LLMs as recommender systems: An alternative for tackling the cold-start problem

Felipe Aravena, Martín Ocqueteau

Abstract

This study investigates using Large Language Models (LLMs) like ChatGPT for the cold-start problem in recommender systems, focusing on movie suggestions. It compares LLMs, a content-based model, and Rotten Tomatoes' recommendations using a Kaggle dataset of 17,712 movies, testing on 60 diverse films. The methodology uses the Jaccard index to compare suggestions from ChatGPT, the content-based model, and Rotten Tomatoes. The results suggest that the recommendations made by ChatGPT are comparably effective to those of the content-based model, highlighting the potential of LLMs as a viable solution for recommender systems, particularly in addressing the cold-start problem. Future research should refine LLMs, expand test datasets, and improve user prompts for more relevant suggestions.

I. Introduction

Recommender systems are applications that suggest products, services, or information to users based on their preferences, tastes, or needs. These systems have become very popular in the field of e-commerce, entertainment, social networks, and other areas, as they help to improve user experience and increase sales or traffic for providers.

However, one of the main challenges of recommender systems is how to generate recommendations for new users or those with little available information, known as the cold-start problem [3]. This issue limits the ability of recommender systems to offer a satisfactory user experience and increase customer loyalty and retention.

Language models (LLMs) are artificial intelligence-based systems that can generate text from a given input, such as a word or a paragraph. These models are trained with large amounts of text extracted from various sources, such as books, articles, blogs, social networks, etc., and learn to capture the structures, vocabulary, and style of natural language.

LLMs can offer a novel and efficient alternative to traditional methods of recommender systems, as they can generate zero-shot recommendations, that is, without prior knowledge of the user's tastes, using only a single data point. This way of making recommendations has several advantages, including speed, flexibility, and personalization.

This work specifically evaluates the task of recommending movies, based solely on a single data point: a movie that a user likes. Movie recommendations from RottenTomatoes are compared with those from ChatGPT and a content-based recommendation model.

II. Dataset

In this academic study, a dataset downloaded from the Kaggle [5] platform was used, containing detailed information about 17,712 movies registered on Rotten Tomatoes (RT). Preprocessing of these data was a crucial step to ensure the quality and relevance of the analysis. Movies with low audience numbers were removed, and columns considered irrelevant, according to the study authors' criteria, were discarded. As a result, specific data about actors, directors, genres, the main genre, rating, and a summary of each movie were retained.

III. Methodology

To contrast ChatGPT's recommendations with a standard technique, the data were split into a test set with 60 selected movies and another training set to develop a content-based model.

The test set encompassed 60 movies from 1970 to 2020 across nine different genres, fairly distributed between these two categories. This methodology of sampling aims to capture a wide range of preferences and trends over five decades, thus improving the accuracy and relevance of the recommendation models to be developed.

A meticulous compilation of the five movies recommended by RT for each film in our test set was carried out. These data were obtained by visiting each film's dedicated page on RT and registering the recommendations by hand.

The training consisted, firstly, in converting the information of each movie into a tf-idf vector. Then, a function was developed that receives a movie and returns the 5 most similar movies to it within the training set. To evaluate ChatGPT's responses, recommendations for the 60 movies in the test set were compiled. These responses were obtained using the prompt[1]:

> "I like the movie '[name]' (year). What other 5 movies would you recommend based on that?"

Finally, Rotten Tomatoes' recommendations were obtained by manually visiting the detail view of each of the movies. There are always 5 movie recommendations found there.

Notably, these recommendations do not change according to the visiting user and remain constant despite page reloads and visits on different days. It is concluded, then, that they are completely based on the movie detailed in the view and not on variable external factors.

IV. Results

To quantify the obtained results, the Jaccard similarity, also known as the Jaccard index, was used. This is a metric used in set theory to measure the degree of similarity between two finite sets. Formally, it is defined as the size of the intersection divided by the size of the union of the sets. Mathematically, if two sets A and B are given, the Jaccard similarity, $\mathcal{J}(A, B)$, is calculated as

$$\mathcal{J}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

This coefficient results in a value between 0 and 1, where 0 indicates that the sets have no elements in common and 1 indicates that the sets are identical.

