Web-users' behavior prediction using markov model with map reduce

1Prof. Vallabh G Patel, 2Prof. Komal D Anadkat
1Assistant Professor, Department of IT, 2Assistant Professor, Department of IT
1Government Engineering College, Bhavnagar, Gujarat, India,
2Government Engineering College, Gandhinagar, Gujarat, India

***Abstract* - As the usage of internet grows, the amount of data produces also increases by time The main challenge. is to store and analyze the data as it is difficult to handle these large amount of data, which proportionally increase the processing time and cost. All user navigation history is stored on the server in form of web log. To predict the next set of web pages that user may visit based on the user's previous web visit, we need to analyze the web log files. This prediction process in turns will reduce the user's web access time and pre fetch the next probable pages. Many prediction models are available like SVM, Neural network, ARM but in this research we used markov model. The main drawback of markov is the training time for markov is very large, which increases the response time. To solve this problem, we have trained the markov model in mapreduce programming model of hadoop framework, as it process the large amount of data parallely. The experimental result shows the improvement in response time without compromising the prediction accuracy.**

***keywords* - Markov model, N-gram ,hadoop ,log mining**

I. INTRODUCTION

As usage of internet increase drastically and the modern technologies invented currently, the amount of data produces is much higher than past many years. The data generated in last two years is 90% of the total data. All these generated information is quite useful in analysis and decision making in future. Big Data means a collection of large information that cannot be analyze using traditional computing techniques. [4] There are many categories of web mining[1] , and one is web log mining, which deals with extracting useful information from web data and use them for page pre fetching, decision making, performance analysis etc. Every time, a user surf the website all the navigation history about which pages user has been visited, is stored on the server in the form of web log.

Web log mining process consists of three steps like below:-
1) Data Preprocessing,
2) Pattern recognition,
3) Pattern analysis.

All the three steps are important but preprocessing is one of the necessary and difficult part [3] as it consists:-
1) Data cleaning, 2) User identification, 3) Session identification, 4) Path completion.

In the cleaning step the irrelevant entries get removed like noise, redundant data and robots, which save large amount of analysis time. All the user related information stored on the web server, but this information is very large in size. To get the useful knowledge from this information is not possible without analyzing this data. This knowledge will be useful for the administrator to improve their performance and services.

To predict the user behavior there are many problems like preprocessing and prediction. Preprocessing problems includes handling large amount of data which computer memory cannot accommodate, to select optimum sliding window size, identifying sessions, and to search domain knowledge. Prediction challenges include memory limitation, large training/prediction time and poor prediction accuracy. To overcome these challenges, I aimed to implement the web page prediction problem using MapReduce programming model of Hadoop framework. To process large data set hadoop is a freeware and famous software framework. MapReduce programming model of the Hadoop framework has the ability to rapidly process large amount of data in parallel. Hadoop use mapreduce programming language for writing distributed application for efficient processing of big amount of data, on large cluster of reliable hardware and fault tolerant manner. All the kind of data stored in the form of (key, value) pairs. This set of pair (key, value) will be given as an input, and first produce the intermediate list of (key, value) pair and finally produces output.[2] A MapReduce job will take input files in HDFS.

The organization of this paper is as follows: In section II, the related work is discussed. In section III, we present the proposed markov model implementation in MapReduce programming model of Hadoop framework. In section IV, we present the experiments undertaken for the web users' browsing behavior prediction. In section V the paper is summarized and future research works are outlined.

II. RELATED WORK

Many researchers have proposed different prediction models like Various models such as fuzzy interference models, support vector machines (SVMs), artificial neural networks (ANNs), association rule mining (ARM), Markov model, Kth order markov model, all-Kth Markov model and modified markov model to handle web page prediction problem. To prepare an efficient prediction model ,the proper comparison of each model is required.

Table1.1: LIST OF PREDICTION MODEL

Sr No	Prediction Model	Advantages	Disadvantages
1	N- GRAM [1]	When n>3, a precision gain on the order of 10% or more	As the n increases, there is an increase in precision and applicability will be decrease.
2	ASSOCIATION RULE MINING[1]	ARM do not generate several models	ARM endures efficiency and scalability problems by generating item sets and it require exponential time with the number of item sets.
3	MARKOV MODEL[1]	Efficiency and performance ,prediction time and good fault tolerance.	Higher order markov model has higher space complexity. Training time for model is large.
4	SUPPORT VECTOR MACHINE[3]	Accuracy in predicting seen and unseen data compare to Markov model.	It does suffer from scalability problem in both memory requirement and computation time.

III. PROPOSED WORK

Classification can be used to predict web page based on user's behavior and the history of his web usage , from this we can predict the next pages or sites which user can view or user will visit in future .Many prediction models are available now a days. Here we will use markov model with mapreduce technology. the main aim to use markov with mapreduce is to overcome the disadvantage of model which is high training time.

