# Movie Recommendation System Based on Machine Learning

Yinghe Zhang*

Department of data science, City University of Macau, Macau, China

*Corresponding author: D22090104300@cityu.mo

**Abstract.** In today's era, with the growth of businesses on an individual basis, many applications are trying to provide more effective services to individual users, and eventually launched personalized services. Simply put, when a user enters certain keywords in a search engine, the recommendation system can push certain information or products to the user. Based on this principle, there are personalized recommendation videos on the homepage of short video platforms, and various music software have launched personalized radio stations. Even if you want to watch a movie on Netflix, it can recommend it to you. These applications are all looking for similar content in the content that users are interested in and recommending it. A simple movie recommendation system only matches similar movies through user history records, and the accuracy of matching is not high and easily affects efficiency. Therefore, a simple movie recommendation system can no longer meet user needs. The movie recommendation system that incorporates Spark and machine learning algorithms can better serve the public. This article will explore the impact of adding different algorithms to machine learning content on the matching degree of recommended content.

**Keywords:** Machine Learning, Movie Recommendation System.

## 1. Introduction

The application of recommendation algorithms is already very widespread, and the scope of application covers but is not limited to short videos, music, movies, and online shopping. And because different needs have to be met, different recommendation system algorithms have been introduced to enrich the functions of the recommendation system and meet the needs of various users. A representative example in the movie recommendation system is the Cinematch movie recommendation system developed by a video company called Netflix in 2003. The movie recommendation system uses a collaborative filtering recommendation algorithm, and the company uses a user rating system to predict members' liking for a movie.

The recommendation algorithm based on collaborative filtering mainly uses user rating data to predict content that users may like. The principle of this algorithm is that similar items will have similar evaluations for users. For example, if user Amy and user Bob are both young women, if user Amy likes movie A, then user Bob may also like movie A; if the type of movie A is similar to movie B, and there is a group of users who like movie A, then this group of users may also like movie B. Therefore, the recommendation algorithm based on collaborative filtering is subdivided into a user-based collaborative filtering recommendation algorithm and an item-based collaborative filtering recommendation algorithm. Through the research of scholars, it has been found that clustering, similarity, and classification techniques can be used to obtain better recommendations, thereby minimizing MAE and improving performance. And with the increase in the number of nodes in the Spark cluster[1], there has been a great improvement in recommendations.

In order to enhance the accuracy and performance of the recommendation system, some scholars have improved the item-based collaborative filtering recommendation algorithm by integrating feature tags[2]. By constructing a user-tag matrix with the user rating set and movie tag set, and combining the movie similarity model, the user's predicted rating for the movie is calculated, and the recommended results are obtained after sorting the user's predicted rating set for the movie in descending order[3]. Users can have multiple preference tags, and movies can also have multiple feature tags. The more the number of preference tags of the user corresponds to the feature tags of the movie, the more the movie conforms to the target movie recommended to the user. Experiments show that the user coverage rate of the algorithm that integrates movie tags reaches 86.52%, and the

accuracy and recall rate of recommended results reach 58.89% and 35.59%, respectively, effectively improving algorithm performance.

## 2. Method

The author will combine the Spark distributed computing platform to use the ALS algorithm, TF-IDF algorithm and K-Means algorithm to create a movie recommendation system with a hybrid recommendation algorithm. The following is an introduction to the Spark distributed computing platform and each algorithm.

### 2.1. Pyspark

In 2004, Google proposed MapReduce, which is the most primitive distributed big data processing framework [4], which effectively solves background problems such as task spanning scheduling. The early big data processing platform is Hadoop, an open-source project from Apache, whose core is MapReduce and HDFS [5]. However, MapReduce has some disadvantages, such as low efficiency: in order to ensure fault tolerance, all intermediate results must be stored on disk. There are also poor asynchrony and low resource utilization: the Reduce task will only start working after all the Map tasks are completed. Therefore, the demand for processing a large amount of data increases, and Hadoop is difficult to meet the demand. Later, scholars such as Zaharia developed a new distributed big data processing framework Spark [6], and Spark satisfies the demand for large data and fast speed. Not only that, Spark puts the intermediate data in memory, iterative calculation efficiency is high, and fault tolerance is high. Spark is a distributed computing engine, and pyspark is a python library dedicated to operating spark. To put it simply, Pyspark adds Python API on the basis of spark, interacts with Python and Java through integrated Py4J, and further realizes the use of Python to write Spark programs.

