Clonal selection-based AIS weighted feature extraction algorithm to identify the multiclass web pages categories

Karunendra Verma, Prateek Srivastava, Amit Jain

Abstract**: *Due to the unbelievable increment in the assess of data on the World Wide Web, there is a solid requirement to optimize a web page cataloging to reclaim constructive information rapidly. Proposed clonal selection based artificial immune system algorithm to select the most excellent weights for every feature in the training dataset and implement the KNN (k-Nearest Neighbour) classifier to categorized the new web pages from testing dataset. In addition, the weight determination process is depended on both term and tag weighting method. Structure features are gathered and appointed weights in this scheme. Results obtained show that projected classifier effectively classified to demonstrate the efficacy of the algorithm with respect to single and multi-class.***

Index Terms**: *Artificial immune system, k-Nearest Neighbour, Tag weighting, Term weighting, Web page classification.***

I. INTRODUCTION

Internet (WWW) information is as site pages and sites made out of pages. Every day billions of website pages are added to the tremendous store of electronic substance as site pages. These site pages contain data about nearly everything [1]-[2]. Grouping of web substance can decrease the endeavours to numerous folds. Website page characterization is a rational task and it requires human interference [3]. Grouped website pages can be utilized for quick and simple recovery of the stuffing from the database store. It additionally helps in focused crawling, creating question-answer sessions, directed inquiry and collecting domain knowledge [4]- [5]. Web page cataloging is the procedure of allocating a web page to one or extra predefined category. Web page categorization is measured a tough problem because of massive and exponentially growing size of WWW [6]. Categorization of web also helps in humanizing the quality of seek out results. Alongside this, imagine of semantic web, categorization of web records open doors for some different applications. Website pages are created using html tags and presentation and layout out information is quite different from the simple text presentation [7]. Conventional classification techniques are not sufficient for web categorization. Novel machine learning schemes and algorithms are needed for web document categorization. Indeed, even the features required for characterization reason for existing are not the same as basic content report. Numerous examination strategies are being created to deliver practical and computational more affordable solution [8]. Classification procedure can be of two kinds in view of the number of classes. Vault of site page can either be isolated into both of the two expansive classes or one of the numerous predefined classes. Quantities of methodologies are created for both the procedures. Classification plans can empower the distribution, interlink and reuse of immense data as site pages and valuable datasets. The general issue of Web page classification can be additionally reached out to more particular issues [9]-[10]. Classification demonstrates a critical part in different data recovery tasks. The web is exceptionally dissimilar in nature, and no guidelines are present on the best way to assemble HTML pages and how to express the whole structure of the pages. Along these website pages classification is an imperative task. Website page classification procedure utilizes an assortment of data to characterize target pages [13]. The essential thought for website page cataloging is the resemblance estimation between web documents. Current website page grouping techniques utilize an arrangement of information to characterize a site page similar to the content of the page, structural data of the site page and the URL of the target page. In this way web page content, formation and URL are minimum costly to accomplish and considerable sources for cataloging. Web page classification procedures utilize different data to characterize a target site page: the content of the site page, website page URL and formation data on a site page [14].

II. PROBLEM DEFINITION

There has been an exponential increment of data accessible on the World Wide Web (WWW), and thusly discovering pages that present data to fulfill necessities is extremely troublesome. If we need to look effectively and rapidly with a web crawler, we need a proficient strategy for characterizing site pages. In this paper, we talk about a structure based arrangement way to deal with classify website pages into one of two classes (e.g., "sport" or "not sport").

Manuscript published on 30 March 2019.
***Correspondence Author(s)**

Dr. Amit Jain, Assistant Professor in Computer Science and Engineering Department, Sir PadampatSinghania University, Udaipur, India.
Prateek Srivastava B.Tech and M. Tech degree from Uttar Pradesh Technical University. He received PhD degree in computer Science and Engineering from Sir Padampat Singhania University,Rajasthan. India.
Karunendra Verma Department of CSE, Sir Padampat Singhania University, Udaipur ,Rajasthan, India.

