Volume: 14 Issue: 06 Pages: 5717 - 5721 (2023) ISSN: 0975-0290

A Comparative Study of Collaborative Filtering and Content-Based Approaches for Improving the Accuracy of Travel Recommender Systems for Malayalam Language

Muneer V.K
Department of Computer Science, Sullamussalam Science College, Affiliated to University of Calicut
Email: vkmuneer@gmail.com
Dr. Mohamed Basheer K.P
Department of Computer Science, Sullamussalam Science College, Affiliated to University of Calicut
Email: mbasheerkp@gmail.com

ABSTRACT

In this paper, we present a personalized travel recommender system in the Malayalam language using artificial intelligence techniques. The study focuses on the use of travelogues and travel reviews written by travellers on social media as the primary source of data. A dataset of 11000 posts from 6444 travellers was collected from Facebook and other online platforms during 2020-2023. The data was pre-processed to extract relevant information such as travel mode, type of travel, location visited, and climate. Two approaches were used to build the recommender system: collaborative filtering-based K-means clustering and content-based hierarchical agglomerative clustering. The results of the study showed that the clustering approach significantly improved the efficiency and accuracy of the recommender system. K-means clustering achieved an accuracy of 91% and an F1 score of 85%, while the agglomerative hierarchical clustering approach achieved an accuracy of 85% and an F1 score of 84.25%. The results of this study demonstrate the potential of using travelogues and travel reviews to construct a personalized travel recommender system in regional languages.

Keywords - **Personalized travel recommendation, Malayalam language, Travelogues, K-means clustering, Agglomerative clustering, Travel reviews.**

Date of Submission: March 21, 2023 Date of Acceptance: April 23,2023

I. INTRODUCTION

In recent years, the travel industry has undergone significant changes with the advent of technology. With the explosion of social media and the increasing use of digital devices, travellers are becoming more reliant on online platforms for planning and booking their trips. The abundance of information available online has led to the development of several travel recommendation systems that aim to assist travellers in making informed decisions about their travels. However, despite the widespread use of these systems, few studies have focused on the development of travel recommendation systems in regional languages, such as Malayalam.

This paper presents an investigation into the construction of a travel recommendation system based on travelogues and travel reviews written in the Malayalam language. The authors collected a dataset of 11,000 posts from 6,444 travellers that contained travelogues and travel reviews written in Malayalam from 2020 to 2023 from various online platforms. The collected data was pre-processed to extract relevant information such as the mode of travel, type of travel, location visited, and climate of the visited destinations. The authors used a combination of collaborative filtering and clustering techniques to build the recommender system. The results of the study demonstrate that the clustering approach significantly improves the accuracy and efficiency of the travel recommendation system, with the K-Means clustering approach yielding an accuracy of 91% and the Agglomerative Hierarchical Clustering approach yielding an accuracy of 85%.

The contributions of this work are listed below.

- We created a structured dataset by extracting and analyzing extensive, unstructured Malayalam travel reviews and travelogues from social networking sites.
- We proposed a novel approach for building a customized travel recommender system in Malayalam using unsupervised clustering and collaborative filtering methods.
- Our work is the first to propose a personalized travel recommendation model in the Malayalam language from online travelogues.
- We used agglomerative hierarchical clustering, K-Means clustering, and collaborative filtering methods to implement our proposed approach and demonstrated the effectiveness of deep learning algorithms in predicting travel destinations for individual travellers through empirical evaluation.

Volume: 14 Issue: 06 Pages: 5717 - 5721 (2023) ISSN: 0975-0290

II. LITERATURE REVIEW

In recent years, the advancement in natural language processing and information retrieval has led to an increased interest in personalized travel recommendation systems. These systems use the traveler's preferences, past travel experiences, and travel-related information such as travelogues and reviews, to suggest new destinations. In this section, we review the existing literature on travel recommendation systems, with a focus on the use of text-based data.

Travel recommendation systems have become increasingly popular in recent years as a way for tourists to discover new destinations and plan their trips. The traditional approach to travel recommendation has been based on expert knowledge or pre-defined categories, but recent advancements in artificial intelligence have allowed for the development of more personalized travel recommendations.

