POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Integrating Semantic and Keyword Search: A Transformer-Based Approach for Content Discovery

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Sofia Martellozzo

Advisor: Prof. Paolo Cremonesi

Co-advisors: Riccardo Biondi, Federico Sallemi

Academic year: 2023-24

1. Introduction

In the domain of streaming platforms, such as the well-known Netflix, Disney+, and Amazon Prime, offering an extensive catalog of films and series, the demand for sophisticated search mechanisms is evident.

Traditional keyword-based search functionalities, while efficient in matching exact terms, are limited in their ability to fully capture and respond to the nuanced intentions behind user queries. Recognizing this limitation, our research introduces an innovative hybrid search model that harnesses precise keyword detection and sophisticated semantic search techniques. Leveraging Transformer-based models, renowned for their deep understanding of natural language, semantic search interprets synonyms, typos, and structured sentences, thus revolutionizing the search experience within these platforms. Our research is motivated by the understanding that users would benefit from the interaction with search engines as they would in a conversation with another human, using natural language and expecting an intuitive understanding of their needs. Traditional systems require users to precisely articulate their search terms, a limitation that restricts discoverability and may miss content that, while not matching keywords exactly, is perfectly aligned with the user's intent.

For example, a search for "space" in a traditional system might return titles like "Space Jam" or "Office Space", based solely on keyword matches. In contrast, a semantic search could suggest films such as "Interstellar" or "Gravity". Consider the refined example: "I want to watch a xmas movie with Bruce Willis". In this scenario, the hybrid model's semantic search capabilities interpret "xmas" as "Christmas", while the keyword aspect precisely targets "Bruce Willis", thereby refining the search results to high relevance.

This thesis work was developed during an internship program at ContentWise, a software company specializing in the development of advanced recommender systems, with the purpose of integrating this novel service into its framework. The ambition is to redefine the user experience, making content discovery on streaming platforms as natural and intuitive as conversing with a friend, thereby enriching user satisfaction and engagement through enhanced discoverability and a deeper, more meaningful connection with content.

2. State of the Art

In Natural Language Processing (NLP) and Information Retrieval (IR), the development of word embeddings, and transformer-based architectures, have contributed to a revolution in the understanding and processing of human language by machines. Embeddings are numerical representations of text that capture semantic relationships and features in a multi-dimensional space. The advent of Transformer architecture and its attention mechanism[4] has enhanced the capability of models to produce embeddings of words or entire sentences. Examples of this advancements include BERT[1] and its variants, such as sBERT[3] and RoBERTa[2]. These powerful models can improve their performance in specific domains through fine-tuning techniques. Semantic search is the process of retrieving information that is most similar to a given query by finding the closest matches in a multidimensional vector space, based on the embeddings of the query. Hybrid search combines traditional term-based search techniques with the power of semantic search, which captures the contextual meaning of a query. It achieves this integration through a fusion algorithm that combines the scores given by keyword search and semantic search into a unified ranked scoring list.

3. Dataset

This study leverages three distinct industrial datasets containing metadata associated with streaming content.
The dataset 1 originates from Contentwise and is constructed by leveraging the TMDB open-source database, entirely in English. Dataset 2, originating from a Northern European company, provides visual media content with multilingual support in five distinct versions, reflecting the regional diversity of its audience. Sourced from a global telecommunications firm, dataset 3 encompasses a comprehensive collection of entries originating from Latin American countries. Due to the substantial size of this dataset, we limited our analysis to the subset distributed in Brazil, encompassing approximately 20% of the original data. The three metadata datasets differ also in information quantity, with dataset 1 being the most comprehensive.
Afterward, we create a synthetic query-answer (QA) dataset to mitigate the absence of suitable open-source data, employing `gpt-3.5-turbo` for generating responses to a list of queries, generated with a template. The subsequent filtering and dataset splitting ensure data integrity and facilitate model training and evaluation, resulting in a total of approximately 40000 query-answer (with a proportion of 60-10-30 for training, validation, and test). Figure 1, illustrate the pipeline employed to generate distinct versions of synthetic datasets.

