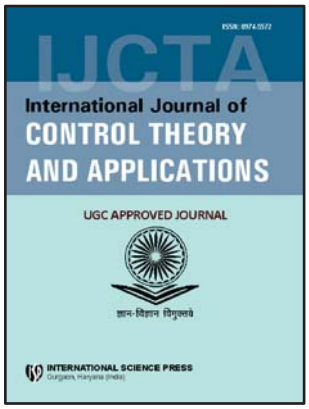

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 31 • 2017

A MapReduce Framework Base Web Sessions Classification using Fuzzy Temporal Classifier

Raghuram Bhukya[a] and Jayadev Gyani[b]

[a]*Department of Computer Science and Engineering, Kakatiya Institute of Technology and Science Warangal, India*
E-mail: raghu9b.naik@gmail.com
[b]*Department of Computer Science and Engineering, Jayamukhi Institute of Technological Sciences Warangal, India*
E-mail: jayadevgyani@yahoo.com

Abstract: Learning from the user click streams and corresponding buying events is indefinitely useful for e-shopping sources for magnifying their business opportunities. Extracting classification models to predict user buying options from such a massive click stream data values and mitigating effect due to huge different timestamps on quality of classifiers are definitely challenging issues for data science community. Applying traditional classification model techniques may lead to soaring time consumption and huge number of unintuitive results. Considering the massive nature of web sessions data set and requirements of efficiency and intuitiveness of a resulting classifier we propose a web session classifier model extraction using fuzzy temporal associative classifier rules using MapReduce framework. Where the MapReduce framework counter the voluminous of data and fuzzy temporal associative classifier take care of efficiency and intuitiveness of results. The experimental evaluation conducted on ACM's Recses click stream dataset reveal that our proposed model is highly competitive in classifying massive datasets with respect to efficiency and intuitiveness.

Keywords: *Web mining; Associative classification; MapReduce; Distributed file systems.*

1. INTRODUCTION

In the current age of digitization the enormous amount of data being digitized in different sectors like social government sector, financial sector, and medical sector and e-services sectors. According survey published by emc.com the yearly data production may exceed to 35 zeta bytes by 2020. The enormous amount of data gathering is because of individual as well societal migration towards digitization is referred as Big-data [1]. In other words Big-data can be referred as collection of data with 4V's characteristics namely, huge Volume, huge Velocity, high Variety and low veracity. The challenges impose by big data includes need of huge data storage resources, efficient processing systems and preserving reliability.

As a response to primary challenge impose by big data that is any single disk can't store such massive data the ever scaling distributed file system is proposed as solution. The distributed file system pave a way for huge data storage and reliability by storing data on parallel disks and by offering common retrieval interface based

on distributed architecture. The Hadoop distributed file system (HDFS) [2] is leading distributed file system offered by Apache as open source, which is adopted by most of the global on-line data service providers. The primary reason for success of HDFS is it can scale on commodity hardware connected with network to store massive amount of ever increasing data. Efficient processing of massive data the other challenge faced by any distributed file system is addressed by HDFS using its adopted MapReduce framework [3].

The MapReduce is the open source computing framework initiated by Google for parallel processing of data stored in distributed file system architecture. The greatness of MapReduce framework is that it can perform its parallel computation on extended number of general computing systems. The MapReduce framework follows a two phase parallel processing approach namely Map and Reduce. Out of which the Map phase is responsible for individual processing of data at local level and Reduce is responsible for collective processing of data at global level. The user is authenticated to decide number of Reducer paradigm to be initiated but the number of Map nodes to be initiated will be decided by framework based on computation load. The all Map and Reduce paradigms organized around input (key, value) value and output (key, value) pairs which makes the easy for developers to implement parallel algorithms on framework. The ease of implementation and efficiency of processing is influencing may field in real time computation including data mining.

