Design of personalized movie and TV recommendation algorithm based on movie and TV big data

Xingfei Cheng[a,*], Yonghui Wang[a]
[a]Nanchang Institute of Technology, Nanchang 330044, Jiangxi, China.

ABSTRACT

With the popularity of the Internet and mobile terminals, the amount of movie entertainment information on the Internet has increased dramatically, and users' demand for personalized movie services is growing. Personalized recommendation service can be used to mine the hidden correlation between user's historical information, movie project information, user's historical operation log and data, and recommend the obtained video resources that users may be interested in to users, so as to better serve users. Among them, video recommendation is an important field of recommendation system technology research. The existing film and television recommendations are mainly popular recommendations and related recommendations. Popular recommendations are easy to lead to Matthew effect, while related recommendations are in line with users' preferences to a certain extent, but the degree of personalization is low. Different users often see the same recommendation list on the same play page. This article first analyzes user behavior data in the user behavior data modeling phase, and establishes an initial user behavior data model by combining user scoring behavior and labeling behavior. Finally, based on the improved recommendation algorithm proposed in the previous article, a video recommendation system is designed and implemented. First, the system requirements are analyzed, and then relevant designs are carried out according to the requirements. The system is implemented using the SS2H framework, and the main data table display and functional interface display of the system are given.

Keywords: Big data of film and television; Personalized film and television; Recommended algorithm

1. INTRODUCTION

With the continuous enhancement of China's Internet strength and the increasing demands of the people, China's Internet products have also been continuously enriched and improved, and China's computer Internet technology has also been more rapidly developed, making our shopping, living, entertainment, and other methods more convenient and intelligent. In this information age, data and information are also showing a rapid increase, but at the same time, it is particularly important for people to obtain relevant recommendations for their unique needs in this vast amount of data. Personalized recommendation technology is a means of information filtering that can mine users' interests and preferences, recommend interested information to users based on their interests, provide personalized services for users, and solve the problem of information overload [1-2]. Collaborative filtering algorithm is a widely used personalized recommendation algorithm that can identify similarities between users and movies based on user evaluations of projects, find the nearest neighbor of a target from similar users or movies, and make recommendations based on the information of the nearest neighbor [3].

Video websites are an important part of the Internet. At present, video websites contain tens of thousands of movies. In the Statistical Report on the Development of Internet in China, which was compiled by CNNIC in August 2019, the number of netizens in China was 854 million before the statistical date, among which the number of online video users reached 759 million, accounting for 88.8% of the total Internet users [4]. Most importantly, the recommendation system realizes a "one-to-one" information service mode, which is more in line with the individual needs of users. Different from the active search for information by users in search engines, the participation of users in recommendation systems is also lower, thus greatly reducing the cost of searching for information by users [5].

*E-mail:2020049@nut.edu.cn

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik, Tao Shen, Proc. of SPIE Vol. 12779, 127791U · © 2023 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2689519

Especially in the field of e-commerce, its role is particularly prominent. E-commerce websites are the earliest fields to use recommendation technology, with representative websites such as Amazon and eBay, and well-known domestic websites such as Taobao, JD.com, and Dangdang [6]. The combination of website operation and recommendation technology has brought enormous benefits to these e-commerce websites, such as Amazon, where about 20% of the revenue is generated by recommendation systems, making it known as the king of recommendation systems[7]. Therefore, the significance of research on recommendation systems is receiving increasing attention. Recommendation systems can not only provide users with excellent product experience effects, but also bring great economic benefits to enterprises. However, current recommendation systems still have problems such as low scalability, sparse data, and low recommendation accuracy. Therefore, studying personalized recommendation algorithms is of great significance [8].

This paper mainly introduces the relevant knowledge of collaborative filtering algorithm in recommendation system, including basic methods and theories, and mainly introduces the recommendation algorithm based on matrix decomposition, and analyzes the matrix decomposition model LFM and its parameter training method random gradient descent method in detail. Then, the distributed processing architecture based on Hadoop and Spark is studied, and a multi-node cluster experimental environment is designed and implemented. The Spark distributed architecture technology based on memory is fully utilized, which not only improves the data processing ability of the system, but also greatly improves the processing efficiency of the whole system, and realizes the movie recommendation system on this experimental platform.