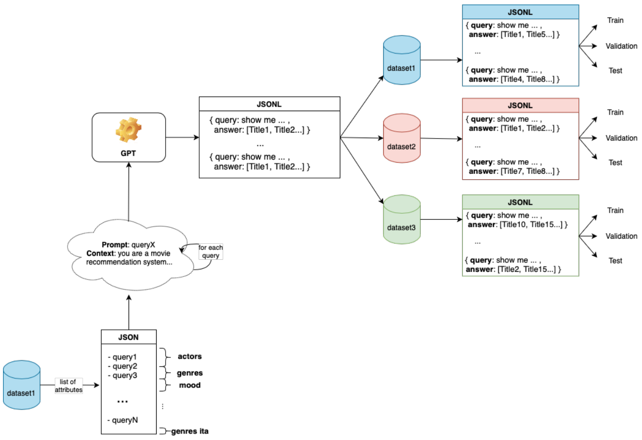
JSONL
{ query: show me ... , answer: [Title1, Title5...] }
...
{ query: show me ... , answer: [Title4, Title8...] }
Train
Validation
Test
dataset1
JSONL
{ query: show me ... , answer: [Title1, Title2...] }
...
{ query: show me ... , answer: [Title1, Title2...] }
GPT
dataset2
JSONL
{ query: show me ... , answer: [Title1, Title2...] }
...
{ query: show me ... , answer: [Title7, Title8...] }
Train
Validation
Test
dataset3
JSONL
{ query: show me ... , answer: [Title10, Title15...] }
...
{ query: show me ... , answer: [Title2, Title15...] }
Train
Validation
Test
Prompt: queryX
Context: you are a movie recommendation system...
for each query
JSON
- query1
- query2
- query3
...
- queryN
actors
genres
mood
genres ita
dataset1
list of attributes

Figure 1: Workflow diagram for synthetic Question-Answer dataset generation

Lastly, a form is spread to retrieve possible queries for manual validation, enhancing the model's real-world applicability.

4. Methodology

This thesis explores the selection, implementation, and fine-tuning of embedding models for enhancing semantic search and information retrieval systems. Our selection criterion relies on open-source models for their customizability and cost-efficiency, notably from the Hugging Face repository and including AWS's `Titan` for benchmarking.
Model selection is guided by performance on the QA synthetic test sets and considerations like model size and speed, aiming for real-time application suitability. The methodology involves a bifurcated testing approach to assess model performance across different data granularities, exploring cross-lingual robustness with our datasets.
The construction and application of similarity matrices provide insights into the embeddings' effectiveness in capturing item nuances,

derived by a subset of selected embedding models. Within a demo environment that simulates a streaming application, each item's visualization page is enriched with multiple carousels, showcasing the ten most similar items as determined by the similarity matrix.

We adopt different fine-tuning strategies (Figure 2) including adapter mechanisms, traditional full-model fine-tuning, and Low-Rank Adaptation (LoRA), with Multiple Negative Ranking Loss (MNRL) and Cosine Similarity Loss (CSL) for efficient parameter adjustment. These methods aim to refine models' semantic search capabilities, with a particular focus on movie and series retrieval based on user queries.

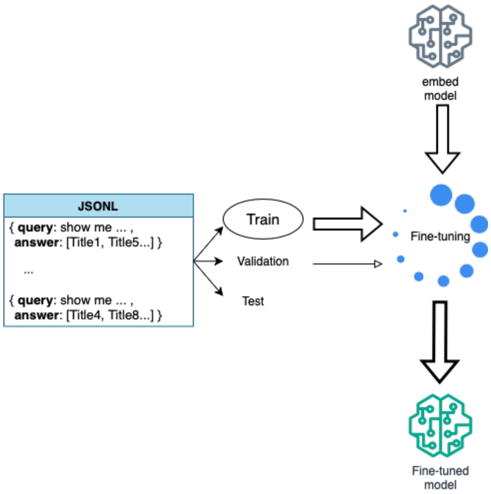
embed model
JSONL
{ query: show me ... , answer: [Title1, Title5...] }
...
{ query: show me ... , answer: [Title4, Title8...] }
Train
Validation
Test
Fine-tuning
Fine-tuned model

Figure 2: Embedding model fine-tuning

5. Implementation

Our customized embedding models serve as a backbone for constructing a search system with a vector database for efficient query and retrieval within a video streaming context. This system architecture involves integrating dataset metadata with their embeddings, processing vectorized queries, and aggregating results from semantic and keyword searches.

First, the system requires uploading all relevant metadata along with our embedding model to ensure that the raw metadata and corresponding embeddings are accurately stored in the vector database, as illustrated in Figure 3. This initial step allows for the strategic organization and indexing of data to facilitate search operations.

The indexing strategy can be customized to enable queries to be performed either in linear time, for precise matching, or through the use of approximate search algorithms, Hierarchical Navigable Small World graphs (HNSW), which are designed to execute query operations in logarithmic time.