Data mining [4] is evolved as one most use full component of data analytics in real time due to its wide spread applications in digital era. Data mining offers pattern recognition as well classification of data which information could be used in decision making and offering targeted services in many fields like service industry, e-commerce, finance, health care, environment and media. The classification technique which is responsible for predicting applicable class label of a new data record using the information in existing data set plays its significant role real time applications of machine learning. So far in the literature we can found several numbers of different types of classification techniques which depends mainly on decision trees, Bayesian, neural network, SVM and rule base. Out different types of classification techniques the associative classification technique [5] plays a vital role by offering high accuracy and good interpretable results to user. These qualities drive the associative classification technique even on newly evolving Big data filed that is extracting association rules driven classifier on Map-reduce framework [6], even the fuzzy version of associative classification which gives intuitive and accurate results is also explored on MapReduce frame work [7]. But no one of these models concentrates on the data generated by World Wide Web which is a primary and most important source of the big data.

The data generated by world wide web includes web contents stored by different websites, web uses data which furnishes data about how users are surfing the different sites and web structures that is how data artifact among the different web pages. In order to extract use full knowledge from different forms of WWW data, there is separate field in data mining that is known with name of web mining [8]. The different field of web mining includes web usage mining, web content mining and web structure mining. In current scenario in order to explore conventional data mining techniques on MapReduce a good amount of work carried out for web content mining with MapReduce framework but a ignorable work carried out with respect to web usage mining on MapReduce framework[9].

Web usage mining mainly includes extracting patterns of web site surfing and predicting user functions by building appropriate classification models from web sessions which is been archived. One of the major challenges in web usage mining is handling huge verity of time steps. For example in crisp handling of time stamps with second are micro second different will be considered as different, so which could result be into huge number of different rules effecting the intuitiveness and efficiency end classification model. In order to counter this worse effect of time stamps, in this work we are proposing usage fuzzy temporal model [10], which can reduce the number of crisp time stamps into a perceptive labels to reduce number of resulting rule set. Considering this lack of web usage mining with MapReduce framework and considering importance of fuzzy associative classification in extracting intuitive and efficient classifiers this work proposes a fuzzy associative classifier extraction for web session extraction using MapReduce framework.

2. REATED WORK

As the distributed file system proved to be efficient in handling massive and ever scaling data sets, the various MapReduce applications are evolved to store and process the big data. For example to store and process data evolved from cloud computing sources Amozon EC2[11] was proposed, to process the stream related data ApacheS4[12] was proposed and there is a Apache Sprak[13] which is an improved version of MapReduce for iterative and fast computing.

Data mining algorithms which are essential part of new generation analytics are getting scaled on to the MapReduce framework. The various types of mining algorithms scaled on MapReduce framework includes & et.al's[14] association rule finind, & et.al's [15] clustering, & et.al's SVM [16], & et.al's boosting [17]. The main objective of these algorithms is to scale the conventional data mining algorithms on MapReduce framework, but the thing they ignored are possibility of dynamic updates to big data set and lack of intuitiveness of results to user. These problems of traditional approaches can be overcome by introducing dynamic associative classification technique for MapReduce framework.

The associative classification technique where classifier extracted from class label based association rules was at first introduced as CBA by Liu & et.al [18] was later upgraded to multi class association rule as MCAR by Li& et.al's [19]. As the associative classification technique proved as efficient with respect to accuracy there are other improvements followed in literature which includes Hu & et.al model [20] for systematic handling for missing values, Yang & et.al [21] model for compact classification and fadi & et.al model [22] for extracting classifier with single scan to database. In other end we can found associative classification method with various applications which includes Ajlouni & et.al proposed association classification model for web mining [23], F.P.Pach & et.al fuzzy intuitive associative classifier [24], Jabbar & et.al [25] associative classification model for image processing, Yoon & et.al [26] associative classification model for text mining, Costa & et.al [27] associative classification model for XML documents, Mokeddem & et.al [28] associative classification model for distributed data sources. Out of these different applications of associative classification the fuzzy associative classification model proved its significance by providing intuitive and efficient classifier for various applications. In literature even we can find fuzzy associative classification approach to handle distributed databases and to handle MapReduce framework [29, 30]. But the downsides of these models are they won't consider the web data sources which are highly reputed source for generating big data.

