Big Data Analytics and Web Based Data Mining

GOKILA. D[1], BRINDHA. K[2]
[1,2] *Assistant professor, Department of Computer Science, Sri Krishna Adithya college of arts and science, Kovaipudure, Coimbatore*

Abstract -- Big data analytics is a major thing that everyday people are indirectly connected with it and the level of frequency is very much higher in this case the Web based data mining is help full to do the work of the data analytics and gives the proper result for that data or a document. The work of web-based data mining is to calculate the value and give the data for the present time and this web-based data mining is a used to do the work as fast as it can and the usage of it is very much applicable to work though the data analytics. For example, the data which is created before 10 years the data that is applied and convert through web-based data mining can get the result for the present period of time and the value will be generated by itself and gives the result which we want that is about the wed based data mining.

I. INTRODUCTION

The web usage mining has various properties that let it to interesting in its own way and challenged too. There is a big amount of data in form of information that is still increasing like anything. The scope finding information of web is larger than and diversifying as it allows to abstract data. the data is available in multiple formats on the web just like images in jpg format, videos and text files also. The information on web is just up a to a level of anything and is also accessible across the world just in a few moments of time. We have done researches on a number of news websites and we have found that there recommendation system have a number of problems .Some recommendation system have not used the html news websites .There is a problem in this approach as with the increase in the news data it has now become very difficult to understand users perspective and also recommend data accordingly .the users interest has now become impossible to judge as because complexity of data available. Existing machine learning approach is not able to determine dynamic information frequently and accurately. Recently support vector machine learning algorithm is used to apply classify the news data sets. There are two methods applied for news data classification and abstraction the first one is naïve bayes classifiers and the other one is C4.5, etc. To make a tool for news abstraction for users, we use content-based rating techniques to classify the usage documents and predict the user behavior.

II. RELATED WORK

There Existing web data mining process facing number of challenges for that there is a great need of researches in this era: Web data with its diverse qualities do comprises of various ambiguity, noisy nature and a huge amount of unstructured data that is inconsistent in nature. Web data is continuing to expand in its own way so it is required to identify and discover the knowledge in particulars for the user interaction and behaviors .as there are various algorithms used in process of web mining so in case to measure the quality of algorithm ,its efficiency and complexity is to be judged .so it is really important to improve the performance of these complex algorithms .as the user data on web usage is increasing naturally.

Mark Werwath discuss the existing work on concerning neural networks to query answering. Enhancing the scheduling of for cancer patients Enhancing the placement of medical means beside the route of the Chicago marathons Building analytics founded fake news detector Building an algorithm for predicting housing prices based on past data Optimizing electric vehicle accusing for city of Evanston Enhancing prediction accuracy. Jun-Jie Zhang et al [2] big data has powerfully influenced theoretical research and applied application of enterprise presentation decisions. Collected enterprise managers and academic researchers essential reproduce to development and alteration of the times, receive and modernize in feature of marketing mix theory Innovation of Marketing Mix Theory after Big Data Perspective.

Chenggang Zhu et. al.[4]Furthermost of the current analysis concentration on construction a common model to predict the attractiveness of convinced content in a precise medium but inattention the enormous gap that advances as content popularity development progresses. As a consequence, those approaches are mostly ineffective for program approval prediction for broadcast TV, particularly when predicting process through initial peaks and later bursts of attractiveness.

III. MAP REDUCING MODEL BASED ON DEEP LEARNING

Fusion Level Training Testing Model (FLTTM): At the establishment of data pre-processing, we analysis with Web (Mix data set (Multi-dimensional)) numerous servers as well as the Web application (active data). To calcified connection the row files and then anonymized the consequential for information classification. Data fusion is an identical standard data processing system to variety up for the deficiencies affected by the missing data or noise information. This is used number of techniques. The principal module analysis (PMA) excerpts the principal module to fuse the training set, and the explicit data set information and implicit information are collective together to personalization to design Deep learning model using genetic algorithm for information classification and behaviour prediction. Our proposed deep learning model based on is a fusion model. Different number of operations performs in our proposed model first is training and testing. Training process perform in deep learning based on back propagation neural network. BPNN was constructed since completely of the weights is specified. Learning parameters and deep learning structure was received from the structure optimization request. If the Training ended when the deep learning using binary classifier error joined to the minimal value.

IV. PROPOSED FUSION LEVEL ALGORITHM

In Almost all machine learning algorithms are suffering from various difficulties of training phase due to increasing amount of data in data sets. These process are really expensive to operate on large scale .the calculation time and the space to store of SVM are majorly determined by vector space. this time taken for estimation and estimation complexity are majorly in a limited factor for machine learning .to overcome all these flaws of complexities, scientist have developed some sort of techniques, methods and various estimation. using some classification technique ,to classify the Semi supervise learning algorithm of particular domain in feature selection with binary neural network classification. by this feature gathering process, feature vector size can be decreased. a fusion learning algorithm is retrieved that is based on feature selection. this feature selection basically solves two problems, the first one increases the performance of resources and also increase the training sets while other method that helps to remove the noisy data and improve the accuracy and classification of data and expand the performance and improve the efficiency of map reduce model. Feature extraction approaches helped to remove the difficulties of dimensionality on increasing dimensionality. Feature extraction methods are used to achieve the curse of dimensionality that refers to the problems as the dimensionality increases. Through this approach, we can convert high dimension data sets to low dimension data sets .it contains number of information classification algorithms such as ICA, PCA, SVD etc. If We are available with big amount of data sets then we need to have a great quantity of computational reports .as existing algorithms have huge amount of classification problems s given by some researchers. now by this method we split data in two a number of sets, so that we can extract the features. Now here is non-support values we select the needed values and retrieve them out of nonsupport values. Now we are available with a large amount of data sets now by customizing them we can precisely obtain support vector. Many researchers have proposed a parallel SVM algorithm for training subsets, but we proposed semi supervise support vector. through which we got to obtain a final result value.