2. KEY TECHNOLOGY AND BASIC THEORY OF RECOMMENDATION ALGORITHM

2.1 Model and application of recommendation system

In existing movie and television recommendation algorithms, it is commonly used to obtain the similarity of preferences among users by analyzing the rating information of users on movie resources, and to make recommendations based on the similarity of users. With the rapid development of the Internet, user tagging information is increasingly widely used in personalized recommendation [9]. First of all, we often encounter or use product recommendation functions in our learning and life. The most common application is to apply recommendation systems to online retail. On domestic e-commerce platforms, recommendation algorithm technologies are beginning to be added, such as the "guess what you like" function on Taobao. Recently, you often view and browse a certain category of products, and Taobao will recommend related products to you. The "guess what you like" function of e-commerce is mainly intended to improve the user's shopping experience, increase the sales of goods in online shopping malls, and thereby increase sales and profits [10]. Of course, with the continuous improvement of internet technology, the product recommendation function is constantly achieving real "personalized recommendation", and the items recommended to everyone are different, because they are all recommended according to everyone's taste. Memory-based collaborative filtering selects some neighbor users with similar interests for the target user, and predicts the target user's score value of the project according to the score of the neighbor users. Typical memory-based collaborative filtering includes nearest neighbor collaborative filtering and its improved algorithm. Nearest neighbor collaborative filtering is the most basic and understandable algorithm in recommendation system, which has been widely used in industry. It can be divided into two categories: one is item-based collaborative filtering algorithm, and the other is user-based collaborative filtering algorithm.

With the continuous growth of the scale of Internet information, the sparsity and scalability issues of the above-mentioned memory-based collaborative filtering algorithms are becoming increasingly serious. Model-based methods have begun to emerge. Model-based collaborative filtering learns a complex model based on training set data, and then deduces the target user's rating values for non rated items based on the model and the target user's rated data, which eliminates the need to call the database every time, Improves the speed and scalability of the system. In the process of recommendation, content based recommendation systems need not only appropriate techniques to describe items and user personal information, but also strategies to compare user personal information and items. The high-level structure of a content based recommendation system is shown in Figure 1.

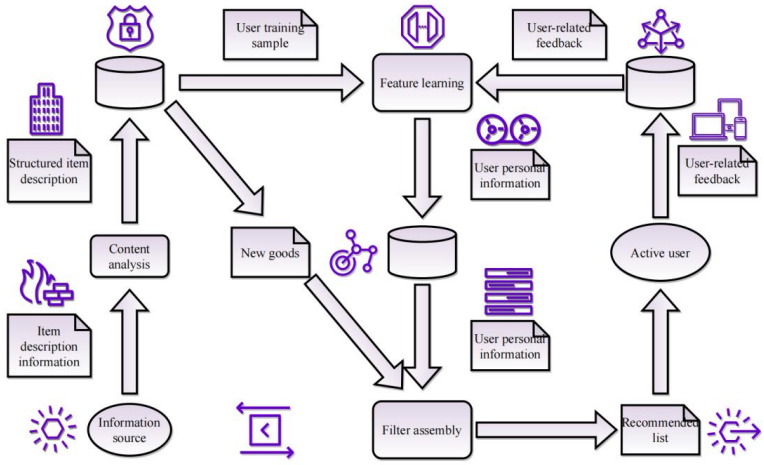

Figure 1. Content based recommendation algorithm model diagram

If the current information is unstructured, such as some text information, some preprocessing programs are needed to process the unstructured information, and the object content from the information source is expressed in a reasonable format, which is convenient for the next operation. Feature extraction and analysis are carried out for data objects, and the original information is converted into an item description format, such as a commonly used keyword vector format. The formatted information description becomes the input information of the information learner and the filtering component.