Web usage mining which extracts user surfing patterns will helps in generating recommendations to web user[31], in order to provide personalized search content to web surfing users [32] and finding suitable customer in targeted campaigning's need to explore in context of MapReduce framework because massive data is gathering as web usage history. At same time web usage mining suffers with huge set of discovered rules due large variety of time stamps, in order to address this problem our proposed model make uses fuzzy temporal associative rules [33] for extracting classification rules. Considering these need of exploring web usage mining on MapReduce framework and significance intuitiveness provided by fuzzy temporal associative classifiers this work proposes the Web Sessions Classification Using Fuzzy temporal Associative Classifier for MapReduce framework.

3. PREPARE YOUR PAPER BEFORE STYLING

The MapReduce architecture adopted for proposed fuzzy temporal rules based web session classifier shown in Fig. 1. According to architecture the web usage data collected from different web sources stored in Hadoop distributed file system (HDFS). The HDFS runs MapReduce paradigm controlled by Name node where the user program for processing the data will be submitted. The Task-tracker function of the Name node will schedule the Map task by initiating by Map nodes according to load of the data to be processed. The data at initiated Map nodes will be processed locally in parallel. Once the Map processing has been done the local intermediate results will be shuffle to Reducer nodes where consolidated processing will carried to produce global result. The number of reducer functions to be created will be decided by the user program logic for efficient transforming of intermediate results into global results.

In order to extract fuzzy temporal rules based web session classifier from the web sessions stored in HDFS our proposed model process the data in following steps:

1. Data preprocessing : Fuzzy temporal sets generation from timestamps of web usage data
2. Fuzzy temporal classification rules generation from pre-processed web usage dataset
3. Prune the resultant fuzzy temporal classification rule set
4. Classify and evaluate the resultant fuzzy temporal classification rule set using test instances

