

CONTEXT-AWARE MOVIE RECOMMENDATIONS WITH LARGE LANGUAGE MODELS

C.B. Attidiya

208523X

Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

Faculty of Information Technology

University of Moratuwa

Sri Lanka

April 2024

CONTEXT-AWARE MOVIE RECOMMENDATIONS WITH LARGE LANGUAGE MODELS

C.B. Attidiya

208523X

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

April 2024

DECLARATION

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date: 2024/07/12

The above candidate has carried out research for the ~~PhD/ MPhil/~~ Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. A.T.P. Silva

Signature of the Supervisor:

Date:

DEDICATION

It is my pleasure to dedicate this thesis to the cherished individuals in my life, including my mentor Dr. Thushari Silva and my beloved parents, teachers, lecturers, and supportive colleagues who have offered their unwavering guidance and assistance throughout my personal and academic journey.

ACKNOWLEDGEMENT

I am immensely grateful to Dr. Thushari Silva, my supervisor, for his invaluable guidance, constructive comments, and vital support throughout the entire research process, which has contributed significantly to the success of this project.

Furthermore, I extend my gratitude to the Department of Computational Mathematics lecturers, my parents, and supportive colleagues for their invaluable assistance, encouragement, and guidance throughout the research process.

ABSTRACT

The rapid growth of online streaming platforms has resulted in a vast information overload of movie content, presenting a challenge for users in their day to day lives. Effective recommendation systems are vital for improving user experience by providing personalizing content to individual users. This thesis introduces an advanced approach leveraging cutting-edge artificial intelligence technologies, employing ChatGPT along with text-embedding-3-large embedding model, integrated with the Weaviate vector database, to create dynamic recommendation model for movie recommendation. We have generated embedding vectors for 26,942 movies with additionally retrieved metadata, to build this personalized recommendation system. Two distinct recommendation models were developed and were used to compare their recommendation capabilities. The first model creates a centroid reference vector for each user based on their movie ratings of three stars or higher. This reference vector is then used to search for similar movies within the vector database using L2-squared distance, aiming to match user movie preferences. The second model refines this approach by adjusting the centroid vectors according to the weight of user ratings, allowing for a more complete reflection of user tastes. Comparative analysis on recall, precision, F1 score, and hit rate (HR) for the top 5 and top 10 recommendations shows Model 1 achieving a recall of 1.8% and 2.81%, precision of 17.01% and 14.03%, and F1 scores of 3.15% and 4.45% respectively. These metrics demonstrate its effectiveness, with HR in the top 5 and top 10 recommendations reaching 48.97% and 63.66%, respectively. This indicates that Model 1 not only identifies relevant movies with relatively better accuracy but also does so more consistently than Model 2 and is computationally efficient. The results of this study highlight the potential of using advanced AI models to significantly enhance the precision and user satisfaction of movie recommendation systems. This approach allows for a deeper understanding of user preferences and the ability to predict user needs, ultimately leading to a more engaging and satisfying viewing experience.

Keywords: movie recommendation systems, large language model, vector databases, embedding model

TABLE OF CONTENTS

DECLARATION	i
DEDICATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xi
CHAPTER 1	1
INTRODUCTION	1
1.1 Prolegomena	1
1.2 Objectives	2
1.3 Background and motivation	2
1.4 Problem definition	3
1.5 LLM based Context-aware Movie Recommender System	3
1.6 System requirements	4
1.7 Structure of the thesis	4
1.8 Summary	4
CHAPTER 2	5
LITERATURE REVIEW ON RECOMMENDATION SYSTEMS USING LLMs	5
2.1 Introduction	5
2.2 Early developments in Recommender Systems	5
2.3 Recommender Systems with Artificial Intelligence	6
2.4 LLMs in Recommender Systems	8
2.5 Analysis of literature	11
2.6 Problem definition	14
2.7 Summary	14
CHAPTER 3	15
TECHNOLOGY ADOPTED	15