### 2.2. ALS algorithm

Weighted Regularized Alternating Least Squares (ASL) is a matrix factorization algorithm commonly used in recommender systems. It decomposes the user's rating matrix of the movie into two matrices: one is the user's preference matrix for the movie's hidden features, and the other is the movie's hidden feature matrix. This process is essentially a dimensionality reduction process. In a nutshell, the ASL algorithm achieves recommendation by decomposing the rating matrix into two matrices, which respectively represent the user's preference for the movie's hidden features and the movie's mapping on the hidden features[7]. The ALS algorithm based on the Pyspark platform, by adjusting the regularization parameter to 0.01, increasing the number of parallelized block calculations, and reducing the number of hidden semantic factors, can minimize the RMSE of the recommendation algorithm and quickly and accurately recommend products that users are interested in.[8]

### 2.3. K-means algorithm

Clustering is another way to recommend movies. By observing the movies that users watch, clustering is used to find similar movies and ultimately recommend the most similar movies. In our research, we used the K-means clustering algorithm for recommendation. The main goal of the K-means algorithm is to reduce the sum of the distances between points and their corresponding cluster centers. Here, the concept of inertia comes into play, which calculates the sum of the distances between all movies in a cluster and its cluster center. The goal is to make the inertia within a cluster as small as possible, while the inertia between two different clusters as large as possible.[9] Therefore, the K-means algorithm can aggregate similar movies to quickly process a large amount of data.

### 2.4. TF-IDF algorithm

TF-IDF is a weighting algorithm. TF refers to term frequency. IDF refers to the inverse document frequency, which reflects the frequency of a word appearing in all texts. If a word appears in many texts, its IDF value should be low. The larger the value of TF-IDF, the more important the feature word is in the article.

The expression of TF is Formula 1

$$\text{word frequency(TF)} = \frac{\text{The number of times a word appears in an article}}{\text{The total number of words in the article}} \quad (1)$$

The expression of IDF is Formula 2

$$\text{inverse document frequency(IDF)} = \log\left(\frac{\text{The total number of documents in the corpus}}{\text{Number of documents containing this word} + 1}\right) \quad (2)$$

## 3. Experiment

### 3.1. Dataset introduction

The data set used in the experiment comes from the famous data set website Movielens. The data set includes credits, keywords, links, movies_metadata, movies, ratings, tags. These include "userId", "movieId", "rating", "title", "overview", "genres", "tagline", "keywords". The data type is text, and the dataset sizes used are 189.9MB, 6.2MB, 198KB, 34.4MB, 494KB, 2.5MB, and 119KB. There are 9,743 movies, 610 users, and 100,837 reviews.

### 3.2. Training process

**Table 1.** Comparison Table of Experimental Variables for ALS Algorithm

| Method 1 | Method 2 | Method 3 |
|---|---|---|
| Parameter 1: rank=10 | Parameter 1: rank=10 | Parameter 1: rank=10 |
| Parameter 2: maxIter=10 | Parameter 2: maxIter=10 | Parameter 2: maxIter=10 |
| Parameter 3: regParam=0.05 | Parameter 3: regParam=0.1 | Parameter 3: regParam=0.2 |

Using the ALS algorithm, firstly, the rating matrix is decomposed into two low-rank matrices, one representing the user factor and the other representing the item factor. Then, it alternately fixes one of the matrices and optimizes the other until convergence. The author sets the parameters of the ALS algorithm to rank=10, maxIter=10 and regParam=0.1. That means it decomposes the scoring matrix into two rank-10 matrices and iterates 10 times to reach convergence. The regularization parameter regParam=0.1 is used to prevent overfitting. Latent semantic model for ASL: When the learning rate is 0.02 and the regularization coefficient is 0.01, the RMSE is the smallest and the model prediction is optimal.[7] At the same time, in order to ensure the recall and accuracy of the ALS algorithm in this recommendation system, the author needs to find a balance point in the parameters of the ALS algorithm, so the regParam parameter in the ALS algorithm is used as a variable for comparative experiments (Table 1).The ALS model uses the recommendForAllUsers(10) function to recommend 10 movies for each user. These recommendations are calculated based on user rating data for other movies. Then using the TF-IDF algorithm, the author used the HashingTF class to convert the preprocessed text into feature vectors. Then, it uses the IDF class to perform inverse document frequency weighting on these feature vectors to get the final feature vectors. Then use the K-Means clustering algorithm to divide the movies that meet the feature vectors into 20 categories, and each user recommends a category of movies. Finally, movies that belong to both ALS recommendation and K-Means clustering are selected.