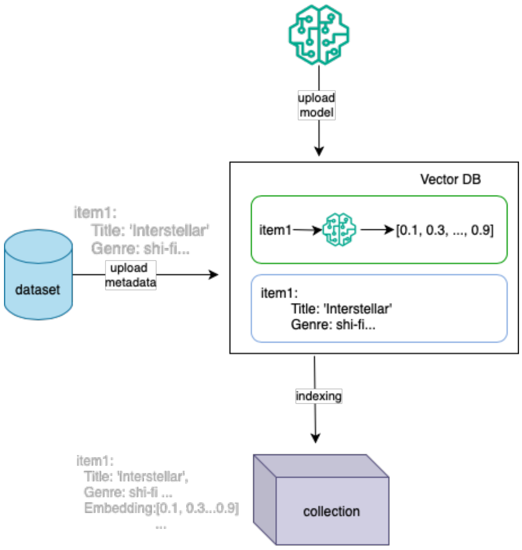

Figure 3: Metadata upload and indexing

Following the indexing phase, the system is ready for users to perform queries through API calls. Our integrated model processes the user query, embedding it to enable the retrieval of the K items most similar to the query. The K scores from both semantic and keyword searches are subsequently merged using a fusion algorithm, producing a single, ranked list that is presented to the user. During the fusion process of the two ranked lists, it is feasible to differentially weight their respective scores. By employing a scale that ranges from 0 (indicating a purely keyword-based approach) to 1 (signifying a fully semantic-based approach), the scores of the corresponding results are multiplied accordingly. It allows for flexible control over the balance between keyword precision and semantic expressiveness in the search results. Figure 4 presents a conceptual visualization aimed at elucidating the mechanics of our hybris search system.

Through automatic and experimental evaluations, we explore different configurations to ascertain optimal setups for hybrid search tasks, demonstrating the system's flexibility and adaptability to varied user needs.

We adopt Weaviate, an AI-native, open-source database, for its comprehensive support of hybrid search methods.

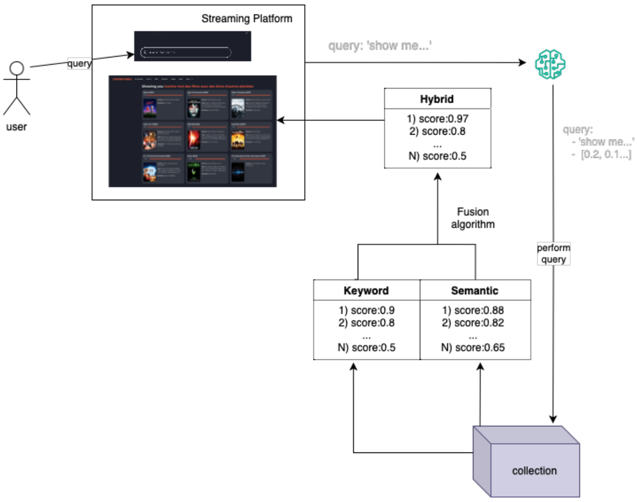

Figure 4: Schema of our information retrieval system

6. Results

The culmination of our research efforts has led to the development of a state-of-the-art solution for Information Retrieval (IR), demonstrating remarkable ability in multilingual semantic search within an hybrid system. This innovative solution has the potential to transform user experiences across various platforms by providing highly accurate and contextually relevant search results (Figigure 5).

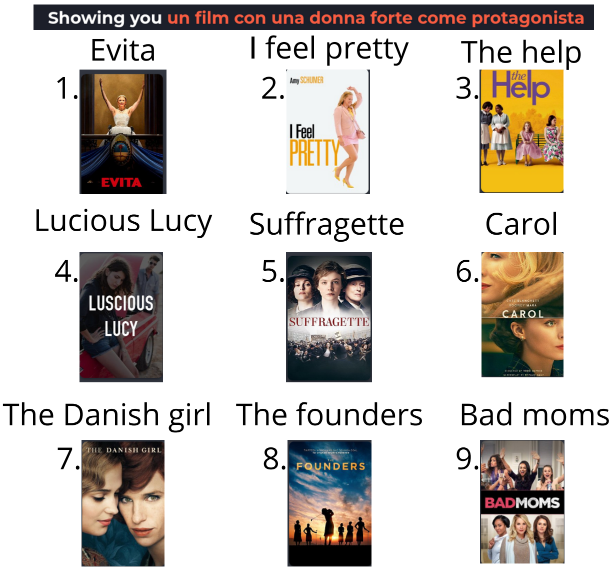

Figure 5: 'Movies with a strong female as main character'

The foundation of this success lies in the strategic selection and optimization of the `multilingual-e5-base` model, fine-tuned using the traditional fine-tuning technique, updating the entire model with MNRL loss. The key accomplishments of our research are highlighted by this model's capacity to analyze and comprehend queries in many languages and its integration into a sophisticated search framework. The decision to adopt this model was validated through extensive evaluations of semantic and hybrid search capabilities, that demonstrated the model's superior performance in real-world search scenarios.