(a) Preprocessing

Here we considered NASA data sets which has total 1,891,714 log records and 1005 total pages. We considered only one month of user log record for prediction. The log data accumulated from NASA web server contains incomplete, noisy, and inconsistent values. Analysis of logs with these properties results in inconsistencies. The first step in data pre-processing is to clean the raw web data. During this step the available data are examined and irrelevant or redundant items are removed from the dataset. Data cleaning is carried out to remove irrelevant entries and reduce the size of original data.

User Identification is the process which identifies the unique user who has access the website. There are some rules to identify the user.[5]

1) Different IP address means different users.

2) The same IP but different operating system or different browser will be considering as different user.

(b) Session identification

The session identification splits all the pages visited by the IP address based on unique identity and timeout, when the time between page requests exceeds a certain time limit. Assuming that the user has started a new session for a particular IP address, the total time accessed by the user must not exceed 30 minutes. If the specified limit is exceeded a new session is considered for the same IP address.

The session is based on IP address, timestamp and URL referred. These fields are extracted from the preprocessed data already collected in HDFS and the same file can be used for other processing too as HDFS contains read once write many times property. For a unique IP that has logged in the server, all the total timestamp is collected along with referred URL. Using the Date function the timestamp is split into hours, minutes and seconds. The session length is calculated first by finding difference between the timestamp of same IP, tracked from login till logout. In Web prediction, the best known representation of the training session is the *N-gram*.

(c) Markov model implementation with mapreduce

In Markov model the probability that the process is in the current state is depends on the just previous state of the process. So a series of transition between different states, is called a markov chain. and the probability of next state, not on how the process arrived that state. The 0th Order Markov Model is based on page to page analysis which reads previous page to predict next page. The input weblog files are initially prepossessed. Let There are N user in log from last month , we will create sessions S(n) for every user and len is length of session for the same day ,so based on this the Probability P(N,S(n),len) can be predicted . From the map reduce Key and Value format we can decide that if an user is on page P(x) then the next page p(X+1) will be suggested by higher probability from session, P(X+1) = Pmax (p1,p2,p3.....pn) Where, pi belongs to P.

IV. EXPERIMENT AND RESULTS

The experiment uses the Mapreduce approach for efficient processing of log files. The process is carried out in Ubuntu OS with Apache Hadoop in pseudo distributed mode. The non hadoop approach of processing log files is performed on java in single jvm. The effect of cleaning process is shown in below table.

Table 1.2: EFFECT OF DATA CLEANING

Phase	No. of Records
Initial Log	18,91,709
After applying cleaning	7,04,279
Irrelevant Log	11,87,430

From the experiment results it is proved that processing of text files in single jvm on java takes more time than processing the same file in Hadoop Mapreduce.. This function is capable to do the processing in less amount of time when compared to java. it is observed that the time to execute **195MB** of dataset in both the environment shows a difference of **254** seconds. When scaling the dataset to terabytes or petabytes the time difference would vary in minutes or hours.

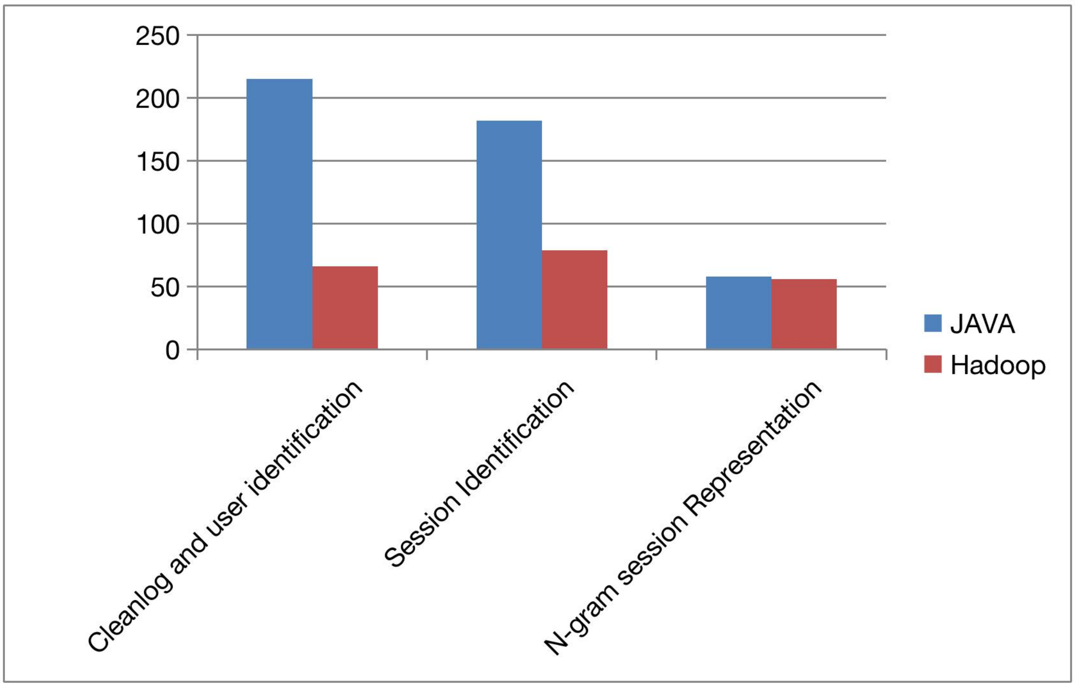

Fig.1 Comparison of time taken using JAVA and Hadoop

Markov model can predict the next set of WebPages based on previously visited pages. We have consider top 100 page model, For the series of experiments we have consider 10k ,20k and 25 k session with traditional java based approach and hadoop as well. The experiment was carried out using different order of markov model.