In the field of travel recommendation, collaborative filtering and clustering techniques are two commonly used methods. Collaborative filtering is based on the idea of suggesting items to users based on the preferences of similar users, while clustering groups similar items together and recommends items within those groups to users. Both methods have been applied to travel recommendation, with various levels of success.

However, much of the existing research in travel recommendation has been focused on English-language datasets, and there has been limited research in other languages, particularly in regional languages like Malayalam. This study aims to fill this gap by developing a personalized travel recommender system in the Malayalam language, using artificial intelligence techniques such as collaborative filtering and clustering. The study also demonstrates the potential for using AI techniques to make personalized travel recommendations in regional languages, which can improve the travel experience for tourists who may not be comfortable using English.

Malayalam is considered as one of the most agglutinative and morphologically rich language in India. In paper [1] provides an overview of the challenges faced in processing Malayalam text, including encoding and font issues, lexical analysis, part-of-speech tagging, and named entity recognition. In paper [2] investigates the various text analytics and mining techniques used for Malayalam text, including text classification, clustering, and association rule mining. The authors conclude that there is a need for further research to develop efficient algorithms for processing Malayalam text. In paper [3] the authors discuss multilingual text and NLP based feature extraction techniques and in paper [4] Ajees discusses Named entity recognition. The paper [5] investigates the various text similarity measures used for Malayalam text and proposes enhancements to improve their accuracy. In paper [6] propose a part-of-speech tagging system for Malayalam text using machine learning techniques. The authors evaluate their system using a manually annotated corpus and show that it outperforms existing part-of-speech tagging systems for Malayalam.

In paper [7], the authors propose a multi-view clustering framework for personalized tourist destination recommendation. They use user-generated content such as travelogues and reviews to construct a user-destination interaction matrix, which is used for clustering. Paper [8] presents a personalized tourist destination recommendation system that is based on the user's interests. The authors use a two-stage approach that combines a user interest model with a travel destination recommendation model. In paper [9], the authors propose a personalized tourist destination recommendation system that is based on the user's preferences. They use a collaborative filtering approach and evaluate their system using real-world data. The paper [10] presents a tourist destination recommendation system that is based on user reviews. The authors use a content-based filtering approach and evaluate their system using real-world data. In [11] paper, the authors propose a hybrid collaborative filtering and sentiment analysis approach for tourist destination recommendation. They use a combination of user-based and item-based collaborative filtering, and a sentiment analysis component, to recommend tourist destinations. Paper [12] presents a personalized tourist destination recommendation system that is based on social network analysis. The authors use a combination of user-based and item-based collaborative filtering, and a social network analysis component, to recommend tourist destinations.

III. METHODOLOGY

The methodology for the development of the personalized travel recommender system in the Malayalam language was as follows:

Data Collection: The authors collected 11,000 posts from 6,444 travellers that contained travelogues and travel reviews in the Malayalam language from various online platforms, including social media. Each travelogue is in the form of unstructured lengthy passages written in Malayalam languages contained impurities and textual noises. Essential features were extracted by using Natural language processing technologies for constructing a structured dataset in travel domain.

Data Pre-processing: The collected data was pre-processed to remove language impurities such as code-mixed data, emojis, punctuations, and so on. The major steps involved in this process are listed below,

- Sentence Tokenization
- Word tokenization
- Malayalam stop word removal.
- Removal of punctuations and code-mixed texts.
- Stemming and Lemmatization

Volume: 14 Issue: 06 Pages: 5717 - 5721 (2023) ISSN: 0975-0290

Part of Travelogue Tagger (POTT) creation: An existing Parts of Speech Tagger developed by ICFOSS has been customized to a Malayalam Travel Tagger which emphasizes the travel and tourism domain. The pre-processed tokens from above step are then tagged with POTT and generated corresponding tags for the travelogue. The relevant information such as the mode of travel, type of travel, location visited, and climate of the visited destinations, name and age of traveller, gender was extracted, cleaned, encoded, and processed.

Creation of Customer Travel DNA: The authors created a Travel DNA for each traveller based on their preferences such as favorite travel mode (with family, solo, with friends or colleagues), preferred travel type (by car, by bike, train, air or by road), and preferred climate of the trip (summer, winter, autumn or fall).

