Multi-Modal Search Engine Final Report

Jake Cooley, Lauryn Fluellen, Andrew Miranda, and Darshan Shah
CS 6200 Information Retrieval Final Project Report
Fall 2023

I. INTRODUCTION

This project focuses on developing a multi-modal search engine with a primary emphasis on movie-related queries. By incorporating text-based information from the IMDB dataset, we aim to provide users with comprehensive and enriched search results. This project is crucial as it addresses the increasing demand for diverse and relevant information in the domain of movies, catering to users who seek detailed insights such as specific scenes or iconic elements.

Team member contributions are listed below:

- Jake Cooley
    - Implementation of movie and cast query retrieval and query expansion code
    - Query collection for expanded annotated queries
    - Video relevancy evaluation
- Lauryn Fluellen
    - Query retrieval
    - Dataset import and download
    - Image, text, and video relevancy evaluation
    - Information collection for final report
    - Recruit participants for external feedback
- Andrew Miranda
    - Text retrieval for actors/directors and movies
    - Implementation of query expansion to text retrieval pipeline
    - Query collection for annotated queries
- Darshan Shah
    - Query processing
    - Image retrieval/crawling for actors and movies
    - Implementation of query expansion to image retrieval pipeline
    - Video and Gif retrieval pipeline with query expansion

II. BACKGROUND

A. *Dataset*

The IMDB dataset represents an integral repository of information within the cinematic

domain, providing invaluable resources for scholarly research, industry analysis, and the broader domain of entertainment. The IMDB dataset, sourced from the Internet Movie Database, encapsulates an extensive array of film-related information, including movie titles, release years, genres, ratings, and comprehensive cast and crew details. With millions of user-contributed entries, IMDB stands as a voluminous and diverse dataset, catering to the diverse needs of researchers, analysts, and enthusiasts. For this system, we will only be relying on actors' names, directors' names, and movie titles.

B. Query Expansion

Query expansion is a pivotal method in information retrieval aimed at improving system performance and relevance by introducing additional terms into a given query [1]. Various strategies exist for query expansion, with prominent approaches including pseudo-relevance feedback and word embedding-based expansion.

Pseudo-relevance feedback leverages highly ranked documents to identify new terms for query expansion, demonstrating effectiveness both with and without enhancements, such as a term proximity heuristic that considers the selection of expansion terms in proximity to the original query terms [2].

In the landscape of contemporary query expansion techniques, word embeddings assume a critical role. Widely employed in natural language processing, word embeddings map co-occurrence information to a lower-dimensional space using models like skip-gram or continuous bag of words. Within this framework, query expansion terms are identified by pinpointing the nearest neighbors to the query terms through these embedding models [3].

Modern techniques, such as word embeddings generated through models like skip-gram or continuous bag of words, contribute to identifying semantically related terms. The nearest neighbors to the original query terms are determined in the embedding space, providing an additional layer of context-aware query expansion.

III. IMPLEMENTATION AND DISCUSSION

A. Query processing

Our system's query processing mechanism involves a two-fold process of parsing and tokenizing user queries. In the parsing phase, the raw input is dissected to extract distinct components, including individual words, phrases, and punctuation. Subsequently, tokenization refines these components into standardized units, or tokens, forming the foundation of the query. This meticulous process is essential for comprehending the user's information needs and transforming the query into a structured format suitable for retrieving

relevant results, facilitating effective information retrieval through subsequent stages in the system's pipeline, such as query expansion and relevance ranking. We expect the queries to look like an actor's first and last name or a movie title.

B. Query Expansion

The query expansion implementation in this study follows a systematic process. Based on the initial query, we find the top 5 movies per the initial query as well as the top 3 cast members per each movie. Then we proceed with retrieving text and image results for the related subjects using our existing retrieval systems. Currently, modern techniques such as word embeddings, generated through models like skip-gram or continuous bag of words, contribute to identifying semantically related terms. Nearest neighbors to the original query terms are determined in the embedding space, providing an additional layer of context-aware query expansion. The query expansion component further enriches user queries by incorporating metadata from the IMDB dataset. Utilizing related films and actors, our system aims to broaden the query scope, leading to a more diverse set of results. Through this combined approach, the implementation seeks to enhance information retrieval precision and relevance, aligning with the research objectives and advancing retrieval system capabilities.

C. Image Processing

Our image processing module focuses on enriching the user experience by incorporating relevant visual content. Leveraging the Google Crawler library, our system retrieves images related to the user's query from various online sources. The process involves querying search engines for images based on both the original user query and the expanded query obtained through our query expansion mechanism.