3.1	Introduction.....	15
3.2	Overview of the Technologies Used for LLM based context-aware Movie Recommendation	15
3.3	Embedding Model.....	16
3.5	Vector Databases	17
3.6	Model Evaluation Methods.....	17
3.7	Summary	18
CHAPTER 4		19
LLM BASED CONTEXT-AWARE MOVIE RECOMMENDER SYSTEM APPROACH.....		19
4.1	Introduction.....	19
4.2	Hypothesis.....	19
4.3	Input to the Recommender System.....	19
4.3.1	The MovieLens dataset.....	20
4.3.2	The TMDb dataset	20
4.4	Output of the Recommender System	21
4.5	Process	21
4.5.1	Model 1 Algorithm	23
4.5.3	Model 2 Algorithm	24
4.6	Users	24
4.7	Features	25
4.8	Summary	25
CHAPTER 5		26
DESIGN OF CONTEXT-AWARE MOVIE RECOMMENDER SYSTEM USING LLMS		26
5.1	Introduction.....	26
5.2	Top Level Architecture	26
5.3	Data Retrieval and Preprocessing	27
5.4	Vector Database	27
5.5	Recommendation Module.....	28
5.6	Presentation Module	29
5.7	Summary	29
CHAPTER 6		30
IMPLEMENTATION OF LLM BASED CONTEXT-AWARE MOVIE RECOMMENDER SYSTEM.....		30

6.1	Introduction.....	30
6.2	Data Preprocessing.....	30
6.3	Embedding model and Vector distance	33
6.4	Vectorizing Data	33
6.5	Recommendation Module.....	36
6.6	Presentation Module	36
6.7	Summary	37
CHAPTER 7		38
EVALUATION.....		38
7.1	Introduction.....	38
7.2	Evaluation parameters.....	38
7.2.1	Recall@K.....	38
7.2.2	Precision@K	38
7.2.3	F1@K.....	39
7.2.4	HR@K	39
7.3	Model Evaluation.....	39
7.3.1	Model Evaluation for Top 5.....	40
7.3.2	Model Evaluation for Top 10.....	41
7.4	Summary	42
CHAPTER 8		43
CONCLUSION AND FUTURE WORK		43
8.1	Introduction.....	43
8.2	Conclusion	43
8.3	Limitations and Further work	44
8.4	Summary	44
REFERENCES		45
APPENDIX A.....		47
DATA RETRIEVAL AND PREPROCESSING SOURCE CODE		47
	MovieLens Data Retrieval	47
	TMDb Data Retrieval	48
	Data Preprocessing for upload	49
APPENDIX B		50

VECTOR DATABASE RELATED SOURCE CODE	50
Vector database creation	50
Data upsert to vector database	51
UUID retrieval	52

LIST OF FIGURES

Figure	Description	Page
Figure 2. 1	Methods of using LLMs in Recommendation systems [1].	09
Figure 3. 1	Text to vector conversion by an Embedding model.	16
Figure 3. 2	Vector Database process architecture	17
Figure 4. 1	High-level framework of proposed solution.	22
Figure 4. 2	Centroid reference vector generation in embedding space.	23
Figure 5. 1	System module architecture for proposed system.	27
Figure 6. 1	Distribution of Movie Ratings.	31
Figure 6. 2	Movie Rating Distribution in millions.	31
Figure 6. 3	TMDb API usage for movie-context information.	32
Figure 6. 4	Sample data after preprocessing.	32
Figure 6. 5	Weaviate Vector Database creation.	34
Figure 6. 6	Weaviate Dashboard	34
Figure 6. 7	OpenAI usage cost.	35
Figure 6. 8	OpenAI API usage	35
Figure 6. 9	Presentation module implementation	37

LIST OF TABLES

Table	Description	Page
Table 2. 1	Analysis of Literature.	11
Table 6. 1	Sample data evaluation	33
Table 7. 1	Recall, Precision and F1 score for models at top 5.	40
Table 7. 2	Hit Rate for models at top 5.	40
Table 7. 3	Recall, Precision and F1 score for models at top 10.	41
Table 7. 4	Hit Rate for models at top 10.	41

LIST OF ABBREVIATIONS

Abbreviation	Description
BERT	Bidirectional Encoder Representations from Transformers
CF	Collaborative Filtering
CoT	Chain of Thought
GPT	Generative Pre-trained Transformer
HNSW	Hierarchical Navigable Small World
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network