Micro-Video Recommendation System Using Improved Slope One Algorithm Based On Big Data

Prof. Minal Nerkar[1], Prajkta Kavitake[2], Shraddha Bhurke[3],Nikita Kadam[4],Komal Handore[5]
Department of Computer Engineering
[1, 2, 3, 4, 5]AISSMS IOIT

Abstract-*Recommender systems helps people in particular item selection, like what item to purchase, what news to read, or which videos to watch according to interest of the viewer. Current slope one scheme is a rating-based recommendation algorithm which is simple, efficient to use even though these algorithm have some disadvantage, like data sparsity and new item problem. To overcome these problem we proposed an improved slope one algorithm. In our algorithm, the data sparsity problem is dealed with by introducing clustering algorithm. By merging several items into a cluster based on the rating of item, we can predict the rating of the item based the cluster it belongs. And content similarity algorithm is used in slope one scheme to tackle the problem of new item. In these we check to which item the user has rated and the content analysis of that video is done and accordingly recommendation of the video with similar content is done. The final result for recommendation of item is the linear combination above two algorithms rating. Experiment done on sample data shows that algorithm is better than the other three slope one algorithms on the predictive performance*.

Keywords-HDFS, Improved Slope One Algorithm, HADOOP Map Reduce Analytics, NLP.

I. INTRODUCTION

Many users, they spend a lot of time to get their favourite micro-videos from amounts videos on the Internet; for the micro-video producers, they do not know what kinds of viewers like their products. The technology of object-oriented classification is now an inevitable trend for the development of High Resolution Video. We are using different techniques for Prediction of Video Recommendation to the registered user. Hence, we are proposing a system, which will take the input as an video and user will be able to comment and like the particular video. Thus, this concept improves the traditional recommendation algorithms, using the popular computing framework to process the Big Data.

At the Bid Data times, the challenges what we meet are data scale, performance of computing, and other aspects. Slope one recommendation algorithm is a parallel computing algorithm based on MapReduce and Hadoop framework which is a high performance parallel computing platform. The other aspect of this system is data visualization. Only an intuitive, accurate visualization interface, the viewers and producers can find what they need through the micro-video recommendation system.

II. RELATED WORK

HDFS

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

Features of HDFS

1. It is suitable for the distributed storage and processing.
2. Hadoop provides a command interface to interact with HDFS.
3. The built-in servers of name node and data node help users to easily check the status of cluster.
4. Streaming access to file system data.
5. HDFS provides file permissions and authentication.

HADOOP Map Reduce Analytics

Map Reduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.

Map Reduce is a processing technique and a program model for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce task is always performed after the map job. During a Map Reduce job,

Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster. The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.

Most of the computing takes place on nodes with data on local disks that reduces the network traffic. After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

TECHNOLOGIES TO BE USED

Hadoop

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage

Natural language processing

Sentiment analysis also known as opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional communication (the emotional effect the author wishes to have on the reader).

Sentiment is not analyzed via artificial intelligence, as some people may be tempted to think. Rather, it is analyzed via a systematic process that involves the use of a sentiment lexicon. This lexicon assigns a degree of positivity or negativity to a word by itself that is then used to give meaning to the entirety of the article. This is a way of analyzing sentiment, then, by considering a type of inherent positivity or negativity of each word that would be used by someone to talk about your business or products. For example, “happy “would be deemed a positive word, as well as “like” and “love” .At the opposite end of the spectrum we can see words like “hate”, “dislike”, etc.

