International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

# Big Web Data Mining for Predicting Usage Behaviour Using Fusion Map Reduce Model

**Anand Singh Rajawat, Dr. Akhilesh R. Upadhya**

Research Scholar, Shri Jagdishprasad Jhabarmal Tibrewala University, Churela, Rajasthan, India

## ABSTRACT

Unique of the greatest common problems that appearance pattern discovery, analysis and recommendation technique is dealing with the huge volumes of information in the form of data on the Web, and consequently the scalability of information classification recommendation and analysis given the write results is currently a big issue. Scalability means the rate of execution time, memory utilization, error control and accuracy required for the task, conferring to the parameters or factors that stimulate the performance of the algorithms, such as number of users or pages. Difficulties with the data itself. Complications in considerate the framework of search requests. Complications using classifying the alterations in user's information requirement. The pre-existing machine learning algorithms are unable to solve this in a better way. The current application of this for data classification is really expensive in nature Improve the recommendation technique using map reduce model based on the machine learning: to proposed technique for Big Web Data Classification For User Behaviour Predicting Using fusion based MR S3VM algorithms. The experimental results show that fusion based map reduce model is extremely appropriate for modelling a classification model among high accuracy , less time , less memory utilization and that its performance is better to that of traditional machine learning classification algorithm.

**Keywords :** Big Data, Neural Network Model, Information Mining Methods, Mapper, Reduce, and BPDL.

## 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 also 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 . As the

recommended data can be of users interest as well as of no any interest so it is really important to have on time feedback to recommend user as per his interest.
To implement our proposed model on a map reduce model and improve the response time to existing algorithm To frame that work as follows. Section 2 presents related work for data classification. Represent the working of existing deep learning model in Section 3. Proposed hybrid Map Reduce model is drafted in Section 4 illustrates a result analysis techniques is discussed finally to conclusion our proposed model very effective exiting one.

## 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 et al.[1] 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 an 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 modernise in feature of marketing mix theoryInnovation of Marketing Mix Theory after Big Data Perspective.

Chengang 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.

Ni GAO et al.[5]they have been proposed intrusion detection model based on the greedy multilayer DBN this model used for unlabelled information extraction. This work discus about DBN can be applying for large data set unlabelled data unsupervised learning.
Shi Cheng et al[6]In this research work they have been discuss about effort of big data analytics problematic is analysed. This work can be classified the four components: handling huge quantity of data, conduct high dimensional data, supervision run time data set. Furthermost real world big data problems can be exhibited as a huge scale. Swarm intelligence has presented important accomplishment on these problems. They have been proposed swarm intelligence, additional current approaches can be intended and operated in the big data analytical problems.

Yi Wang in et al.[7]proposed sparse coding-based data density method cannot first diminish the size of a dataset but similarly successfully mine PUPs after massive load profiles for dissimilar applications. As well clustering also the furthermost normally used technique for classifying the characteristic pattern of every customer to permit that customer's consumption behaviour to be labelled in relationships of numerous distinctive load patterns.

## 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 anonymised 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 technique. 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 operation 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. To show in figure 1 process of changing the Training weight and learning parameter. Produces by conception alterations in weight values by expending gradient method preliminary at the output layer then moving backward complete the hidden layers of the network and henceforward is prone to lead to troubles such as local minimum problem, measured merging pace and convergence instability in its training process. The respectable belongings of binary dissimilar technique (deep learning based on binary neural networks and back propagation neural network) by relating them to difficulties to resolve proficiently this are improving by the fusion algorithm. An infrequent process termed mutation similarly fluctuations particular traits. Primarily, behaviour predicting of online newspaper complete separate model. Furthermore, concept a deep learning neural network model captivating forecasting consequences as input value and definite consequence as output value. Lastly, train the model expending BPNN.