By integrating images into the search results, our system provides users with a comprehensive view of the queried entities, enabling better understanding and engagement.

D. Video Processing

The video processing component of our system aims to extend the retrieval capabilities to multimedia content. Utilizing the YouTube API, we retrieve video search results based on both the original user query and the expanded query derived from our query expansion process. This two-pronged approach broadens the search scope, encompassing a diverse range of relevant videos.

The retrieved videos are then presented to the user, offering a mix of content that aligns with their information needs.

IV. RESULTS

A. Relevance Criteria and Result Organization

Relevant results for queries are defined based on the match to the user's query,

leveraging information from IMDB. The organization of results will be presented as the query text search result and the image search results for each of the terms. The image search results will display ten relevant images. These results are then evaluated and later passed to external candidates for feedback in the relevance evaluation.

B. Evaluation Metrics

To assess the effectiveness of our system, we will employ human judges to hand-evaluate the results. The human judges were both members of the group as well as four additional external students to aid in the feedback of our system. A binary scale will be used to gauge the relevance of each result to the user's query. This ensures a nuanced evaluation process, capturing the varying degrees of relevance.

The criteria for judging relevance will employ a binary relevancy measure, where a score of 0 indicates non-relevance and 1 signifies relevance. For movie queries, an image receives a relevance score of 1 if it matches the intended movie or includes one of the top 3 actors identified during query expansion. Notably, in the case of a movie series, an image is marked as not relevant if it depicts a movie from the series other than the specific one sought. When the user's query pertains to an actor, expanding to the top 5 movies they've been in, the system assigns a score of 1 if the image portrays the queried actor or any of the top 5 movies.

This binary system ensures a straightforward and clear assessment of the system's ability to deliver relevant content in response to user queries, offering a robust evaluation metric for performance analysis.

C. Performance Analysis

Evaluating our sample queries, we see that there was not a significant increase in text relevancy when performing query expansion, as the original queries already obtained a positive binary result for text relevancy. However, we see that query expansion had a significant impact on image relevancy and video relevancy for sample queries that are movie titles. Before performing query expansion, the image and video relevancies of our samples ranged from 0% to 75%. After expanding the queries; however, all sample queries had an image and video relevancy of 100%, indicating that query expansion has a significant impact on image and video retrieval performance.

V. CHALLENGES

Despite encountering challenges related to the video component and limitations in the IMDB dataset, our team has been actively addressing these issues. Textual results related to actors in the IMDB dataset have presented limitations. However, our ongoing efforts include exploring alternative solutions and

strategies to ensure a robust and cohesive project outcome.

VI. ACCOMPLISHMENTS

The project has earned an A grade by achieving a consistently high external evaluation score to validate the system's quality, implementing a feature to retrieve and display YouTube clips related to the queried movies or actors, and implementing feedback refining its functionality. We had additional external help within the relevancy evaluation to help ensure an unbiased evaluation of our system. Achieving an A- grade, the project extended query expansion to include diverse metadata and tags, accompanied by well-defined relevance judging criteria, contributing to a thorough evaluation process.

Securing a B+ grade, the project enhanced query processing with advanced parsing techniques and expanded query richness through a wider range of metadata and tags. Refinements in the evaluation process ensured more consistent relevance judgments, while external input broadened perspectives. Earning a B grade, the project developed a fundamental mechanism for parsing and tokenizing user queries, transforming them into a structured format. A basic query expansion mechanism using the IMDB dataset laid the foundation for result diversity, demonstrated through sample queries and initial evaluations focused on relevance.

VII. FUTURE WORK

In future work, we aim to refine and extend query expansion techniques, specifically exploring enhancements to pseudo-relevance feedback and word embedding-based approaches. Fine-tuning the term proximity heuristic in pseudo-relevance feedback and investigating alternative methods for identifying expansion terms from highly ranked documents could optimize this strategy.

Additionally, we plan to explore advanced word embedding models, such as contextual embeddings to improve term identification precision in the embedding space. These additional changes would offer avenues for tailoring the query expansion process to specific contexts. These efforts aim to further advance our information retrieval system, enhancing its effectiveness in delivering accurate and relevant results.

VIII. SAMPLE QUERIES, NARRATIVES, AND RELEVANCE JUDGMENT

Text Results Folder

1.