III. SYSTEM ARCHITECTURE

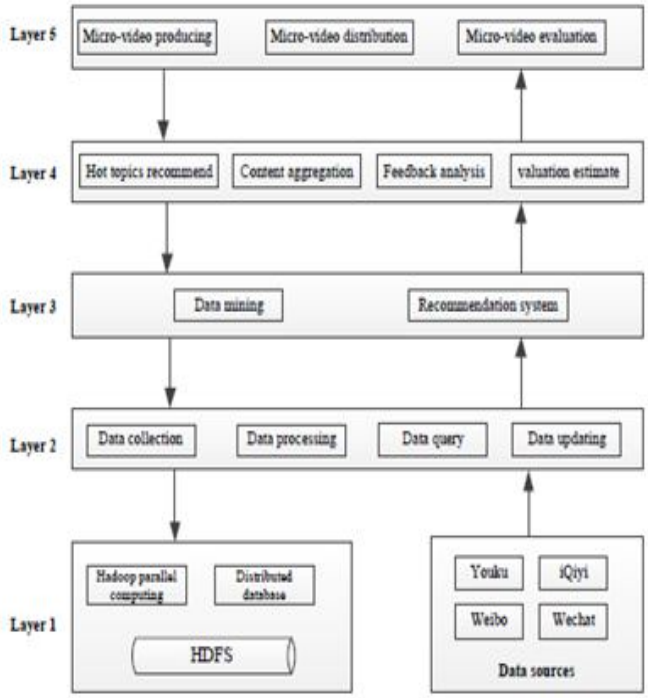

Fig. System Architecture

Layer 1

Data source is the basic of MRS, since all the data what we need are download from the web sites. The video websites are the main data source from which we can download, such as youku, iQiyi, and so on. The contents that we need download include the video ID, brief description, click rate, ranking list and so on. We also download video comments from social networking service web sites, for example, Weibo, Wechat, and micro-video forum .platform. All the data what we download from the web sites are stored in the HDFS (Hadoop distributed file system) .

Layer 2

Layer 2 is data interface layer, which contains data collection, data processing, data query, and data updating. The main idea of data collection is to download microvideo information from the data source automatically, i.e. web

crawler. Data processing transfer the download data format that the data mining algorithms or recommendation system can process. Data query is an interface for users to query the database. Data updating is a partner of data collection, which can update the download data simultaneously.

Layer 3

Layer 3 is the core parts of this system, including data mining algorithms, recommendation algorithms, and so on.

Layer 4

The layer 4 has two main functions. One of them is push recommend videos to user according to MRS results. The other function of this layer is to reflect the processing results to the micro-video producers who will take into account what kind of videos are on their next schedule.

Layer 5

The layer 5 is the interface layer for users. This layer contains many interfaces for many applications. The microvideo producers can use the micro-video producing interface to push their new video to target users through this platform. We can also use the micro-video evaluation interface to evaluate a video whether the audiences like it or not. Another function of Layer 1 is Big Data processing. We use Hadoop framework as Big Data storage and processing

IV. ALGORITHMS

Slope one algorithm

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 ij dev of ij {dev }, ij dev 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 .

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

(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 ij uj dev r + as the prediction rating of user u to the item i , but a more reasonable predictor might be the average of all such predictions:

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

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:

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

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

Improved Slope one algorithm

Slope one is a family of algorithms used for collaborative filtering, introduced in a 2005 paper by Daniel Lemire and Anna Maclachlan Arguably, it is the simplest form of non-trivial item-based collaborative filtering based on ratings. Their simplicity makes it especially easy to implement them efficiently while their accuracy is often on par with more complicated and computationally expensive algorithms. They have also been used as building blocks to improve other algorithms. They are part of major open-source libraries such as Apache Mahout and Easyrec.

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.

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

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

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:

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

Collaborative filtering algorithm

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.

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

Algorithm - 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 jth centroid is most similar to item i , we will assign item i to the jth 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.

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 , $i \in I$ indicates that item i is not a new item, $i \notin I$ 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

The core function of the system is the recommendation algorithms. The commonly recommendation algorithms are suite for tradition date sets, such as content-based recommendation, collaboration recommendation, and so on.

However, with the development of Big Data, the recommendation algorithms should can deal with the Big Data. The Slope one algorithm is a Big Data recommendation algorithm based on Hadoop framework. The MapReduce programming model is based on Hadoop framework, which can process Big Data. Thus, this concept use MapReduce programming model to implement the Slope one algorithm.

VI. APPLICATIONS

We can used for publicity of product.
We can used in Bollywood industry.

FUTURE SCOPE

The main idea of Slop one algorithm is to find the similarities between two items among the users. If we

compute the average rating of the two items, and compare the difference of them, the results can help us to predict another user's rating of those items.

With the development of Big Data, the recommendation algorithms should have the ability to deal with the Big Data. The Slope one algorithm is a Big Data recommendation algorithm based on Hadoop framework.

The MapReduce programming model is based on Hadoop framework, which can process Big Data. Thus, this paper use MapReduce programming model to implement the Slope one algorithm.

ACKNOWLEDGMENT

We would like to thank Miss. M. P. Nerkar for the valuable guidance she provided for completing this paper and the feedback she provided improved the scope of this paper.

REFERENCES

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

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

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

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

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

[6] A Micro-video Recommendation System Based on Big Datahttp://ieeexplore.ieee.org.sci-hub.cc/document/7550932/

[7] Big Data Analysis: Recommendation System with Hadoop Framework117.55.241.6/.../Big%20Data%20Analysis%20Recommendation%20System%20with....

[8] An Improved Collaborative Filtering Recommendation Algorithm Combining Item Clustering and Slope One Schemehttp://www.iaeng.org/publication/IMECS2015/IMECS2015_pp313-316.pdf

[9] Multithreaded Implementation of the Slope One Algorithm for Collaborative Filteringhttps://pdfs.semanticscholar.org/f560/5386c699b3be76e560700c560eef52b8da20.pdf