2.2 User behavior data modeling

In the research of personalized recommendation strategies, user behavior data is usually divided into two types: explicit feedback behavior and implicit feedback behavior. Explicit feedback behavior mainly involves the behavior of users explicitly expressing their preferences for items or resources, and the main form of expression is user ratings. Many websites like to use a 5-point or 10-point scoring system to let users express their preferences for items, such as domestic Douban.com and Time.com. The level of user ratings often expresses the user's preference for resources, while the user's annotation data expresses the user's preference. The combination of the two can effectively improve the degree of personalization. There are differences in user behavior data in different fields. For example, the accumulated user behavior data in the e-commerce field mainly includes user purchase records, shopping carts, wish lists, and so on; The user data that movie recommendation focuses on is the user's record of watching movies, rating values, tag data, and so on. Contrary to explicit feedback behavior, implicit feedback behavior usually cannot clearly reflect the real preferences of users, such as users' browsing behavior. If a user browses the page of an item, it doesn't mean that the user likes the item displayed on this page, but it may be because the link on this page is on the home page, so it is more convenient for the user to click on it. In order to explain more intuitively what explicit feedback data and implicit feedback data are, Table 1 lists examples of these two kinds of user behavior data in various websites.

Table 1. Examples of implicit feedback data and explicit feedback data in various websites

Website type/feedback information type	Implicit expression	Explicit formulation
video website	Users watch videos and browse the logs of related video pages.	Video rating, like or step on
News portal	Browse the portal's log	Comments on user-related reports
E-commerce / electronic commerce	Browse the log of the product page	Evaluation, praise or bad review of the purchased goods
Online music	A log of listening to music	A score or comment on music

Extracting effective behavior data can be used to model corresponding user behavior data, and then the similarity between users can be calculated through the processing of the established model to make recommendations. Therefore, analyzing user generated behavioral data is an important step in personalized recommendation, which will directly affect the final recommendation effect of the recommendation strategy. The degree of user preference for movie resources is usually reflected by the user's scoring behavior for the movie. The user's labeling behavior, that is, labeling watched movie resources with a certain label, also reflects a certain degree of user preference. At the same time, labeled labels often reflect the content characteristics of movie resources. Therefore, comprehensive consideration of users' scoring behavior and labeling behavior can more accurately capture users' preferences, which is conducive to improving the recommendation effect.

3. RESEARCH ON PERSONALIZED VIDEO RECOMMENDATION ALGORITHM DESIGN BASED ON VIDEO BIG DATA

3.1Traditional recommendation algorithms and technology platforms

Currently, recommendation algorithms, regardless of their technical ideas, have certain limitations. In content recommendation algorithms, it is generally the projects that users like or have followed that have certain similarities in content. Just as you have seen movie speed and passion 8 before, the content based recommendation system has detected a series of movies such as movie speed and passion 1, speed and passion 2, speed and passion 3, and so on. This series of movies has strong correlation with the content of the movies that users have watched, Whether it's the actors who starred in the movie, or the theme of the movie, the recommendation system then prioritizes these movies to you through calculation and processing. For example, if some users have similar ratings for a category of items, the recommendation system will assume that current users have similar ratings for these items. Therefore, according to the characteristics of this algorithm, in practical applications, the data is very sparse, which leads to the overall performance and recommendation accuracy of the recommendation system becoming worse and worse as projects and users continue to increase. User based collaborative filtering uses historical data to find user neighbors, and generates recommendations to target users based on the rating data of the nearest neighbor with similar ratings. The principle is that the nearest neighbor's rating of an item is very similar to that of the target user, so the weighted average of the nearest neighbor's rating of the item can be used to approximate the target user's rating of the unrated item, and recommendations can be generated for the target user based on the scored information of the nearest neighbor.

The functional modules of personalized recommendation system for movies are divided into three layers: interface layer, business logic layer and core data layer. The interface layer provides external service interfaces; The business logic layer realizes various corresponding algorithms and information management functions, which can be divided into two parts: user information management and film information management; The core data layer stores the core business data of the whole system, such as user information and movie information, user similarity list, movie similarity list and so on. As shown in Figure 2.