We embrace HTML labels and terms as characterization includes and use the clonal determination based AIS calculation to decide ideal loads of each element in the preparation dataset. Consequently, our proposed novel AIS calculation to take care of the issue of ideal element determination and defeat the downsides prior highlights choice procedures. After the feature selection phase, the weights are then used to classify new web pages. The design of the proposed framework is shown in Fig.1.

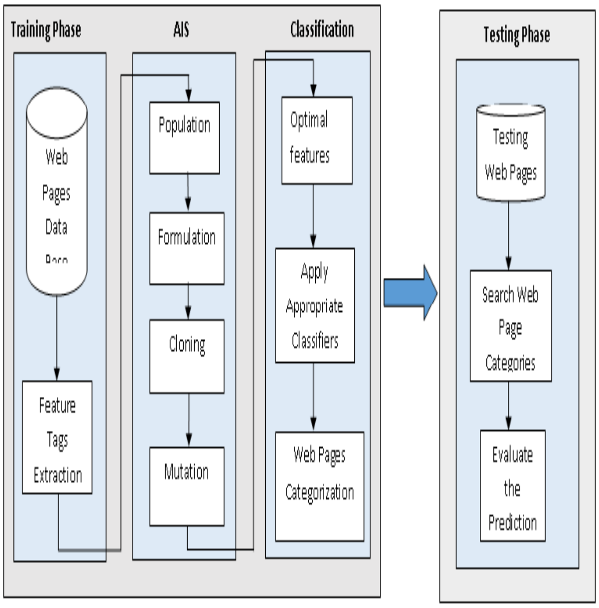

Fig.1 Proposed System Architecture

III. WEB PAGES DATABASE

Web pages are collected from Bank Search dataset [11], is mainly deliberate to help a widespread range of web document categorization . The dataset contains two thousand two web documents arranged into 10 similarly estimated groups like A: Commercial Banking and investment, B: Building societies and investment, C: Insurance agencies and investment, D: Java language pages, E: C/Cpp language pages, F: VB language pages, G: Biology Science, H: Astronomy knowledge, I: Soccer Sport, J: Motor etc [12]. and each contain two hundred web archives. The investigations include order of two classifications from the dataset. Since the dataset does not have test set, the initial 140 reports are utilized as preparing set and the last 60 archives are utilized as test set. A few classifications are very comparative (for example classification A: Commercial Banks and B: Building Societies), while a few classes are very different (for example classification J: Motor Sport and A: Commercial Banks).

A. Term weighting schemes

In web classification systems, term-weighting techniques like Term Frequency (TF), Document Frequency (DF), and Inverse Document Frequency (IDF) are commonly used [16].

B. Term frequency (TF)

Term frequency is identified to the weight of an assured term in the web page. Here, the normalized TF is considered. The weight of the term ‘x’ in website page ‘p’ (signified as TFxp) is controlled by finding the quotient of the raw frequency (occurrences) of the term ‘x’ in the website page ‘p’ and the Euclidean mean of the website page. Where the Euclidean mean of the website page is defined as a square root of the summation of the square of frequencies of all terms in the website page.

C. Document frequency (DF)

DF is mass of a term‘x’ (signified by DFx) denotes the collection of web pages in which the term x is found.

D. Term frequency–inverse document frequency (TF–IDF)

TF–IDF [12] is a valuable standing computing method that denotes the terms inside the web pages database and reflect the statement that a less frequent word in the group is a more noteworthy term in the web page and vice versa.

TF-IDF of a term ‘x’ in document d, denoted by (TF-IDF xp), is the multiplication of the term frequency (TF) and the inverse document frequency (IDF) of the word.

$$\text{TF–IDF}_{x,p} = \text{TF}_{x,p} * \text{IDF}_{x} \qquad (1)$$

$$TermScoreij = \frac{nij}{\sum_{k} nkj} \times \log \frac{|D|}{|\{d : xi \in d\}|} \qquad (2)$$

Where

n_{ij} is no. of frequency of word x_i in document d_j ;
n_{kj} is the sum of occurrences of all term in document d_j ;
D is total documents;
d is no. of documents which included word x_i.

