in traditional IR systems.

Recent advancements in vector-based representations have transformed the field of information retrieval. Semantic embedding models, such as Sentence Transformers, encode textual data into dense vectors that capture both contextual and semantic meaning. These vector representations allow IR systems to measure similarity based on meaning rather than surface-level keyword overlap, making them especially suited for short or ambiguous queries.

Efficient indexing and search are critical for handling the large vector datasets generated by semantic embeddings. FAISS offers a powerful solution by enabling high-speed similarity search across billions of vectors, making it ideal for large-scale healthcare literature datasets. By combining semantic embeddings with FAISS, this paper presents a vector-driven retrieval system that significantly enhances the precision and scalability of healthcare literature search, addressing the shortcomings of traditional IR approaches.

III. Method

A. Data Collection

Our dataset consists of healthcare-related research abstracts (N = 200,000), obtained from a public GitHub repository

curated from PubMed research articles. Each article is divided into structured sections such as Background, Methods, Results, and Conclusions. The dataset is available at: https://github.com/Franck-Dernoncourt/pubmed-rct/tree/master

B. Approach

We evaluate the effectiveness of contextual embeddings with vectors by comparing an experimental model containing both, and 2 baseline models: A boolean retrieval with an inverted index and a TF-IDF model with an inverted index. All these models will be evaluated on precision and recall given a ground truth set of abstract IDs.

C. Data Processing

The following steps were applied to prepare the data for the experimental model:

- **ID-Abstract Association -** Abstracts were grouped by their unique IDs, consolidating all sections into a single document for each article. This ensured retrieval would operate on complete abstracts rather than individual sentences.
- **Semantic Vectorization -** Using a pre-trained Sentence Transformer model (`all-MiniLM-L6-v2`), each abstract was transformed into a high-dimensional vector representation. These embeddings encode the semantic meaning of the abstracts.
- **Index Construction with FAISS -** Vector embeddings were indexed using FAISS, using the IndexFlatIP index for scalability.

Both baseline models utilized the same abstract grouping function as the experimental model. For the boolean baseline model, we tokenized the text using split and regex, excluding special characters and numbers, and turned each token into a fully lowercase version.

The second baseline model utilizes `TfidfVectorizer` from the sklearn library with stop words from the English vocabulary. The text

D. Query Processing

- **The Experimental model** queries were processed similarly to its index creation, transforming them into vector representations using the same Sentence Transformer model.
- **The Boolean model** queries were tokenized using the same function as for the index. The tokens were then matched on logical operators.
- **TF-IDF model** utilizes the inbuilt `.transform()` function of `TfidfVectorizer`.

E. Evaluation

We use **precisions** and **recall** to evaluate the effectiveness of the systems. A ground truth set was established by manually selecting relevant abstracts.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

IV. Results

In this chapter we will discuss the results of the comparison between the different systems. For Boolean Retrieval and TF-IDF, we ran a single system, while for the FAISS system we performed the same query with 2 parameters, a low threshold to allow for more articles, and a higher threshold to check for quality of documents.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT SEARCH TECHNIQUES.

Search Technique	Precision	Recall
Boolean	0.680	1.000
TF-IDF	0.0364	1.000
FAISS (high threshold)	0.0789	0.1765
FAISS (low threshold)	0.0020	0.9412

Based on the result in Table I, the results reveal notable differences between the basic Boolean retrieval system and more advanced queries such as TF-IDF or FAISS. The original boolean retrieval system achieves the highest score in both precision and recall, with respective 0.68 and 1.00. The TF-IDF system receives the lowest scores in precision with a 0.0364, however it is able to maintain a perfect recall with 1.00 as well.

The FAISS retrieval system shows significant different results depending on the level of threshold applied, in both Precision and Recall. The high threshold performs better than TF-IDF for precision, however, the low recall shows that a significant portion of relevant documents falls outside the threshold. When lowering the threshold for the scores, the recall improves significantly, at a cost of heavily reducing the precision.

V. Discussion

A. Outcome insights

The results of the experiments show the increased complexity in developing an information retrieval system using more advanced techniques such as TF-IDF or FAISS. The weakness of a boolean retrieval system, which is the inability to find documents with synonyms and not counting the relevance of a token, seem to not be significant enough in the field of medical scientific papers. This also results in the strength of a vector-space model not being utilized, and falling short to their weakness of mismatching the weights of a document.

B. Challenges and Limitations

The use of vector-based retrieval systems shows significant challenges compared to traditional information retrieval systems in the medical field. A big challenge in vector-based approaches comes in possible mismatching of terms. The low results of our FAISS system shows there is a big disconnect between the vectorization of the expected documents and the results the system provides. With the low threshold, the recall improves significantly , however the precision falls down to where the system seems to grab many health care-related documents.

The results also show the vector approach does not seem to capture the semantics between terms well in this specific domain. This could either be overstating the combinations due to the nature of the domain being specific, or missing connections between keywords and tokens.

With an advanced system like FAISS, we unfortunately lose control of the detail which keywords hold a significant detail, and which keywords are being left out. This means in the case of getting odd results like our system, it is hard to exactly pinpoint what the system determined relevant from our query, and from the documents when calculating the vectors.

C. Opportunities for improvements

The use of advanced information retrieval systems allows for many unique approaches to leverage different strengths.

- **Knowledge Graph** - The use of a knowledge graph could improve the results of an advanced system like FAISS. By providing an extensive ontology within the domain, the accuracy of the retrieval could improve significantly.
- **Thesaurus Integration** - The use of a medical thesaurus could improve the vectorization of the documents, allowing the application to help determine domain-specific synonyms and group them as such.
- **Adaptive Threshold** - By adjusting the threshold, the applications can fine-tune the right trade-off between precision and recall, providing the right amount of relevant results to the user.

D. Further Research

For further research, more complex data could be provided to improve the Precision of these systems. One of such methods is to implement a thesaurus. Providing a thesaurus will allow the application to understand similar keywords and group them together as such, increasing the precision.

Another one of such applications could be the implementation of a knowledge system. By providing the semantics between medical terms, an advanced system such as FAISS is able to leverage ontologies and improve its accuracy.

Additional research and application could be in the use of building an advanced retrieval system with a user-centric approach. A big advantage could be for patients to understand the symptoms, and take action earlier if the problem could be cause of serious health implications. Unlike experts in the medical field, the lack of familiarity with medical terms could be a significant limitation to potential users of boolean retrieval systems.

VI. Conclusion

The use of Information retrieval systems in Healthcare Literature is an increasingly more critical part of research. With the exponential growth of published papers year by year, it's important to develop strong systems with accurate document retrieval, that maintain a high recall rate to provide the most inclusive data.

This paper highlights the challenges associated with building and implementing advanced information retrieval systems. While current approaches struggle to compete with original boolean retrieval systems in the domain of healthcare literature, we have identified potential pathways for improvement.

We believe that it is possible to implement a more precise information retrieval system, and have suggested different approaches, that in a combined hybrid approach could provide a new level of accuracy for literature retrieval. Through the use of knowledge graphs, thesauri, and threshold balancing systems, there's potential to gain significant improvements for information retrieval systems.