Due to the unavailability of a live demonstration of our system and subsequent online metric collection, we primarily leverage offline metrics for performance evaluation. We employ classic order-unaware metrics, such as Precision, Recall, and F1 Score, to quantify basic retrieval effectiveness. Additionally, given the nature of our task, we incorporate order-aware metrics, namely Mean Reciprocal Rank (MRR), Mean Average Precision at a cutoff of 10 (MAP@10), and Normalized Discounted Cumulative Gain (NDCG).

FT	dataset	Precision	Recall	F1-score
None	1	0.0168	0.0439	0.0243
LoRA	1	0.0227	0.0594	0.0328
MNRL	1	0.0969	0.2535	0.1402
None	2 (Da)	0.0033	0.0092	0.0048
LoRA	2 (Da)	0.0054	0.0153	0.0080
MNRL	2 (Da)	0.0257	0.0724	0.0380
None	3	0.0038	0.0081	0.0052
LoRA	3	0.0040	0.0084	0.0054
MNRL	3	0.0094	0.0199	0.0127

Table 1: `multilingual-e5-base` finetuned performances, order-unaware metrics

FT	dataset	MRR	MAP@10	NDCG
None	1	0.0768	0.0668	0.04200
LoRA	1	0.0977	0.0880	0.05211
MNRL	1	0.3118	0.2740	0.21060
None	2 (Da)	0.0196	0.0178	0.00950
LoRA	2 (Da)	0.0221	0.0198	0.01440
MNRL	2 (Da)	0.1045	0.0927	0.06200
None	3	0.0179	0.0162	0.00860
LoRA	3	0.0178	0.0163	0.00900
MNRL	3	0.0307	0.0271	0.01850

Table 2: `multilingual-e5-base` finetuned performances, order-aware metrics

Our final selection favored traditional fine-tuning. This approach has demonstrated the ability to enhance performance and maintain more generalized outcomes, improving results by three to four times in automatic tests, as confirmed by manual evaluations. To exclude the possibility of overfitting, we validated the model across diverse datasets. While Low-Rank Adaptation (LoRA) showed only a modest improvement in automatic evaluations, it yielded promising outcomes in manual tests. Conversely, the Adapter technique did not significantly enhance the model's performance.

The selection of the `multilingual-e5-base` model was the result of extensive testing across multiple open-source models under varied scenarios, during which it consistently surpassed its competitors. This decision was not solely based on performance metrics. Our analysis revealed a positive correlation coefficient between the embedding size and evaluation metrics, indicating that larger models generally yield better performance. However, for practical applications, we must also consider model latency, which is critically influenced by the size of the model. Therefore, in selecting the optimal model, we prioritized a balance between performance and operational efficiency. This consideration led us to choose the base version of the multilingual E5 model over its larger counterparts. Additionally, the multilingual capabilities of the E5 model demonstrate strong performance across languages compared to other models that may have exhibited greater precision in IR tasks but were limited to English. This multilingual proficiency ensures broader applicability and inclusivity in global contexts.

The tests ranged from altering the number of retrieved elements (either 10 or 20) to limiting the metadata available. Generally, a reduction in performance was observed across all models when faced with a smaller K value and diminished input information.

7. Conclusions

This thesis introduces a novel hybrid search system for enhancing content discoverability on streaming platforms, leveraging Natural Language Processing (NLP) and Information Retrieval (IR) technologies. It integrates semantic search with traditional keyword-based methods to improve user interaction with digital content catalogs.

A custom embedding model tailored for streaming content is developed, incorporating advanced fine-tuning techniques, significantly enhancing the system's performance. The work establishes a comprehensive evaluation framework for embedding models, validating the efficacy of the proposed techniques. Despite facing limitations such as subjectivity in effectiveness, data quality issues, the trade-off between model performance and operational efficiency, and challenges in multilingual support, this research marks a significant step towards more intuitive content discovery mechanisms.

Future directions include integrating reranking models to refine search results and developing query refinement models to address user query inaccuracies, alongside exploring advancements in embedding technologies for improved linguistic comprehension and search performance. Additionally, our work has the potential to be trained and adopted in other domains, such as e-commerce, cosmetics, and pharmacy.

8. Acknowledgements

Special thanks go to Prof. Paolo Cremonesi, and Doc. Riccando Biondi, Federico Sallemi for their support and guidance during this work. In addition, all the staff of Contentwise, my family, and friends for their direct or indirect help.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.