The whole MR Based Fusion Level S^3VM is a stereotype ,through which we can solve the various complex problems such as uncertain problems that can occur at any moment of time during the time of information classification .SVM is unable to rectify the thousands of bugs in training data sets .the researchers had previously developed a certain kind of

algorithms that are expensive to execute on a larger scale .by using this approach we provide optimize solution using map reduce model to classify the cloud computing data classification projects. Improved support vector machine for improving map reduce model using MR Based Fusion Level S^3VM

Algorithm description the following

Mapper: 1-Algorithm: 1

Step 1: Input: dataset1 datasate2…datasetn

Output: using Fusion Level S^3VM based on BNNC

Separated the data set given the reducer as an input

Step 2: Select the appropriate dataset as an input

Step 3: Perform the training

Step 4: Applying the appropriate weight and applying the activation unit.

Step 5: Generated training data set produce after preprocessing testing data set.

Step 6: Producing the output otherwise applying the step 4.

Mapper 2:-Algorithm 2

Step 1: Input the training set TD1, TD2, TDn

Step 2: Use the Binary neural network classifier to train the labeled sample set S to improvement the classification model CM1.

Step 3: Usage CM1 to train the unlabeled sample set P to label the samples;

Step 4: recursive train unlabeled sample set P till completely samples remained labelled; Step 5: Reprocess the entirely labelled training set TD to improvement the enhanced classification model CM2

Step 6: Contribution the training set TD into CM2

Step 7: Output the consequence applying reducer.

Step 8: Generated the output through the reducer Step 9: applying the back propagation learning for compute the error

Step 10: then generated the final output (classified results)

Here is the ten data mining process that has been done through the Web based data analytics.

V. EXPERIMENT RESULT

In this paper to use the online data set for performing the simulation which is involves in Entertainment, Politics, Social, Education, Research, and Online new. To distributions of the precise data are made known in the subsequent showing in graph. To perform the experiment using Apache Hadoop 2.7.2 framework and create up by 10, 20, 30 and 50 nodes etc. Data in size of 10000MBare stored disseminated in the platform. Every node is prepared with the Intel(R) CPU 2.2GHZ, 8GB memory and 500GB hard disks. In this research to perform the experiment nodes are connected through every other by Ethernet. Red Hat Linux used in the system. To evaluation of the performance of our proposed Fusion Level S^3VM with traditional BNNC, to execution to gather techniques under comparable conditions. To designate the consequence through the minimum error. Subsequent, we exclusion the proposed hybrid algorithm simply once, permitting the similar complete runtime as the traditional algorithm. Evaluation the modernization errors of the two algorithms, the proposed technique dependably generated a reduced modernization error.
Accuracy: To simulating and evaluation of number of semi supervise support vector diminution up to particular expand accuracy level and finding the accuracy number of different node. to evaluation and getting through the simulation the data size is very less that and number of node very large in this situation not more effect the training time but increasing the level of accuracy.

In a machine learning based information extraction in form of classification the number of current and accurately classified the pattern. To compute the level of accuracy in the term of performance to calculated following methods .

Accuracy = (accurately classified patterns)/(total input patterns) X100

Error rate: To compute the error rate using the dataset sample coming the results in the form of misclassified after the classification. To calculate the error rate using that formula

$$error\ rate\ \% = \frac{total\ misclassified\ patterns}{total\ input\ patterns} X100$$

Or

$$error\ rate\ \% = 100 - accuracy$$

Memory used: To computing the memory utilization in the execution of our proposed algorithm. our proposed algorithm run on the number of node that node applying in the simulation to calculate the total amount of memory utilize for evaluation the performance and find out the total free memory for allocation the different number of nodes.

$$memory\ consumption = total\ memory - free\ memory$$

To find out through the memory utilization computation cost of our proposed algorithm. To show the performance of huge amount of data Classification.

VI. CONCLUSION

We presented big web usage mining for predicting behaviour using fusion based map reduce model. This technique is used to extract the information using retaliation. the relevant documents that are of interest area to user are strapped and obtained from internet. The users interest areas are also linked by the user's usage keywords that sometime similar and linked so a user's interest can be easily determined conclude to use fusion based MR S^3VM algorithms based on back propagation neural networks to improve the prediction and recommendation.

REFERENCES

[1] Mark Werwath," Implications of Big Data for Data Scientists and Engineers".

[2] Liu Shangdong, JiYimu , Zhang Dianchao , Yuan Yongge, Gong Jian ,Wang Ruchuan ," An Online Prediction Algorithm of Traffic in Big Data Based on the Storm".

[3] Jun-Jie Zhang, Li Yang," A Simple Analysis of Revolution and Innovation of Marketing Mix Theory from Big Data Perspective".

[4] Chengang Zhu, Guang Cheng, and Kun Wang," Big Data Analytics for Program Popularity Prediction in Broadcast TV Industries".

[5] Ni GAO, Ling GAO, QuanliGao, Hai Wang," An Intrusion Detection Model Based on Deep Belief Networks".

[6] Shi Cheng, Yuhui Shi, Quande Qin, and RuibinBai," Swarm Intelligence in Big Data.

[7] Yi Wang, Student Member, IEEE, Qixin Chen, Senior Member, IEEE, Chongqing Kang, Senior Member, IEEE, Qing Xia, Senior Member, IEEE and Min Luo.