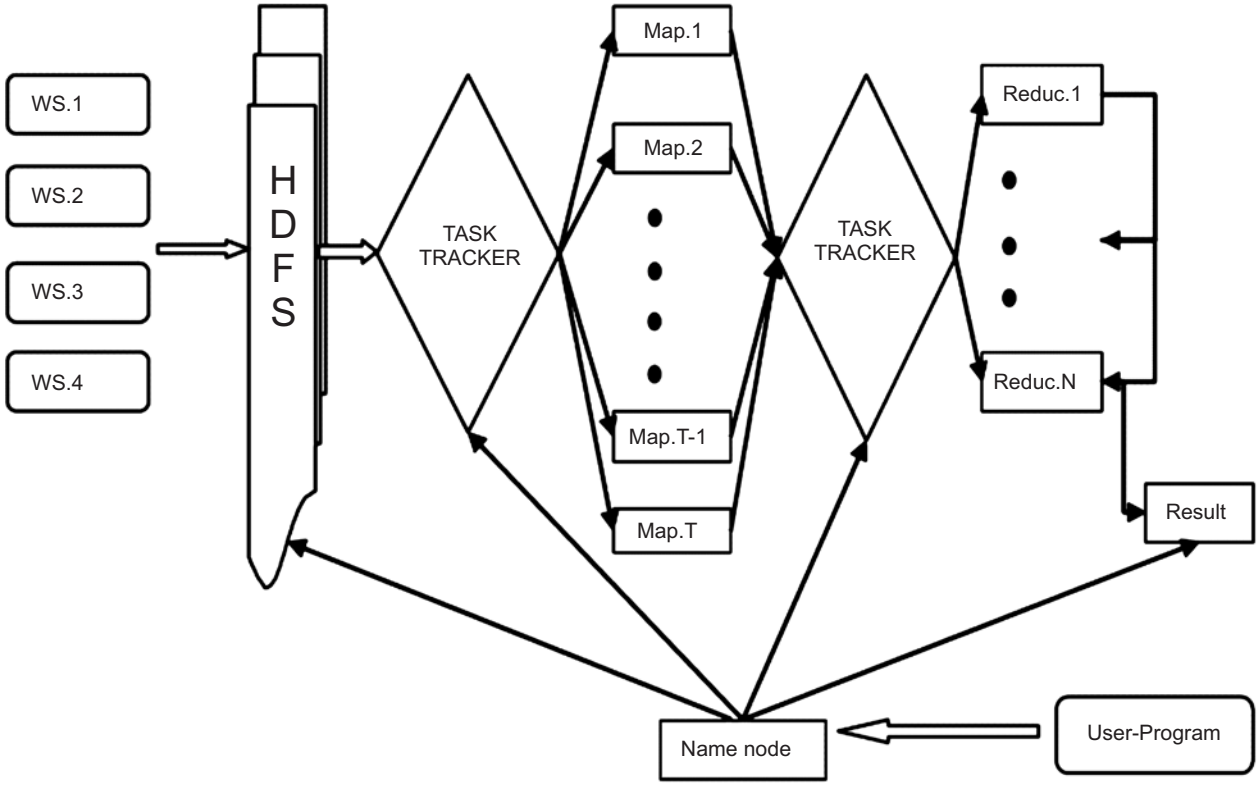

Figure 1: MapReduce architecture adopted for extracting web session classifier

3.1. Fuzzy Temporal Sets base Discritization of Web usage Data

In web session dataset the huge number of different time stamps can affect resulting classifiers in view of accuracy and intuitiveness. That is because of too many different numbers of timestamps will result into huge number of item sets will leads to multiple number of iteration in processing and also cause to produce huge number of rules which are hard to interpret by users. In order to overcome this, the proposed model transforms the time stamps present in web usage dataset into fuzzy temporal sets, which generalizes the set of timestamps into human cognitive label. Due to use of fuzzy temporal sets generalized label, the number of item sets can reduce which result less iterations in further processing and generates less number of easily interpretable rule set.

In generating fuzzy temporal sets from timestamps of web usage data set as the timestamps are huge depending up on experts view to decide fuzzy parameters may lead inaccurate cluster boundaries in order to avoid that we make use of the Map reduce based fuzzy c-means clustering [] model. According to our proposed discritization approach after identifying the temporal attributes from the given web usage data, the

fuzzy c-means clustering algorithm will run on the dataset with respect to temporal attribute to found best boundary values for clustering dataset. In this approach the expert role is reduced to only deciding number of clusters and giving appropriate labels to fuzzy temporal clusters which could greatly improve the efficiency of the cluster process. Once the fuzzy temporal cluster boundaries extraction process over then the using these values the discritization process will be applied to reproduce entire training data set according to fuzzy temporal representation.

3.2. Fuzzy Temporal Classification Rules Generation using Mapreduce Framework

The preprocessed web usage dataset will be stored in HDFS, using which the fuzzy temporal classification rules need to be generate using map reduce paradigm. In order to generate classification rules we propose a parallel associative classification based approach using MapReduce paradigm. According to proposed approach first at Map nodes the Tid-list of allotted sub set of data will produced, which will be further consolidated at Reducer node to extract globally supported fuzzy temporal associative classification rules. The algorithm stated in Table.1 shows the MapReduce procedure for extracting globally supported Tid-list driven class label base association rules. According to the procedure once the data chunks from distributed file system loaded into the respective Map node, it will read transactions and generate corresponding class label base Tid-list representation for corresponding data chunk as shown in Map procedure. Then using final Tid-list generate class label based itemsets and their local support count. The process will be continued in parallel for all data chunk loaded in Map node.

Algorithm 1: Algorithm for Tid-list based globally supported class item set generation.

Input : Data set loaded in HDFS

Output : classifier rules

Map:

1. Read the dataset
2. Tid-list Gen function(): for each record Ri do
    a) Create a new column label in Tid-list for each new combination of attribute value and class label of record Ri
    b) Create a new row in Tid-list representing Ri and store corresponding Tid values
    c) If column label present as combination of attribute value and class label in Ri then store value 1 else store value 0.
3. End Tid-list Gen function ().
4. Generate local class label based item-sets using final tid-list

Reduce :

1. Read all local class label based item-sets of corresponding level
2. Combine all same item-sets to generate global item-set and update support
3. Drop item-sets not satisfying global support threshold
4. Generate class label base association rules form globally supported item-set and corresponding confidence of rules.
5. Drop rules without satisfying confidence threshold.