[1]This prompt has been translated to english for the sake of consistency. However, it was originally written in spanish.

For the test set, the similarity results with RT's recommendations obtained by each model were as follows:

Recommender	$\mathcal{J}_{RT}$
ChatGPT	0.0056
Content	0.0074

Tabla 1: Jaccard Coefficients

ChatGPT achieved a 75 % match in its recommendations compared to the content-based model. This is also visualized in Figure 1.

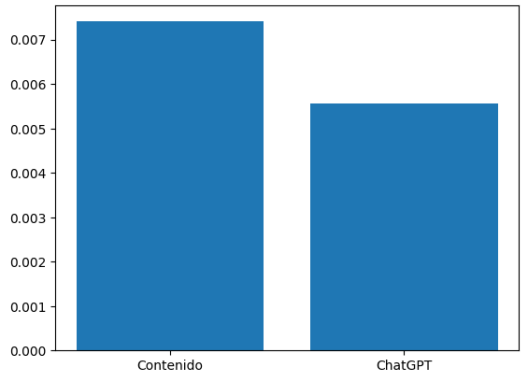

Figura 1: Jaccard Similarity Graph

V. Analysis

The analysis of the academic study reveals interesting comparisons between movie recommendations made by language models, such as ChatGPT, and those generated by traditional systems, including content-based systems and renowned platforms like Rotten Tomatoes (RT). The findings indicate that the recommendations provided by ChatGPT have a degree of similarity to those of Rotten Tomatoes comparable to that of the content-based model.

It is important to highlight that the effectiveness of the recommendations could have been influenced by the popularity of the evaluated movies. In this study, focus was placed on movies whose audience exceeded the median of the database used, which could impact the representativeness of the results.

Another relevant aspect is that most free language models operate with a "temperature" setting above zero by default. This implies that the responses

generated will not always be identical to the same prompt, introducing an element of variability in the recommendations.

Finally, it should be mentioned that Rotten Tomatoes is not necessarily the definitive source of the best film recommendations. Therefore, considering other sources as a reference or "ground truth" for evaluating the quality of recommendations could have enriched the study, offering a broader and possibly more balanced perspective.

VI. Conclusions

Given that a significant number of decisions were made arbitrarily (such as the number of recommendations, the choice of ground truth, the LLM to be chosen, and the size of the movie test set), the results obtained with different parameters of the experiment might vary considerably. It's important to consider that if the movie test set had been larger, this could have significantly impacted the results, providing a broader vision and possibly different insights.

The exploration of LLMs as a potential solution for the cold-start challenge has opened up interesting research opportunities. Although the effectiveness of LLMs is not yet fully established in this context, further investigation into this strategy is recommended to better understand its potential and applicability.

VII. Future Work

In the context of the future development of Language Models (LLMs), a series of strategies are proposed to enhance their effectiveness. Firstly, it is suggested to perform fine-tuning of these models by incorporating specific examples of recommendations. This approach has the potential to significantly improve the performance of LLMs. Additionally, repeating the experiment with an expanded test data set is recommended, which would provide a more detailed perspective on the degree of similarity between recommendations generated by LLMs and those produced by traditional recommendation systems. Lastly, it is advised to enrich user prompts by searching for and adding relevant information about the movie mentioned in the input. This prompt enrichment could contribute to generating more accurate and contextually appropriate recommendations.

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language models are few-shot learners.* arXiv preprint arXiv:2005.14165.

2. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners.* OpenAI blog, 1(8), 9.

3. Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). *Recommender systems survey.* Knowledge-based systems, 46, 109-132.

4. Tan PN, Steinbach M, Kumar V (2005). *Introduction to Data Mining.* ISBN 0-321-32136-7.

5. *Rotten Tomatoes movies and critic reviews dataset.* (2020, November 4). Kaggle. https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset? select=rotten_tomatoes_movies.csv