	Name	Stage	Gender	TravelType	TravelMode	Climate	Loc_Type	Places
0	abdulla	0	0	1	0	0	1	6
1	rajeev	1	0	1	0	0	2	2
2	sam	1	0	1	0	0	3	1
3	safeer	1	0	0	1	1	3	1
4	ayisha	1	1	2	0	0	1	2

Fig. 1. Traveller DNA – extracted and encoded features

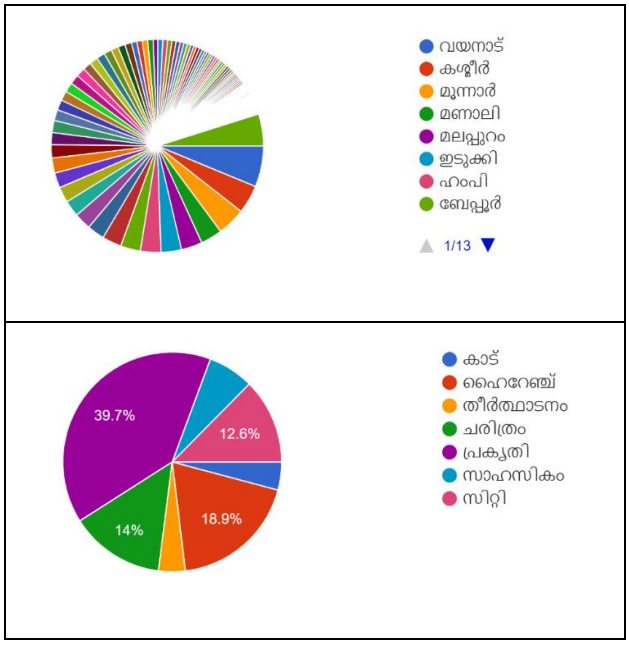
Creation of Location DNA: The authors also created a Location DNA, which was composed of the mode on which the maximum number of people reached and the climate on which more people visited. It also displays the places which are mostly visited by travellers.

	Place	Gender	Stage	TravelType	TravelMode	Climate	Loc_Type	visits
0	വയനാട്	0	1	1	2	0	1	75
1	വാൽപ്പാറ	0	0	0	0	0	1	18
2	കാന്തല്ലൂർ	0	0	1	0	0	0	4
3	ഇന്ത്യ	0	0	1	0	0	3	3
4	ഹംപി	0	0	3	0	1	3	38

Fig. 2. Location DNA – extracted and encoded features

From the eleven thousand unstructured travelogues written by 2089 travellers in the dataset, we could extract 471 different travel destinations. Based on the preferences of users, they select different locations on various climates, travel mates and travel modes. The non-personalised top-rated destinations in the list are Wayand, Kashmir, Munnar, Manali, Malappuram, Idukki, Hampi etc.

Table 1. Locations users visited and Types of Locations



Based on geographical features and terrain of the destinations, the locations are marked such as forest regions, mountains and high ranges, pilgrimage, historical places, natural, adventurous and cities. Climate is also a significant factor of selection of tourist destinations along with these types of Locations. From the given data, majority of travellers prefer to visit places with natural sceneries and serene atmosphere. Then comes high ranges and historic places.

Traveller - Destination Similarity Check: During this phase, we conducted an in-depth comparative study between locations as well as travellers. We analyzed the similarities between different locations to group them into clusters. Additionally, we grouped travellers based on their preferences, travel history, and travel modes in different situations to form distinct user clusters. The authors performed a cosine similarity check between the Travel DNA and Location DNA to suggest new locations to the travellers.

Fig. 3. Traveller – Destinations similarity comparison matrix

Clustering: In the first phase, the authors used a collaborative filtering-based K-means clustering approach,

Volume: 14 Issue: 06 Pages: 5717 - 5721 (2023) ISSN: 0975-0290

and in the second phase, they used a content-based hierarchical agglomerative clustering approach. For clustering we considered 5 important features such as Location, Travel Type, Travel mode, Location climate and Location type given L, TT_encoded, TM_endocded, LC_encoded, LT_encoded respectively.