Table 1.3: TIME TAKEN FOR MARKOV

size	1-gram		2-gram		3-gram	
	Hadoop	Java	Hadoop	Java	Hadoop	Java
10k	21	17.3	22	20.7	27	18
20k	21	24	21	27	29	32
25k	20	26	21	29	28	32

We have also checked the accuracy of generated model with randomly sessions, and it confirms that: by doing the markov parallel (hadoop) there is no such negative effects on accuracy. It has almost same accuracy as with traditional approach.

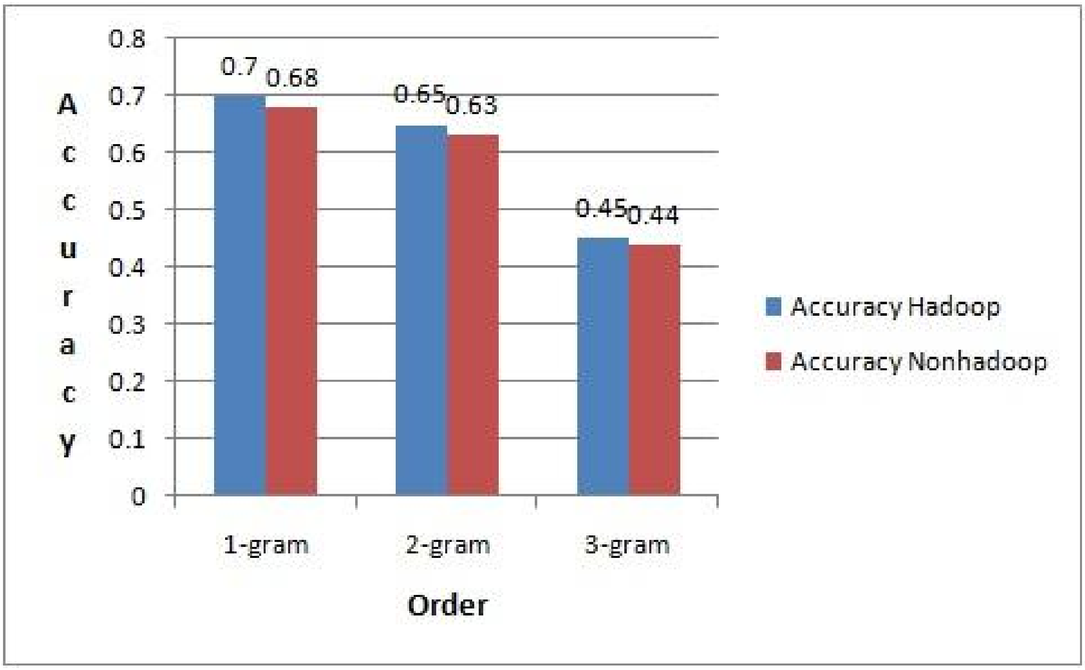

Fig.2 Comparison of accuracy with different order

V. CONCLUSIONS

The approach we applied will improve the quality of user behavior prediction by applying the technologies of distributed environments that is MapReduce. They will perform parallel operation for the prediction of the next page user may visit. This approach will also provide low space complexity because each page will have key and value in state or probabilities of the next page. We have implemented the cleaning, user identification, session identification in java as well as in hadoop mapreduce. N-gram session representation technique is useful for the representation of session. Experimental results shows the benefits of using hadoop over java as it takes less response time. Then in second phase we have created the different order of markov model and result shows that, if the dataset size is small traditional approach like JAVA is better than hadoop but as the data size grows,hadoop takes lesser time than non hadoop approach. So to deal with the bigdata hadoop provide less time to predict the user behavior. The accuracy is almost same in both the approach.

REFERENCES

1. Bing, L., Web Data Mining, 2nd ed., Berlin: Springer-Verlag, 2011.
2. J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Commun. ACM, vol. 51, 2008, pp. 107–113.
3. Zi-jun, C., Xin-yu, W., Wei, L., “Method of web log session reconstruction,” Journal of Computer Engineering, vol.33, 2007, pp.95-97.
4. ”Tutorials point”,may 2014, http://tutorialspoint.com/
5. He Xinhua,WangQiong,” Dynamic Timeout-Based a Session Identification Algorithm” Electric Information and Control Engineering (ICEICE), 2011 International Conference on ISBN 978-1-4244-8036-4,15-17 April 2011.
6. Mamoun A. Awad and Issa Khalil, “Prediction of User’s Web-Browsing Behavior: Application of Markov Model”, IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, vol.42, no. 4, pp. 1131-1142, August 2012.