E. Structure weighting scheme

A Structure-arranged Weighting Technique (SWT) can be utilized to weight the critical features in site pages. The possibility of SWT is to allot more prominent weights to terms that identified with the components that are progressively fitting for demonstrating web pages, (for example, terms encased in TITLE labels). SWT is characterized by the capacity as pursues

a. Tag Frequency [12]

$$tfx(x,d) = \sum_{i \in z} \left[tfx(xi,d) \times \frac{ai}{\sqrt{\sum_{j=0}^{k} ai^{2}}} \right] \qquad (3)$$

Where

tf_x (x,d) is the tag frequency of term x in document d ;
tf_x (x_i,d) is the tag frequency of the term x in tag i ;
a_i is the tag weighting cofficient and i ϵ z and z is the set of tags .

*Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)*

b. Tag weight [12]

$$Wx(x,d) = \frac{tfx(x,d) \times \log(N/nx + 0.01)}{\sqrt{\sum_{x \in d} \left[tf\, x(t,d) \times \log(N/n\, x + 0.01) \right]^2}} \qquad (4)$$

Where

W_x (x,d) is the feature tag weighting of term x in document d ;

tf_x (x,d) is the frequency of the word x in document d ;

N is the total number of documents;

n_x is the number of documents which included term x.

F. Feature Selection

Proposed novel Clonal selection-AIS (artificial immune system) algorithm, solve the web page's feature selection problems. To enhance the effectiveness of AIS algorithm, Clonal selection incorporated to explore local search ability.

IV. ARTIFICIAL IMMUNE SYSTEM ALGORITHM

Artificial Immune System (AIS) [15] which has the characteristic of high self-adjustment and self-development roused from function of biological immune system. It contains the capacity of learning, recognizing, memorizing and characteristic extraction. AIS are being utilized as a part of numerous applications, for example fault detection, adaptive control, data mining, anomaly detection, computer security and pattern recognition. Biological inspired AIS models are as follows: Danger hypothesis, Clonal choice, Immune systems, Negative determination. Among these, clonal choice together with the proclivity development forms has been connected to clarify how the invulnerable framework reacts to the bacterium, and how it enhances its ability of executing the attackers. A solution for optimal feature selection can be expressed by the assignment of evaluating term and tag scores for each web page. In the present paper, both term and tag scores used to represent the antibody (solution). The count of an antibody is equal to the total number of the website pages.

A. Population Initialization

Proposed AIS produce determined sub-antibodies by the arbitrary task of the feature values. The created sub-antibodies are arranged to produce the initial population of the antibodies.

B. Formulation

Formulation is a process of finding fitness value to minimize the population. Fitness evaluation is performed based on features; every antibody (solution) has an affinity value (fitness). The affinity value of the antibodies is resolute by neglecting the antibody having zero value.

C. Cloning

The cloning process is implemented to take each formulated antibody values as an input, apply mean and variance to create a clone (new posterity) for every antibody. The quantity of clones is controlled by the quantity of antibodies and the proclivity estimation of the antibody.

D. Mutation Process

Mutation steps are done by select non negative values from cloning. The mutated antibody replace its original despite of its fitness value.

E. Clonal Selection based AIS Algorithm Steps

(1) Initializing: arbitrarily create a primary population of antibodies (ab).

Each ab in Ab population is characterize by term score and tag score.

(2) Formulation: calculate each ab in Ab population according to fitness value.

(3) Fitness evaluation based on selection of non negative value

Do: {cloning, mutation}

(4) Cloning: for every ab є Ab, generate a set of clones (clone_Ab) by appling the mean value and variance values.

(5) Mutation: mutate every non negative clone and add mutated population to Ab.

(6) Steps 2 to 5 are repeated until a pre-defined stopping condition is reached.

Fig.4. Shows the result after applying above mentioned algorithm on extracted feature tags weighting values.

