Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

"Emerging Technologies in the Computer World", January -2017

A Micro-video recommendation system using improved slope one algorithm based on Big Data

Prof. M. P. Nerkar[1] , Shraddha Bhurke[2], Nikita Kadam[3], Prajkta kavitake[4], Komal Handore[5]

[1]*Computer Department, AISSMS IOIT*
[2]*Computer Department, AISSMS IOIT*
[3]*Computer Department, AISSMS IOIT*
[4]*Computer Department, AISSMS IOIT*
[5]*Computer Department, AISSMS IOIT*

Abstract—*Asper today's generation the demand of internet and social networking service, the micro-videos are becoming more popular. The users spend their lot of time to search their favourite micro-videos from the amount of videos on the web, and also the producers of the micro-videos. Therefore, we have proposed a micro-video recommendation system. The recommendation algorithm are the important part of this system. The slope one scheme algorithm is based on rating of the users, and it is simple, efficient, easy to implement. However, the slope one scheme suffers from both new item problems and data sparsity, which affect the performance of recommender systems. To overcome the drawbacks of slope one scheme, we proposed an improved slope one algorithm which contain collaborative filtering algorithm. The slope one scheme is improved by introducing content similarity computation to overcome the new item problem. In item-based collaborative filtering algorithms, the target users rating to the target item can be predicted based on the ratings that the target user has rated and the content similarities of items. And clustering algorithm is used to overcome the problem of data sparsity. By combining the set of items into several different clusters based on the item rating data, the target users rating to the target users rating to the target item can be predicted based on which cluster the target item belongs to. The rating of target user to the target item is the linear combination of the above algorithms used.*

Keywords-*Micro-video, recommendation system, collaborative filtering, item content similarity, improved slope one scheme.*

I. INTRODUCTION

Micro-video is a short time video which lasts for 30 seconds to 5 minutes (300 seconds). For micro-video producers they don't know how many time their videos have been watched and also how many people like their videos. The micro-video are popular with teenagers, as they prefer to watch short video on their free time through mobile devices. Therefore this paper proposes a micro-video recommendation system (MRS) based on improved slope one algorithm. One of the purposes of MRS is to recommend videos to users. Another purpose of MRS is to provide overview of micro-video for the producer. In this way, the producer knows how many times their videos are on-demand and how many users love their video.

Big data is becoming popular as the development with the internet technology, which means the data sets cannot be handled by the current technology, to capture, manage and process the data within specified time. To enhance the MRS accuracy we need to collect large datasets about how many times the micro-video are on demand. In MRS first step is to collect data as far as possible from internet. We can download the data from video forum, video websites and different video chat websites and so on.

After the data collection step we get enough information of the micro-video, which contains the user watching information and micro-video information. According to the users history the micro-video can be automatically pushed to the user history the micro-video can be automatically pushed to the user by the MRS that the user is interested in. In this paper we use Hadoop framework as an offline big data analyzing platform for storing and processing large data.

II. RELATED WORK

In the field of recommender systems, collaborative filtering (CF) is a very successful recommendation technique. CF algorithms are divided into two categories: model-based CF algorithms and memory-based CF algorithms. Memory based CF algorithm find the nearest neighbors similar to the target user using some similarity based metrics, then predict the rating of the target user to the target item based on the ratings of his nearest neighbors, then recommend the item that has got the highest rating to the target user. In model based CF algorithm use the rating of a user to learn the user preference model, which is then used to make rating prediction. Modal-based CF algorithms are typically faster at query time though they might have expensive cost in learning and updating phases then memory-based CF algorithm.

In addition there is item-based CF algorithm-slope one scheme. In this algorithm we used K-means clustering algorithm to partition the set of items into several small clusters and then used the slope one scheme to get the video ratings. As well there is improvement in slope one algorithm in which content similarity of the item based on semantic is done.