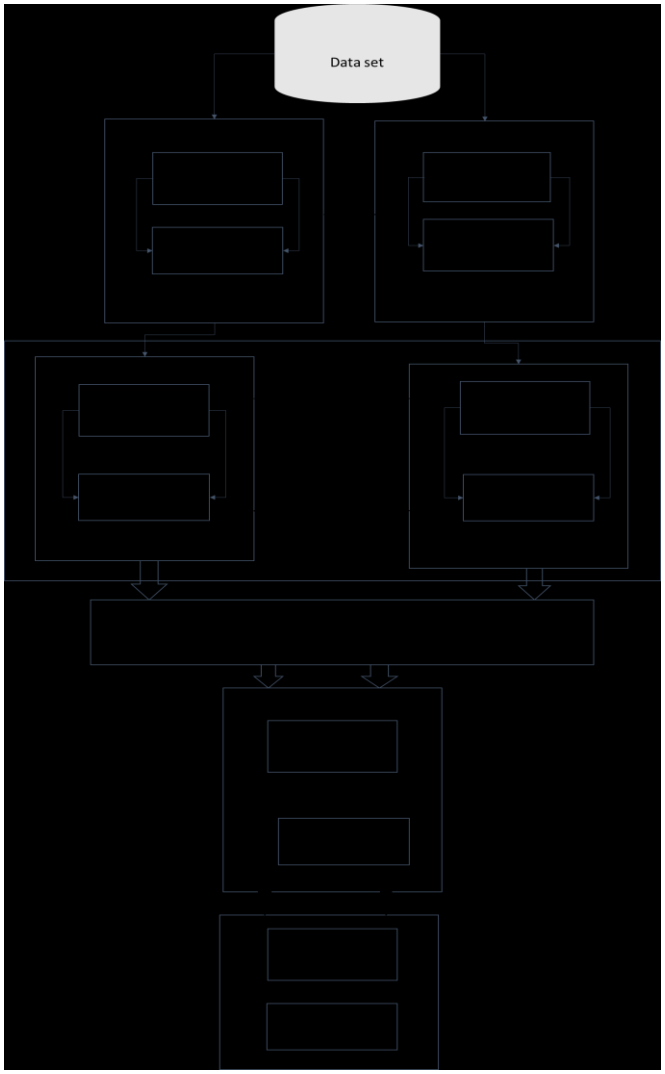


**Figure 1:** Proposed Fusion Level $S^3VM$

## 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 non support 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: $dataset_1$ $datasate_2$…$dataset_n$

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 pre-processing testing data set.

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

Mapper 2:-Algorithm 2

Step 1: Input the training set $TD_1$, $TD_2$, $TD_n$

Step 2: Use the Binary neural network classifier to train the labeled sample set S to improvement the classification model $CM_1$.

Step 3: Usage $CM_1$ 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 $CM_2$

Step 6: Contribution the training set TD into $CM_2$

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 we have selected a number of subsets and then we have tested their efficiency and then evaluated their testing and on that basis we have evaluated the errors. Now we have improved the efficiency and

stability of data models and their map reduced web model based on S³VM.there are almost a number of data sets available in the vector subspace .the operation and process we perform will be in tenfold cross validation. The number of iteration that we perform will be seven.

The approach that we apply is rely very much in aspects to other approaches. There are two task we are performing in this approach .first task is that we are going to select whole data set and will perform training on those particular sets. and in the second step by using the binary classification we are going to improve the classification accuracy .this binary classification that we have done here has included deep leaning process. This approach that we have applied is really simple and up to the logic to bring the result. by using the different languages we can implement this approach in an easy way. We can also apply this logic at different levels such as social websites and other level such as to find the human genome ,etc .this all has eliminated the big data problems on a short time note by this process we can also resolve the multi classification problem

## 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 S³VM with traditional BNNC, to execution to gather techniques under comparable conditions. To designated 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 modernisation errors of the two algorithm, 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