Once the class label base local item-set generation is over then they set as input to Reducer node to generate globally supported item-sets. The input to reducer node scheduled based upon level of item-set. That is each different level of item-set will allotted to different reducer node. So at each Reducer node same level class label base item-sets generated from different Map nodes will gather, using which Reducer node unifying the all similar class label base item-sets and generate their global support count. Using global support threshold the Reducer node drop the class label base item-sets whose frequency is less than threshold to generated globally frequent class label base item-sets. The globally frequent class label base item-sets will used by reducer nodes for generating globally confident class label based association rules according approach stated in [5].

3.3. Rule Pruning

In order to prune the similar kind of rules and form compact fuzzy temporal classifier from globally confident class label base fuzzy temporal associative classification rules we adopt approach stated in [5] and extend for MapReduce architecture. According to adopted approach at first rule will be divided class wise and sorted according to their confidence from high to low, then if a subset of rule present with lower confidence than its superset rule it will prune. In order to implement same procedure on MapReduce architecture the map function will assign the task of segregating rules according their classes and then sort rules in descending order according to their confidence value. Then each set of class label base association rules allotted to an individual Reducer node. Where in parallel each reduce node checks that if any subset rule is present with less confidence than its super set, if so it will drop the rules. In rule pruning if any super and subset of rules with same confidence found then to take pruning decision their support levels will be considered.

Algorithm 2: Algorithm for classification of test instances

Input : Test instance without class label

Output : Predicted class label of the test instance

Map:

1. For each test instance t_i do
   a) Find all rules that applicable to t_s and place in R
   b) If all the rules in R indicating same class then assign that class label to record
   c) Else calculate sum of firing strengths with respect to each class label
   d) And Assign class label to record with highest firing strength
   e) Check actual class label or record and calculate success and false rate

Reduce:

1. Reducer: calculate final set of success and false rate combining all the results of all the map nodes

3.4. Classifying Test Instances

The MapReduce procedure for classifying test set of web usage data instances and efficiency calculation algorithm is shown in Algorithm 2. Web usage data instance loaded into Map node, in order to calculate efficiency of generated fuzzy temporal associative classifier. At Map node fuzzy temporal associative classifier rule set will be applied on test instance without considering their class record. While classifying a test instance if all applicable rules to that particular record indicates same class label then the same class label will be assigned to that record. Instead if applicable rules indicate different class labels, then the class label of rule set with highest firing strength that is rule set with highest total of confidence will be applied to record. Once the test record classified by rules set, it will compare with actual class label if both are same it will increase success rate count else it will increase false rate count. The process will be repeated at all Map nodes in parallel. In the next step all the Map nodes success and failure count will forward to a single reducer node where final associative classifier accuracy will be calculated.

4. IMPLIMENTATION AND EVALUATION

The realization of the proposed fuzzy temporal classification model carried out using the MapReduce framework configured with one name node acting as server and 6 other systems with Pentium-*i*3 processors, 2GB RAM are acting as computing nodes. These systems are interconnected with 1GBPS Ethernet and supported by 1Terabyte external hard-disk. The systems are working with Ubantu 14.4 version operating system and Hadoop2.0.0-cdh4.4.0 version is installed for providing MapReduce framework. The experiment is conducted in three phase to evaluate accuracy, scalability and dynamism of proposed model.

The evaluation process was conducted on ACM's RecSys Challenge 2015 datasets [35] which is open sources classification data set consist of 2 files yoochoose-clicks data and Yoochoose-buys data with more than 5 milion data records. The Yoochoose-clicks file stores the visitor's data of an *e*-shopping website with attributes Session_id, Timestamp, Item_id and Category whereas Yoochoose-buys file stores the visitors buying information of same websites including attributes Session_id, Timestamp, Item_id, Price and Quantiy. Both databases can be linked using common session ids. So the classification task is to use the data set in building the prediction model which should learn the classification rules to predict which of the sessions will end with buying of item.