III. ARCHITECTURE

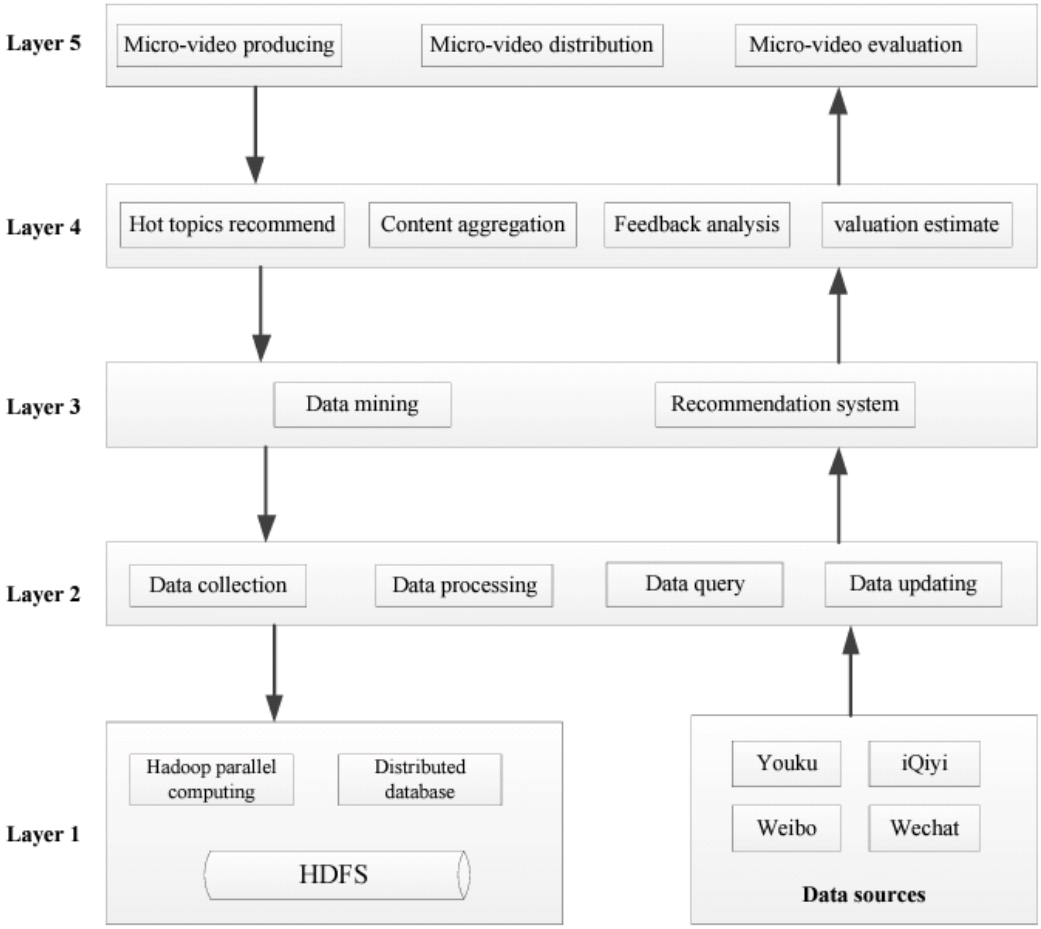

Figure 1.Architecture Diagram

IV.ALGORITHM

4.1. Similarity-Based Algorithm

Suppose that there are two items i and j , the feature vector of item i is (fi1,fi2,....fin)and the feature vector of item j is (f1j,f2j,...fnj) similarity between item i and item j is computed as in formula.

$$sim(i,j) = \frac{\sum_{k=1}^{n} f_{ik} \times f_{jk}}{\sqrt{\sum_{k=1}^{n} f_{ik}^{2}} \sqrt{\sum_{k=1}^{n} f_{jk}^{2}}} \qquad (1)$$

$$sim(i,j) = \frac{\sum_{k=1}^{n} f_{ik} \times f_{jk}}{\sqrt{\sum_{k=1}^{n} f_{ik}^{\ 2}} \sqrt{\sum_{k=1}^{n} f_{jk}^{\ 2}} + 1} \qquad (2)$$

We use the ratings of the target user to other items to predict the rating of the target user to the target item. The specific prediction formula is as following:

$$P_{ui} = \frac{\sum_{j \in I(u)} r_{uj} \times sim(i,j)}{\sum_{j \in I(u)} sim(i,j)} \qquad (3)$$

4.2. Slope One Scheme-

The slope one scheme is a rating-based recommendation algorithm which works on the intuitive principle of a deviation between items for users. And the deviation between any two items can be got by subtracting the average rating of thetwo items. The prediction process of slope one scheme consists of two sections:
(1).Calculate the average deviation matrix ij {dev }. Given a training set, we use formula 4 to compute the value ijdev of ij {dev }, ijdev is the average deviation of item i with respect to item j
In formula 4, user u rates both item i and item j . The deviation matrix ij {dev } is a symmetric matrix, and the matrix can be computed once and updated quickly when new data is entered .

$$dev_{ij} = \sum_{u \in U(i,j)} \frac{r_{ui} - r_{uj}}{N(U(i,j))} \qquad (4)$$

(2).Predict the rating of the target user to the target item. After the deviation matrix ij {dev } has been computed, we can use it to compute the prediction rating. Given the rating uj r of user u to item j , we can use ijujdev r + as the prediction rating of user u to the itemi , but a more reasonable predictor might be the average of all such predictions:

$$P_{ui} = \frac{1}{N(R_i)} \sum_{j \in R_i} (dev_{ij} + r_{uj}) \qquad (5)$$

where $R_i = \{j \mid j \in I(u), j \neq i, N(U(i,j)) > 0\}$ is the set of all relevant

This algorithm uses the number of the users that have rated both item i and item j as the weight. And the prediction formula is as follows accordingly:

$$P_{ui} = \frac{\sum_{j \in R_i} (dev_{ij} + r_{uj}) w_{ij}}{\sum_{j \in R_i} w_{ij}} \qquad (6)$$

where $w_{ij} = N(U(i,j))$.