V. K-NEAREST NEIGHBOUR (KNN) CLASSIFIER

The K-Nearest Neighbour classifier [17] typically applies either the Euclidean distance or the cosine resemblance among the training set and the test set. In this work, the Euclidean distance implementing the K-NN model for web page categorization scheme. The Euclidean distance between a training page and a test page can be derived as follows:

```
Classify (X,Y,z)
//X: Training data, Y :category labels of  X, z :unknown sample
    for i=1 to m do
        Compute distance  d(Xi,z)
    end for
  Calculate set I contain index for the k smallest  distance d(Xi,z).
```

VI. EXPERIMENTAL RESULTS

To calculate the performance of this algorithm, a computational experiment is conducted and the results are presented as follows:

A. System Evaluation Metrics [16]

To evaluate the system performance following below mentioned parameters are measured.

B. Accuracy

The accuracy referred the ratio of the true positive and true negative to the sum of true positive, true negative and false negative. Thus, the accuracy equation is:

*Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)*

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (5)$$

.

C. Precision

It is defined as the ability of the evaluation to find the true positive rate in the validation. Precision is numerically represented as follows.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

D. Specificity

Specificity refers to as one of the corresponding parameter which measures the probability that the result obtained from the evaluation is negative when error is not found. The following specifies the mathematical evaluation of the parameter.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (7)$$

From the equation (7), the specificity is being calculated as ratio between the true positive values to the addition of the false positive to that of the true negative value.

E. Matthews correlation coefficient (MCC) & F measure:

It's the quality of binary (two-class) classifications. It is an association coefficient between the experiential and predicted binary classifications. The following represents the numerical evaluation of the parameter.

$$MCC = \frac{TP * TN - FP * FN}{Root((TP + FP)(TP + FN)(TN + FP)(TN + FN))} \quad (8)$$

"Harmonic mean of precision and recall is called F measure or F1or F score".

$$\text{F - measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

(9)

Figure 2 shows the tag frequency graph of training data set in which image, title, html, table and head tags are important.

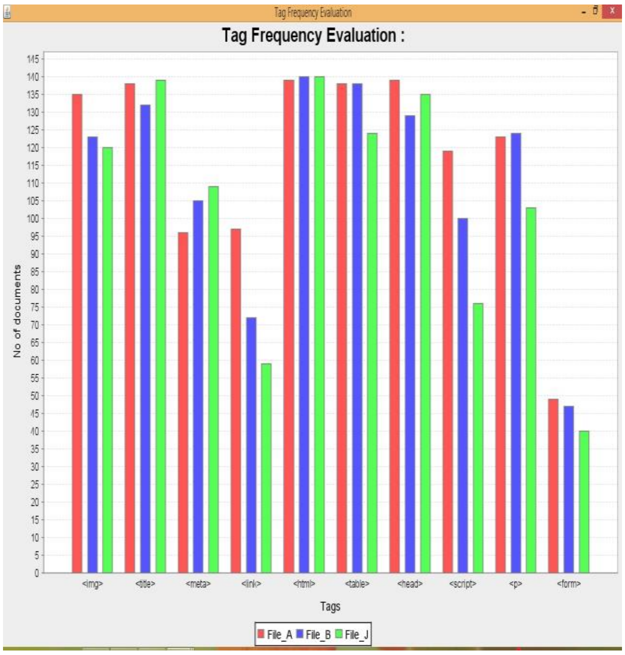

Fig.2 Training set Tag frequency graph

Figure 3 is showing the tag frequency graph of testing data set in which image, title, html, table and head tags are important.

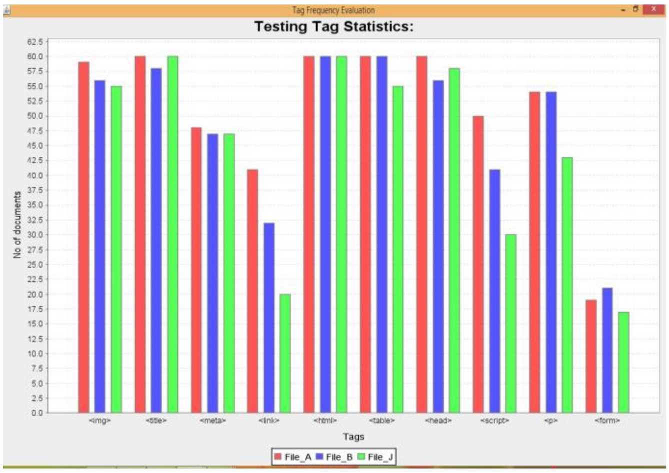

Fig.3 Testing set Tag frequency graph

After finding the important tags and evaluating the term weight and tag weight, Applied the AIS algorithm to extract the optimal features. Results are shown in figure 4 and table 1.