## 4. Evaluation

The experiment evaluates the performance of the proposed method based on the accuracy of the recommendation results (Accuracy), which refers to the proportion of correctly predicted movies among all movies.

The accuracy rate expression is Formula 3. Details are shown in Table2.

$$\text{ACCURACY} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{3}$$

**Table 2.** Details for Formula of Accuracy

| Label / Prediction | Positive | Negative |
|---|---|---|
| True | TP | TN |
| False | FP | FN |

Recall is represented by R. The more result sets are retrieved, the better. This is the pursuit of "recall rate". The larger the value, the better.[10]

Recall expression is Formula 4

$$R = \frac{TP}{TP + FN} \tag{4}$$

**Table 3.** Comparison Table of Experimental Results for ALS Algorithm

| Method 1 | Method 2 | Method 3 |
|---|---|---|
| Recall: 0.79 | Recall: 0.82 | Recall: 0.789 |
| Accuracy: 0.69 | Accuracy: 0.71 | Accuracy: 0.69 |

According to the experimental results (Table 3), the recall rates of the three methods are ranked from largest to smallest as Method 1, Method 2, and Method 3. Comparing the accuracy of the three methods, the accuracy is ranked from largest to smallest as Method 3, Method 2, and Method 1. According to the definition of recall rate, the larger the value, the better, indicating that the 'completeness' is higher. At the same time, according to the definition of accuracy, the larger the value of accuracy, the more accurate the movies recommended to users. Therefore, balancing the accuracy and recall rates of the three methods, it can be concluded that Method 2 has better experimental results, so the ALS algorithm parameters 'rank=10 maxIter=10 regParam=0.1' are more in line with this experiment.

## 5. Conclusion

In this study, various recommendation systems were learned from other scholars, including the algorithms used in the recommendation systems. A hybrid recommendation system using Spark and machine learning algorithms was proposed in the article. Compared with traditional recommendation systems, this recommendation system can take into account both user ratings and movie content. It has been confirmed that movies with user preference features can be screened through user ratings, similar movies can be matched through movie content, and the final intersection is the movie recommended to users. According to this recommendation algorithm, the accuracy and recall rate of the algorithm can be effectively improved, and movies can be recommended to users more accurately. The next work is to improve the accuracy of matching similar movies using movie content, that is, to consider matching synonyms or near-synonyms of keywords. Explore a more stable and efficient recommendation algorithm.

## References

[1] Ying-jie ZENG. Design and Implementation of Movie Recommendation System Based on Spark and User Preference [D]. Zhejiang University of Technology, 2020.

[2] Smitha, N, et al. "A Review on Movie Recommendation System Using Machine Learning." IEEE Xplore, 1 Feb. 2021, ieeexplore.ieee.org/document/9388619.

[3] MA Shuai. (2022). Research on improvement of movie recommendation algorithm based on user behavior. Electronic Design Engineering (19), 60-64. doi: 10.14022/j.issn1674-6236.2022.19.013.

[4] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters[C]//Proceedings of 6th Conference on Symposium on Operating Systems Design and Implementation. San Francisco: IEEE, 2004: 107-113.

[5] YANG Dong-chen. Research and Design of Micro - Video Recommendation System Based on Spark Big Data [D]. Nanchang University, 2021.

[6] Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: Cluster computing with working sets. HotCloud, 2010, 10.

[7] ZHANG Xiao-qian. Recommendation System Based on Machine Learning Algorithms [J]. CHINA SCIENCE AND TECHNOLOGY INFORMATION, 2023(11): 104-107.

[8] XU Wen-ying, XIANG Qiang. Application and implementation of collaborative filtering recommendation algorithm based on Pyspark platform [J]. Journal 0f Southwest Minzu University (Natural Science Edition), 2018, 44(2): 202-207. DOI: 10.11920/xnmdzk.2018.02.013.

[9] Hrisav Bhowmick, Ananda Chatterjee, Jaydip Sen. "Comprehensive Movie Recommendation System." ArXiv.org, 23 Dec. 2021, arxiv.org/abs/2112.12463?context=cs.IR. Accessed 12 Aug. 2023.

[10] YAN Shui-ge, FU Zhen-zhen. Research on Dining Windows Recommendation AIgorithm for Students in Higher Education Based on Mahout [J]. Modern Educational Technology Centre, Nantong University, Nantong 226000), 2015(8): 24-27, 34. DOI: 10.3969/j.issn.1007-1423.2015.23.005.