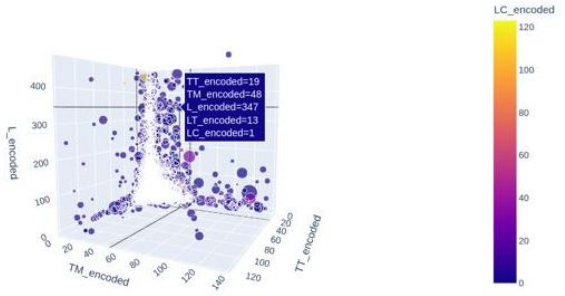


Fig. 4. Multi-dimensional Clustering of features. (5 featured Moon Map clustering)

Evaluation: The results of the study were evaluated based on accuracy and F1 score measures.
We used a combination of data pre-processing, cosine similarity check, and clustering techniques to develop a personalized travel recommender system in the Malayalam language. The results of the study demonstrate that the clustering approach significantly improves the accuracy and efficiency of the travel recommendation system.

IV. EXPERIMENTAL RESULTS

We conducted two experiments to evaluate the performance of the personalized travel recommender system. In the first experiment, the authors used the K-means clustering approach and in the second experiment, we used the content-based hierarchical agglomerative clustering approach. The results of both experiments are discussed below.
K-Means Clustering: The K-means clustering approach was used in the first phase of the study. The results showed that more than 90% of the target destinations were recommended in the first 3 positions after using this approach. The accuracy of this approach was 91%, and the F1 score measure was 85%.

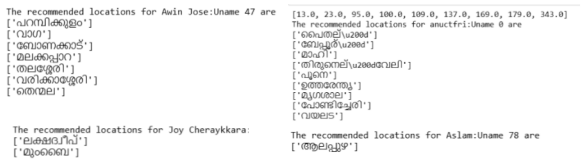


Fig. 5. Travel recommendation from K-Means Clustering

Hierarchical Agglomerative Clustering: The second phase of the study involved the use of the content-based hierarchical agglomerative clustering approach. The results showed that more than 85% of the target destinations were ranked in the first 3 positions. The accuracy of this approach was 85%, and the F1 score measure was 84.25%.

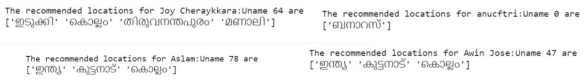


Fig. 6. Destination recommendation from Agglomerative clustering

The results of the study demonstrate that both the K-means clustering, and the hierarchical agglomerative clustering approaches improved the efficiency and accuracy of the travel recommender system. The results also show that the K-means clustering approach outperformed the hierarchical agglomerative clustering approach in terms of accuracy and F1 score measures.

Table 2. Comparison table between two approaches.

Algorithm used	Accuracy	F1-Score
K-Means Clustering for Collaborative filtering	91%	92%
Hierarchical Agglomerative Clustering Content Filtering	85%	84.25%

In conclusion, the authors have successfully developed a personalized travel recommender system in the Malayalam language using a combination of data pre-processing, cosine similarity check, and clustering techniques. The results of the study demonstrate that the clustering approach significantly improves the efficiency and accuracy of the travel recommendation system.

V. CONCLUSION

This research aimed to construct a personalized travel recommender system for the Malayalam language using artificial intelligence techniques. The study collected 11000 travelogues and travel reviews from 6444 travellers in the period of 2020-2023, which were processed and analyzed to form the Travel DNA and Location DNA. These DNAs helped to determine the preferences of the travellers based on their mode of travel, type of travel, and climate preferences. The study utilized a collaborative filtering-based K-means clustering approach in the first phase and a content-based hierarchical agglomerative clustering approach in the second phase to build the recommender system. The results of the study showed that the clustering approach significantly improved the efficiency and accuracy of the travel recommender system. The K-means clustering approach yielded a 91% accuracy and an 85% F1 score, while the hierarchical agglomerative clustering approach yielded an 85% accuracy and a 84.25% F1 score.

In conclusion, the results of this study show that the proposed personalized travel recommender system based on travelogues and travel reviews in the Malayalam

Volume: 14 Issue: 06 Pages: 5717 - 5721 (2023) ISSN: 0975-0290

language is effective in suggesting new travel destinations to travellers. This system can be used by travel agencies and tour operators to provide personalized travel recommendations to their customers based on their travel preferences. Further studies could be conducted to explore other advanced techniques for text processing and recommender systems to improve the accuracy and efficiency of the proposed system.