In order to implement the proposed model at first level the data in both files are loaded to HDFS and using the time stamps attribute of the data set fuzzy clusters centroid's and boundaries extracted by applying fuzzy c-means. The data from both the sets are integrated by their transaction ids. The obtained clustering centroid's used for data discritized process to transform dataset into fuzzy temporal dataset. Then the using fuzzy temporal dataset the fuzzy temporal classification rules extracted for predicting the web session which is going to buy or not by applying proposed algorithms for classification rule extraction and rule pruning for MapReduce framework. Once the classifier rule extraction is over on training dataset then the same rules applied on testing data to evaluate the accuracy. As there is no similar model found in literature proposed for web session classification with MapReduce framework, we implemented distributed CBA which is Map Reduce version of CBA on web usage session data to carry out comparative evaluation. In comparative evaluation of the proposed Fuzzy temporal model has given 79% accuracy which is comparatively acceptable with respect to the 81% accuracy given by distributed CBA. In case time efficacy the proposed Fuzzy temporal model generate final set of rules with in 346seconds which is significantly well in compare to distributed CBA which took 1308seconds to generate final set of classification rules. The other advantage of the proposed model is it extracted only few easy interpretable rule set.

5. CONCLUSION

As the trends of e-shopping enormously increasing worldwide, a massive amount of click stream data has been gathering from e-shopping web sites and portals which could be easily considered as big data analytical challenge. Handling timestamps which is attached with each session is could be the major challenge in producing intuitive and effective classification rules for predicting user buying patterns in stipulated time period. These challenges overcome in our proposed model by generating fuzzy temporal classification rules for web session classifier on MapReduce framework. The proposed model experimentally evaluated on ACM's Recses15 click strem dataset shown that the proposed model efficient enough to produce efficient and intuitive classification rules over web session's big data set.

REFERENCES

[1] C.L. Philip Chen , Chun-Yang Zhang, .Data-intensive applications, challenges, techniques and technologies: A survey on Big Data Information Sciences 275 (2014) 314–347

[2] Apache Hadoop Project, http://hadoop.apache.org/

[3] MapReduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawa, Google Labs, pp. 137–150, OSDI 2004

[4] C.W.Tsai, C.F.Lai, M.C.Chiag, L.T. Yang " Data mining for internet of the things" IEEE communications Survey & Tutorials, 2014, Vol.16.Issue.1, 77-97.

[5] Neda Abdelhamid, "Associative Classification Approaches: Review and Comparison", Journal of Information & Knowledge Management, Vol. 13, No. 3 (2014) World Scienti¯c Publishing Co.

[6] Alessio Bechini, Fransceco Marcelloin, Armado Segatori, " A MapReduce-based Fuzzy Associative Classifier for Big Data" Information Sciences Volume 332, 1 March 2016, Pages 33–55

[7] Pietro Ducange, Fransceco Marcelloin, Armado Segatori, " A MapReduce-based Fuzzy Associative Classifier for Big Data ".2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Pages 2-5 Aug. 2015, DOI: 10.1109/FUZZ-IEEE.2015.7337868.

[8] QINGYU ZHANGRICHARD, S. SEGALL, Int. J. Info. Tech. Dec. Mak. 07, 683 (2008). WEB MINING: A SURVEY OF CURRENT RESEARCH, TECHNIQUES, AND SOFTWARE, DOI: http://dx.doi.org/10.1142/S0219622008003150.

[9] Sheela Gole; Bharat Tidke A survey of big data in social media using data mining techniques 2015 International Conference on Advanced Computing and Communication Systems IEEE Conference Publications, Pages:1-6, DOI: 10.1109/ICACCS.2015.7324059.

[10] Zadeh, L. A.: Fuzzy sets. Inf. Control, 8, 338–358 (1965).

[11] http://aws.amazon.com/ec2/.