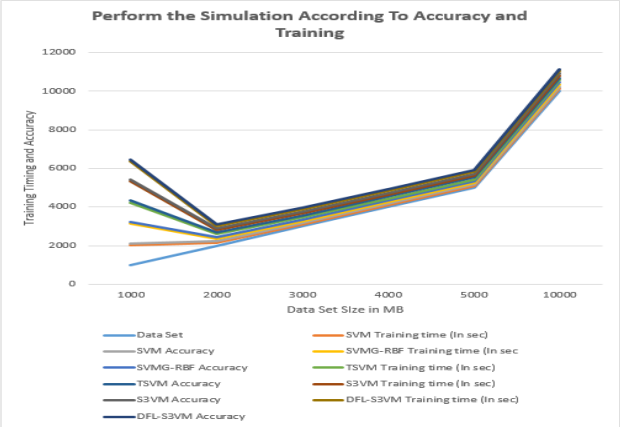


**Figure 2 :** Performance Evaluation different number of algorithm according Accuracy and training time

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

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

Or

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


**Figure 3:** The Classification Result of the Algorithm Proposed

**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

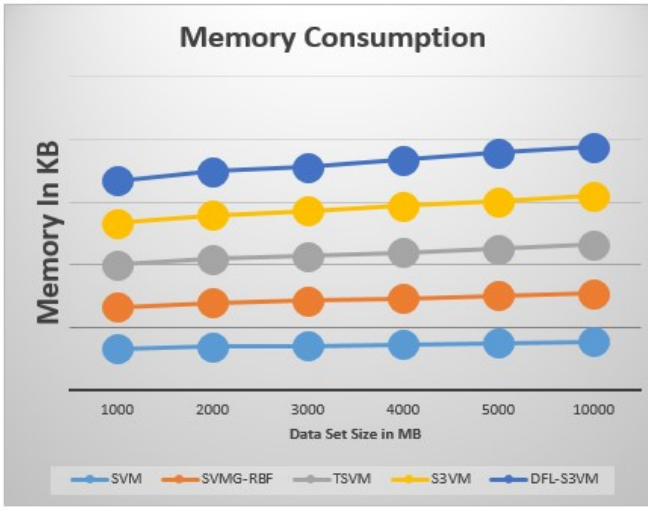


**Figure 4 :** Memory Used Memory Consumption

Using deep learning with binary neural network and genetic algorithm this paper not only increases the precision of the classification model, but correspondingly increases there call rate. The classification effect is enhanced. This is significant for a prediction service to preserve a reasonable response time and predominantly the recommendation calculation to have a limited running time. We should measurement that our proposed algorithm generated the real time results. We have comprehensively tested the effectiveness of our system for growing amount of users, data and complex queries and it returns high speed outcomes.

## 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 users usage keywords that sometime similar and linked so a users interest can be easily determined conclude to use fusion based MR S$^3$VM algorithms based on back propagation neural networks to improve the prediction and recommendation. This all resulted in to the increase in accuracy in big data classification for reduce of training time The results recommend that map reduce based learning model in common should be measured in the future for this task and additional similar tasks. In the near future, to plan to strategy and implement additional machine learning algorithms on the cloud computing platform in mandate to make a more support for the mobile usage information prediction.

## VII. REFERENCES

[1]. Mark Werwath," Implications of Big Data for Data Scientists and Engineers" Ieee Engineering Management Review, vol. 45, no. 3, third quarter, September 2017 IEEE doi 10.1109/emr.2017.2734323.

[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" Fifth International Conference on Advanced Cloud and Big Data"

DOI 10.1109/CBD.2017.30 978-1-5386-1072-5/17.

[3]. Jun-Jie Zhang, Li Yang," A Simple Analysis of Revolution and Innovation of Marketing Mix Theory from Big Data Perspective" IEEE 2nd International Conference on Big Data Analysis 978-1-5090-3619-6/17/2017 IEEE.

[4]. Chengang Zhu, Guang Cheng, and Kun Wang," Big Data Analytics for Program Popularity Prediction in Broadcast TV Industries" DOI 10.1109/ACCESS.2017.2767104, IEEE Access.

[5]. Ni GAO, Ling GAO, QuanliGao, Hai Wang," An Intrusion Detection Model Based on Deep Belief Networks" Second International Conference on Advanced Cloud and Big Data -2014.

[6]. Shi Cheng, Yuhui Shi, Quande Qin, and RuibinBai," Swarm Intelligence in Big Data Analytics" IDEAL 2013, LNCS 8206, pp. 417–426, 2013.

[7]. Yi Wang, Student Member, IEEE, Qixin Chen, Senior Member, IEEE, Chongqing Kang, Senior Member, IEEE, Qing Xia, Senior Member, IEEE and Min Luo ,"sparse and Redundant Representation-Based Smart Meter Data Compression and Pattern Extraction" DOI 10.1109/TPWRS.2016.2604389, IEEE.

[8]. Haojin Yang, Martin Fritzsche, Christian Bartz, ChristophMeinel,"BMXNet: An Open-Source Binary Neural NetworkImplementation Based on MXNet"arXiv:1705.09864v1 27 May 2017Conference'17, Washington, DC, USA.

[9]. Zhang, W., Li, Z., Xu, W., & Zhou, H. (2015). A Classifier of Satellite Signals Based on the Back-Propagation Neural Network, (Cisp), 1353–1357. https://doi.org/10.1109/CISP.2015.7408093

[10]. Zhao, Y., & Bai , S. H. (2012). Research on optimizing recommend system for agriculture information personalization based on user clustering. Proceedings of the 2012 International Conference on Industrial Control and Electronics Engineering, ICICEE 2012, 1477–1480. https://doi.org/10.1109/ICICEE.2012.388

[11]. Yildirim, P. (2017). Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 193–198. https://doi.org/10.1109/COMPSAC.2017.84

[12]. Yin, C., Xiang, J., Zhang, H., & Wang, J. (2016). A New Classificaiton Method for Short Text Based on SLAS and CART. Proceedings - 2015 1st International Conference on Computational Intelligence Theory, Systems and Applications, CCITSA 2015, 133–135. https://doi.org/10.1109/CCITSA.2015.13