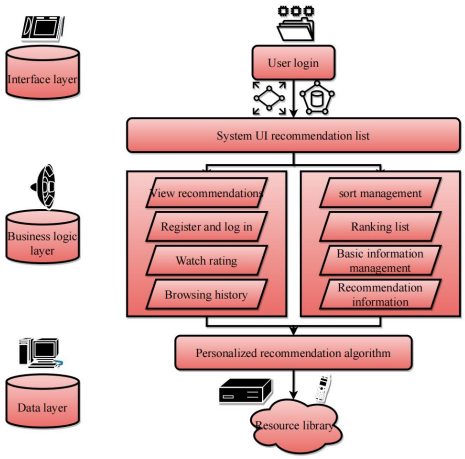

Figure 2. System Function Module

The modeling method used in this article is the VSM model modeling method introduced. It assigns values to users and program models based on user information data in the dataset, generates user interest data, and then completes the program recommendation process for channel users. The modeling part is mainly divided into channel modeling and program object modeling. Firstly, the process of establishing program model is introduced. The program attribute fields in the dataset are all structured and stored, so it is relatively convenient to describe programs using spatial vector models. In the process of film and television data modeling, the set P is used to represent the program set. Use vector representation as:

$$P = \{p_1, p_2, p_3, ..., p_m\} \tag{1}$$

Where PL and P2 respectively represent program 1, program 2 and m represent the total number of programs in the data set. The program represented in that data set are represen by a plurality of characteristic keywords. Let the program be represented by n attributes, and then the program PI can be represented by an L * n dimensional matrix, where n represents the number of attributes describing the program, such as director, screenwriter, starring, and so on. You can remember to do:

$$p_i = \{pa_1, pa_2, pa_3, ..., pa_n\} \tag{2}$$

Wherein the influence factor of each attribute field in the set D corresponds to each program attribute in the vector P. For example, wl represents the weight of the attribute represented by the program attribute pal.

Calculate the minimum value of RMSE as needed, so for the matrix elements of M and UV, so:

$$t_{ij} = \sum_{k=1}^{d} u_{rk} v_{kj} = \sum_{k \neq s} u_{rk} v_{kj} + x v_{sj} \tag{3}$$

At this time, x in the formula replaces u_{rs}, and $\sum_{k \neq s}$ is set here to represent the sum of $k = 1,2,...d$ except k = s.

Where $R'_s(u,i)$ is the predicted value of the user's rating on the goods. Therefore, the following formula is used in the ALS algorithm based on spark:

$$\min \sum_{i,(i,j) \in R} (a_{ij} - u_i^T v_j)^2 \tag{4}$$

Therefore, based on the above, it can be concluded that the optimal parameter value can be obtained through calculation during SVD matrix decomposition and dimensionality reduction. After code implementation, the recommended movie results can be obtained, and then the recommended results after training based on the ALS algorithm model can be linearly added to obtain the final recommendation results.

3.2 Analysis of experimental results

This paper designs a comparative experiment on the improved algorithm proposed previously. This algorithm is improved on the existing User based CF to improve the personalization and effectiveness of recommendations. The user's rating reflects the user's preference for movies, while the user's labeling behavior reflects the user's preference for movies. Combining the two can effectively improve the personalization of recommendations and improve the recommendation effect. This data set collects the actual ratings of users on the system over a period of nine months through a small Internet movie recommendation system. The processed data is divided into five small parts, and each part is divided into training sets and test sets based on an 8-to-2 ratio. In addition, the additional data set also contains basic information about users and movies, The user information includes the user's gender, age, native place, etc. The movie information includes the name of the movie, release year, etc. As shown in Table 2.

Table 2. Basic information of data set

Data set	Number of users	Number of movies	Scoring number	Sparsity
MovieLens 100k	945	1652	100000	6.32%
MovieLens 1M	6231	3571	1002130	4.56%
MovieLens 10M	70126	10268	1000096	1.4343%

The data sets used in the experiment are all preprocessed, including filtering users with less than 20 scores and users with only low scores and more than 20 scores. Pre-processing is to remove users who may have bad habits and destroy the system, and to remove other noises.