AIS Starting :

^^^^^^^^^^^^^

=========CatagoryA=========

Mutuated values :

TermScore TagScore

TotalFile Size : 140 MutuatedFile Size : 56

=========CatagoryB=========

Mutuated values :

TermScore TagScore

TotalFile Size : 140 MutuatedFile Size : 43

=========CatagoryJ=========

Mutuated values :

TermScore TagScore

TotalFile Size : 140 MutuatedFile Size : 29

End of AIS Part...

Fig. 4 AIS based results

S.NO.	DOCUMENT CATEGORY	POPULATION SIZE	TERM & TAG BASED OPTIMAL FEATURES (AFTER MUTUATION)
1	A	140	56
2	B	140	43
3	J	140	29

Table 1. Optimal Features extraction values

Cataloging of category A and B, considering both terms & tags				
	A	B	Existing SVM System Accuracy (%)	Proposed AIS-KNN System Accuracy (%)
A	58/60	2/60	93.90	95.83
B	3/60	57/60		

Table 2. Term & Tag based accuracy comparisons in between A & B Category

F. Classification using three Categories

For various and optimal value of k Applied the KNN classifier on Retrieved the optimal features using AIS from training set and retrieved the value of term and tag score from testing data set, categorized the web page class.

Results are shown in table 2 and table 3. From table 2, AIS-KNN algorithm is given 95.83% accuracy to categorize the A & B web page class which is improved from existing SVM classifier. From table 3, AIS-KNN algorithm is given 98.33% accuracy to categorize the A & J web page class which is improved from existing SVM classifier.

Table 3. Term & Tag based accuracy comparisons in between A & J Cateory

Cataloging of category A and J, considering both terms & tags				
	A	J	Existing SVM System Accuracy (%)	Proposed System Accuracy (%)
A	59/60	1/60	97.70	98.33
J	1/60	59/60		

Figure 5 shows the accuracy comparison graph between SVM and AIS-KNN algorithm based on Term & Tag weight.

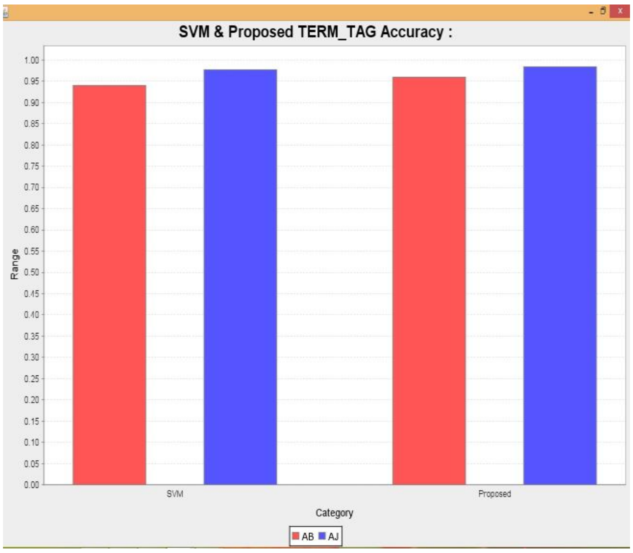

Fig.5 Term & Tag based accuracy comparison graph

G. System Performance analysis

Measures/Category	Category (A& B)	Category (A&J)
Sensitivity	0.9583333333333333	0.9833333333333333
Precision	0.9584606835232008	0.9833333333333333
Specificity	0.6696428571428572	0.6666666666666666
Fmeasure	0.9583304396138621	0.9833333333333333

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

MCC	0.6291966520252016	0.6444444444444445
Accuracy	0.9583333333333333	0.9833333333333333

Table 4. Various System performance measures with KNN classifier

Table 4 shows that various systems measures performance analysis among the various web page categories.