[12] L. Neumeyer, B. Robbins, A.Nair, A. Kesari, S4: distributed stream computing platform, in: Proceedings of 2010 IEEE International Conference on Data Mining Workshops (ICDMW),, 2010,pp.170–177, doi:10.1109/ICDMW.2010.172

[13] M. Zaharia, M. Chowdhury, M.J. Franklin, S.Shenker, I.Stoica, Spark: Cluster computing with working sets, in: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, Hot Cloud'10,USENIX Association, Berkeley,CA,USA,2010.10–10

[14] M.-Y.Lin,P.-Y.Lee,S.-C.Hsueh, Apriori-based frequent itemset mining algorithms on MapReduce, in: Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, ICUIMC'12, ACM, NewYork, NY, USA, 2012, pp.76:1–76:8.

[15] W.Zhao,H.Ma,Q.He, Parallel KMeans clustering based on MapReduce, in : Lecture Notes in Computer Science, 5931, Springer Berlin Heidelberg, 2009, pp.674–679.

[16] Q. He,C.Du,Q. Wang,F.Zhuang, Z.Shi,A parallel incremental extreme svm classifier,Neurocomputing74(16)(2011)2532–2540.

[17] I. Palit, C.Reddy, Scalable and parallel boosting with MapReduce ,IEEETrans. Knowl.Data Eng.24(10)(2012)1904–1916.

[18] B.Liu, W..Hsu& Y.Ma." Integrating classification and association rulemining" Jouranl of Knowledge discovery and data mining, KDD-98 , 1998.

[19] W.Li, J. Han & J.Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules ", International conference on data mining, pp. 369–376, 2001.

[20] H.Hu, J.Li. "Using association rules to make rule-based classifiers robust". Proceedings of the sixteenth Australasian database conference pp. 47–54, 2005.

[21] J.Yang, " Classification by association rules: The importance of minimal rule sets", The twentieth international conference on machine learning, 2003

[22] F.Thabtah, P.Cowling and S.Hammoud, "MCAR: multi-class classification based on association rules", Proceedings of the ACS/IEEE 2005 International Conference on Computer Systems and Applications, Washington, DC, USA, pp. 33–38, 2005.

[23] Ajlouni, M. I. A., Hadi, W., & Alwedyan, J. (2013). Detecting phishing websites using associative classification. European Journal of Business and Management, 5, 36–40.

[24] F.P.Pach, A.Gyenesei, J.Abonyi, Compact fuzzy association rule-based classifier, Expert Systems with Applications, Vol.34, 2008, pp.2406–2416

[25] Jabbar, M., Deekshatulu, B., & Chandra, P. (2013). Knowledge discovery using associative classification for heart disease prediction. In Intelligent informatics (pp. 29–39). Springer.

[26] Yoon, Y., & Lee, G. G. (2013). Two scalable algorithms for associative text classification. Information Processing and Management, 49, 484–496.

[27] Costa, G., Ortale, R., & Ritacco, E. (2013). X-class: Associative classification of XML documents by structure. ACM Transactions on Information Systems, 31, 3.

[28] Mokeddem D, Belbachir H.2010. Distributed classification using class association rules mining algorithm. IEEE International conference on Machine and web intelligence, Algeria

[29] Raghuram Bhukya and Jayadev Gyani, "Fuzzy generalised classifier for distributed knowledge discovery", International Journal of Business Intelligence and Data Mining, Vol.8, No.3, pp.227 – 243, 2013

[30] Raghuram Bhukya and Jayadev Gyani, "Fuzzy associative classification algorithm based on MapReduce framework," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere,2015,pp.357-360, doi: 10.1109/ICATCCT.2015.7456909.

[31] Rana Forsati, Mohammad Reza Meybodi, Afsaneh Rahbar, An efficient algorithm for web recommendation systems IEEE/ ACS International Conference on Computer Systems and Applications, 2009. AICCSA 2009.

[32] Nicolaas Matthijs, Filip Radlinski, Personalizing web search using long term browsing history Proceeding WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining Pages 25-34

[33] Lee WJ, Lee SJ, Discovery of fuzzy temporal association rules, IEEE Trans Syst Man Cybern B Cybern. 2004 Dec;34 (6):2330-42.

[34] Mapreduce base fuzzy c-means Springer

[35] ACM RecSys Challenge 2015

[36] W. Peizhuang Pattern recognition with fuzzy objective function algorithms (James C.Bezdek), SIAM Review, 1983.