In the single-node mode, each set of data is run five times repeatedly, and then the average of the five experimental results is taken. In the cluster mode, the experiment is repeated five times for each group of data, and finally the average value is taken. This can reduce the error of experimental results to some extent. By comparing the experiment between the cluster computer and the single-node mode, and comparing the number of different data sets with the number of computing nodes, the purpose of this paper is to study and analyze the advantages of the distributed cluster architecture environment designed and implemented in this paper with the increasing processing data, as shown in Figure 3.

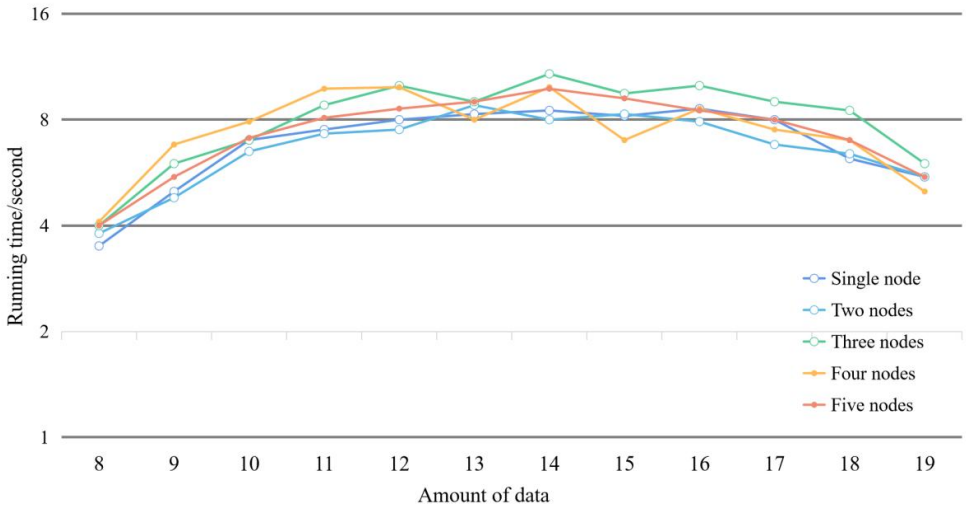

Figure 3. Runtime for Different Node Numbers

As can also be seen from the above figure, with the continuous growth of data volume on a single node, the system running time in the single node mode increases rapidly. In five clusters, with the growth of data volume, the system running time also increases, but compared to the growth speed of a single node, there is a significant improvement. From the trend of this data graph, it can be concluded that when the amount of data increases to a certain extent, such as tens of millions or even hundreds of millions of pieces of data, the advantages of the multi node cluster mode based on Spark will emerge, and the processing efficiency must have been multiplied compared to the single node mode.

In the experiment, experiments were carried out in stand-alone mode and cluster mode respectively. When there were 20,000 pieces of data, the hybrid algorithm proposed in this paper was used to carry out five experiments in stand-alone mode and cluster mode until there were 100,000 pieces of data, and then the RMSE value of each output was calculated, and the average value was taken as the final result. The following is a data comparison diagram after experiments on stand-alone and clusters with different node numbers, and the result is shown in Figure 4.