4.3. Collaborative filtering methods

Following formula, is used in some collaborative filtering methods for similarity among users where the difference in each user's use of the rating scale is taken into account. where, Ri,s is the rating of item s by user i, As is the average rating of user i for all the co-rated items, and Ii,j is the items set both rating by user i and user j.

$$sim(i,j)=\frac{\sum_{s\in I_{ij}}(R_{is}-A_s)(R_{js}-A_s)}{\sqrt{\sum_{s\in I_{ij}}(R_{is}-A_s)^2\sum_{c\in I_{ij}}(R_{js}-A_s)^2}}$$

4.4. K-means clustering

Input: the training data set and the number of clusters k
Output: k clusters
1. According to the number of users that have rated items, we sort all items of the training data set in descending order. Then we select the top k items as centroids rather than k random items.
2. For each item i , find the centroid that is most similar to it, and if the j -th centroid is most similar to item i , we will assign item i to the j -th cluster C j . Here, we use Pearson's Correlation Coefficient as following formula 7 to compute the similarity between any two items.

$$S(i,j)=\frac{\sum_{u\in U(i,j)}(r_{ui}-\overline{r}_i)(r_{uj}-\overline{r}_j)}{\sqrt{\sum_{u\in U(i,j)}(r_{ui}-\overline{r}_i)^2}\sqrt{\sum_{u\in U(i,j)}(r_{uj}-\overline{r}_j)^2}}$$

Where, the larger the value Si j (,) is, the higher the similarity between item i and item j is. 3. For each cluster, we take the average of all the items assigned to it as the new centroid of it. 4. Repeat step 2 and 3 until the k clusters stop changing.

4.5. The fusion algorithm

Input: The training data set, the feature vectors of all items of the training data set and the test data set, the number of clusters k , the combination parameter λ , the target user u and the target item i .
Output: the prediction rating P (ui)
1. Firstly compute the item's content similarity matrix in formula 2, and then compute the prediction rating 1 Pui of the target user u to the target item i in formula 3.
2. Apply algorithm 2 to compute the prediction rating 2 P (ui) of the target user u to the target item i .
3. Finally, use the linear combination of the prediction ratings in step 1 and step 2 as the prediction rating of the target user to the target item, as shown in formula.

$$P_{ui}=\begin{cases}\lambda P_{ui}^{\ 2}+(1-\lambda)P_{ui}^{\ 1}, & i\in I\\ P_{ui}^{\ 1}, & i\notin I\end{cases}$$

where, I is the set of all items in training data set , i∈ I indicates that item i is not a new item, i∉ I indicates that item i is a new item, the value range of λ is (0,1). For different data sets, the value of λ that can lead to the best prediction performance is different.

V.CONCLUSION

According to viewers browsing or watching history, this system can recommend the favourite videos. The recommendation system can collect the reaction and give some suggestion for micro-video producers with how many viewers like the video. The users have no need to search their favourite video from the amount of videos. The slope one algorithm has the drawbacks of data sparsity and new items problem so, the improved slope one algorithm is used to overcome the drawbacks of it and improves the efficiency of the recommendation system.

VI.REFERENCES

[1] Songtao Shang, Minyong Shi, "A Micro-video Recommendation System Based on Big Data", IEEE, June 2016

[2] Jiyun Li, Pengcheng Feng, "An Improved Slope One Algorithm for Collaborative filtering", 2013 ninth International Conference on Natural Computation(ICNC).

[3] Y. Z. Li, T. Gao, "Design of video recommender system based on cloud computing", Journal on Communications, Vol. 34, No. Z2, pp. 138-140, 147, 2013.

[4] Y. Li, "Development mode of micro-video communication", Academic Exchange, Vol. 248, pp.177-181, 2014.

[5] S. M. Meng, W. C. Dou, "KASR: a keyword-aware service recommendation method on MapReduce for Big Data application", IEEE Transactions on parallel and distributed system, Vol. 25, No. 12, 2014.

[6] D. M. Zhou, Z. J. Li, "Survey of high-performance web crawler", Computer Science, Vol. 36, No. 8, pp.26-29, 53, 2009.

[7] G. Y. Su, J. H. Li, "New focused crawling algorithm", Journal of Systems Engineering and Electronics, Vol. 16, No. 1, pp.199-203, 2005.