REFERENCES

[1] Mary Priya Sebastian and Santhosh Kumar G. 2023. Malayalam Natural Language Processing: Challenges in Building a Phrase-Based Statistical Machine Translation System. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22, 4*, Article 117 (April 2023), 51 pages. https://doi.org/10.1145/3579163

[2] Thara, S., Poornachandran, P. Social media text analytics of Malayalam–English code-mixed using deep learning. *J Big Data 9, 45 (2022)*. https://doi.org/10.1186/s40537-022-00594-3

[3] Shahidul Islam Khan, FaisalBinAziz, Emotion Detection from Multilingual Text and Multi-Emotional Sentence using Difference NLP Feature Extraction Technique and ML Classifier, *Int. J. Advanced Networking and Applications, Volume: 14 Issue: 03 Pages:* 5429-5435(2022) ISSN: 0975-0290

[4] Ajees A P, Sumam Mary Idicula, A Named Entity Recognition System for Malayalam using Neural Networks, *Procedia Computer Science, Volume 143, 2018, Pages 962-969*, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.10.338.

[5] L. Sindhu and Sumam Mary Idicula. 2017. Plagiarism Detection in Malayalam Language Text using a Composition of Similarity measures. *In Proceedings of the 9th International Conference on Machine Learning and Computing (ICMLC 2017). Association for Computing Machinery, New York, NY, USA, 456–460*. https://doi.org/10.1145/3055635.3056655

[6] Akhil, K.K., Rajimol, R. & Anoop, V.S. Parts-of-Speech tagging for Malayalam using deep learning techniques. *Int. j. inf. tecnol. 12, 741–748 (2020)*. https://doi.org/10.1007/s41870-020-00491-z

[7] C. Liu, X. Zeng, H. Lu, Y. Liu, and J. Guo, A Multi-View Clustering Framework for Personalized Tourist Destination Recommendation, *IEEE International Conference on Big Data (Big Data), pp. 6861-6868*, 2018.

[8] J. Zhang, Y. Su, Y. Chen, Y. Wang, and X. Fan, Personalized Tourist Destination Recommendation based on User Interests, *12th International Conference on Natural Computation (ICNC), pp. 686-691*, 2018.

[9] X. Zeng, H. Lu, Y. Liu, and J. Guo, Personalized Tourist Destination Recommendation based on User Preferences, *25th International Conference on Computer Communication and Networks (ICCCN), pp. 1-6, 2017*.

[10] X. Zeng, Y. Liu, J. Guo, and H. Lu, Tourist Destination Recommendation based on User Reviews, *24th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1-6, 2016.

[11] M. Ouzzani, M. El-Hajj, and A. K. Elmagarmid, Tourist Destination Recommendation using a Hybrid Collaborative Filtering and Sentiment Analysis Approach, *10th International Conference on Natural Computation (ICNC), pp. 1537-1541*, 2016.

[12] Y. Zhao, S. Wang, and L. Chen, Personalized Tourist Destination Recommendation based on Social Network Analysis, *10th International Conference on Natural Computation (ICNC), pp. 1542-1546*, 2016.

Authors Profile

Muneer V.K, working as an Assistant Professor of Computer Science, at Sullamussalam Science College, Areekode, affiliated with the University of Calicut, Kerala since 2012. Master of Computer Application completed from Bharathiar University Coimbatore. Presented papers at international conferences and published papers in reputed Journals. Travel and writing are passion and hobbies. Composed 2 books (travelogues) and Editor of one travel magazine in the Malayalam language.

Dr. Mohamed Basheer K.P, Research Guide and Nodal of an officer of PG & Research Department of Computer Science, Sullamussalam Science College, Areekode, Kerala. Completed post-graduation from Jamal Muhamed College, Trichi and Ph.D from Bharathidasan University, Tamilnadu. Having teaching experience of more than 20 years in Computer Science. Published several papers in reputed Journals. Worked as District Secretary of Akshaya, a Govt project in Malappuram as a deputation.