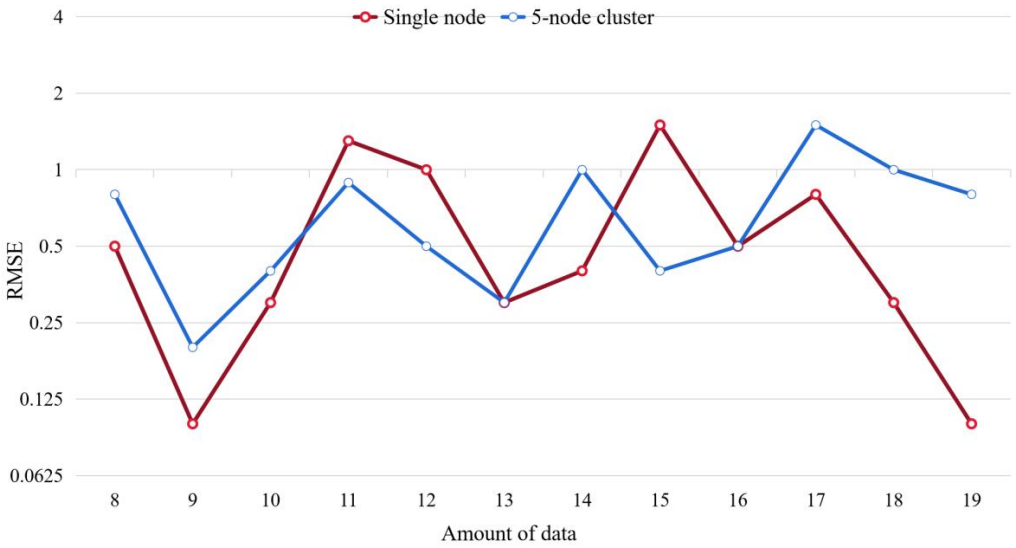

Figure 4. Recommendation results of single machine and cluster

When the amount of data in this experiment is relatively small, the RMSE value calculated by the distributed processing architecture has not changed very much, but the range of change can be predicted in the trend of data change. Because five computing nodes are used this time, if the number of nodes increases to a certain number, the recommendation accuracy will also increase obviously.

4. CONCLUSION

Personalized recommendation systems are currently a hot research area in big data and machine learning. With the requirements of high accuracy, efficiency, and massive data required for big data processing, personalized movie recommendation algorithms have moved from theoretical research in the laboratory to commercial applications, and personalized recommendation algorithms have been widely used in many products at home and abroad. The user's rating reflects the user's preference for movie resources, while the user's labeling behavior reflects the user's preference for movie resources. The combination of the two can effectively improve the personalization of recommendations. Generally, the entered search information may not fully express the user's wishes, and the obtained information may not necessarily achieve the desired effect of the user. Personalized recommendation service is the best way to solve this problem. Through very strict experimental comparison and verification, it shows that the hybrid recommendation algorithm based on Hadoop-Spark distributed cluster adopted in this paper has a certain acceleration effect, and the recommendation accuracy has also been improved to some extent. The personalized movie content recommendation system has a certain improvement effect, and the overall research results of this paper have certain significance and reference value for personalized recommendation system and distributed big data analysis.

REFERENCES

[1] Huang, J., Zhang, J., Chen, N., et al., Preference Degree Based Personalized Recommendation Algorithm. Shanghai Jiaotong Daxue Xuebao/Journal of Shanghai Jiaotong University, vol.52, no.7, pp.15 (2018).
[2] Chen, S., Huang, L., Lei, Z., et al., Research on personalized recommendation hybrid algorithm for interactive experience equipment. Computational Intelligence, vol.35, no.3, pp.12 (2020).
[3] Tripathi, A., Ashwin, T. S., Guddeti, R., EmoWare: A Context-Aware Framework for Personalized Video Recommendation Using Affective Video Sequences. IEEE Access, vol.7, no.5, pp.33 (2019).
[4] Liu, J., Choi, W. H., Liu, J., et al., Personalized Movie Recommendation Method Based on Deep Learning. Mathematical Problems in Engineering, vol.69, no.10, pp.34 (2021).
[5] Zhu, S., User Model-Based Personalized Recommendation Algorithm for News Media Education Resources. Mathematical Problems in Engineering, vol.44, no.7, pp.24 (2022).

[6] Li, C., A personalized recommendation algorithm based on large-scale real micro-blog data. Neural computing & applications, vol.32, no.1, pp.15 (2020).
[7] Cai, D., Qian, S., Fang, Q., et al., Heterogeneous Hierarchical Feature Aggregation Network for Personalized Micro-video Recommendation. IEEE Transactions on Multimedia, vol.44, on.8, pp.46 (2021).
[8] Xiao, J. S., Zou, W., et al., Video denoising algorithm based on improved dual-domain filtering and 3D block matching. IET Image Processing, vol.55, no.8, pp.11 (2018).
[9] Yang, J., Personalized Song Recommendation System Based on Vocal Characteristics. Mathematical Problems in Engineering, vol.10, no.6, pp.20 (2022).
[10] Ning, H., Li, Q., Personalized Music Recommendation Simulation Based on Improved Collaborative Filtering Algorithm. Complexity, vol.36, no.10, pp.47 (2020).