Original Query: "Scrooge: A Christmas Carol"
Query Intent: Find out information about the “Scrooge: A Christmas Carol” movie
Text Relevancy: 1
Video Relevancy:

Videos	1	2	3	4
Relevance	1	1	1	0

Image Relevancy:

Images	1	2	3	4	5	6	7	8	9	10
Relevance	0	0	1	1	1	0	1	1	0	1

Expanded Query: Scrooge: A Christmas Carol + Stephen Donnelly + Luke Evans + Scrooge + Olivia Colman + Past + Jessie Buckley + Isabel Fezziwig
Query Intent: Find out information about the “Scrooge: A Christmas Carol” movie and/or it’s cast
Text Relevancy: 1
Video Relevancy:

Videos	1	2	3	4	5
Relevance	1	1	1	1	1

Image Relevancy:

Images	1	2	3	4	5	6	7	8	9	10
Relevance	1	1	1	1	1	1	1	1	1	1

Expected Relevant Results:

- Details about the movie "Scrooge: A Christmas Carol," its adaptation, and key cast members
- Potentially, images related to the film's portrayal of characters and scenes

Expected Non-Relevant Results:

- Details about the movie "Scrooge: A Christmas Carol," its adaptation, and key cast members from the wrong Scrooge movie.
- Potentially, images related to the film's portrayal of characters and scenes from the wrong Scrooge movie

2.

Original Query: "The Tower"
Query Intent: Find out information about the “The Tower” tv show
Text Relevancy: 1
Video Relevancy:

Videos	1	2	3	4	5
Relevance	0	1	0	0	0

Image Relevancy:

Images	1	2	3	4	5	6	7	8	9	10
Relevance	0	0	0	0	0	0	0	0	0	0

Expanded Query: The Tower + Gemma Whelan + DS Sarah Collins + Emmett J Scanlan + Inspector Kieran Shaw + Tahirah Sharif + PC Lizzie Adama
Query Intent: Find out information about the “The Tower” tv show and/or it’s cast
Text Relevancy: 1
Video Relevancy:

Videos	1	2	3	4	5
Relevance	1	1	1	1	1

Image Relevancy:

Images	1	2	3	4	5	6	7	8	9	10
Relevance	1	1	1	1	1	1	1	1	1	1

Expected Relevant Results:

- Information about the tv show titled "The Tower," covering cast, plot, and genre.
- Images related to the TV show, such as screenshots or promotional materials.

Expected Non-Relevant Results:

- Information about the other TV shows and movies titled "The Tower," covering cast, plot, and genre.
- Images related to the wrong tv show, such as screenshots or promotional materials.

3.

Original Query: "Love Affair"
Query Intent: Find out information about the “Love Affair” movie
Text Relevancy: 1
Video Relevancy:

Videos	1	2	3	4	5
Relevance	0	0	0	0	0

Image Relevancy:

Images	1	2	3	4	5	6	7	8	9	10
Relevance	1	1	0	1	0	0	1	1	0	0

Expanded Query: Love Affair + Warren Beatty + Mike Gambril + Annette Bening + Terry McKay + Katharine Hepburn + Ginny
Query Intent: Find out information about the “Love Affair” movie and/or it’s cast
Text Relevancy: 1
Video Relevancy:

Videos	1	2	3	4	5
Relevance	1	1	1	1	1

Image Relevancy:

Images	1	2	3	4	5	6	7	8	9	10
Relevance	1	1	1	1	1	1	1	1	1	1

Expected Relevant Results:

- Information about the movie titled "Love Affair" covering cast, plot, and genre.
- Images related to the movie, such as screenshots or promotional materials.

Expected Non-Relevant Results:

- Information about the other TV shows and movies titled "Love Affair," covering cast, plot, and genre.
- Images related to the wrong TV show or movie, such as screenshots or promotional materials.

IX. REFERENCES

[1] E. N. Efthimiadis, “Query expansion.” Annual review of information science and technology (ARIST), vol. 31, pp. 121–87, 1996.
[2] Y. Lv and C. Zhai, “Positional relevance model for pseudo relevance feedback,” in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’10. New York, NY, USA: ACM, 2010, pp. 579–586. [Online]. Available: http://doi.acm.org.libproxy.mit.edu/10.1145/1835449.1835546
[3] F. Diaz, B. Mitra, and N. Craswell, “Query expansion with locally trained word embeddings,” arXiv preprint arXiv:1605.07891, 2016.

X. PROJECT CODE LINK

https://github.com/Darshan1510/CS6200-Final-IR-Project.git