It shows that how the proposed scheme performs on multiclass web page categories concerning different chosen parameters. Kind of classifiers, include set choice, preparing informational collections prerequisites, figuring expenses, and execution are the key parameters which are chosen for performing a general correlation of the proposed plan regarding other existing plans.

VII. Conclusion

With a boost in the user's need on internet, response time and accuracy are the main concern. To overcome these issues, novel clonal selection based AIS method for website page cataloging using weighted features to identify multiclass has proposed. Compared to SVM web document cataloging method, combine the full text with structure information achieve nearly 1.93% accuracy improvement in the case of similar categories and 0.63% accuracy improvement in the case of different categories. The execution of the projected plan is assessed utilizing diverse evaluation metrics where its execution has discovered acceptable regarding the chosen parameters. Later on, we will investigate different features for multiclass classifier for efficient website page classification. We can also compare these results with other classification algorithm by applying similar categories and different categories web pages as in input.

REFERENCES

1. A .Matthew.,Web 2.0, An argument against convergence. In Media Convergence and De-convergence, Palgrave Macmillan, Cham, (2017),pp. 177-196.
2. Rogers, Richard, and Noortje M.,Landscaping climate change: A mapping technique for understanding science and technology debates on the World Wide Web. Public Understanding of Science (2016).
3. Bhalla, K .Vinod, and K. Neeraj, An Efficient Multiclass Classifier Using On-Page Positive Personality Features for Web Page Classification for the Next Generation Wireless Communication Networks. Wireless Personal Communications 93, no. 2, (2017),pp. 503-522.
4. Gani, Abdullah, A.Siddiqa , S. Shahaboddin, and H. Fariza ,A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowledge and information systems 46, no. 2, (2016), pp. 241-284.
5. Pérez, Serge, A. Sarkar, R.Alain, D.Sophie, B. Christelle, and A. Imberty., Glyco3D, A Suite of Interlinked Databases of 3D Structures of Complex Carbohydrates, Lectins, Antibodies, and Glycosyltransferases. In A Practical Guide to Using Glycomics Databases, Springer, Tokyo, (2017),pp. 133-161.
6. Malhotra, Ruchika, and A. Sharma, Quantitative evaluation of web metrics for automatic genre classification of web pages. International Journal of System Assurance Engineering and Management 8, no. 2, (2017), pp. 1567-1579.
7. Khalil, Salim, and F . Mohamed, RCrawler, An R package for parallel web crawling and scraping. SoftwareX 6 , (2017), pp. 98-106.
8. Li, Huakang, Zheng X., Tao L., Guozi S., and R. C Kim-Kwang.,An optimized approach for massive web page classification using entity similarity based on semantic network. Future Generation Computer Systems 76, (2017), pp.510-518.
9. Khatami, Reza, G. Mountrakis, and V. S. Stephen,A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. Remote Sensing of Environment 177, (2016), pp. 89-100.
10. Bader, Sebastian, and O. Jan, Semantic Annotation of Heterogeneous Data Sources, Towards an Integrated Information Framework for Service Technicians. In Proceedings of the 13th International Conference on Semantic Systems, ACM , (2017), pp. 73-80.
11. M.P.Sinka and D.W.Corne BankSearch dataset. Retrieved from http://www.pedal.reading.ac.uk/bansearchdataset Accessed January 15, 2005.
12. K. Verma, P. Srivastava, P. Chakrabarti, Exploring structure oriented feature tag weighting algorithm for web documents identification, Soft computing system. kollam, India : Springer, 2018; pp. 169-180.
13. Fielding, T. Roy, R. N. Taylor, J. R. Erenkrantz, M. G. Michael, W. Jim, R. Khare, and O. Peyman, Reflections on the REST architectural style and principled design of the modern web
14. architecture (impact paper award). In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ACM, (2017),pp. 4-14.
15. Arya, Chandrakala, and S.K. Dwivedi, News web page classification using URL content and structure attributes. In Next Generation Computing Technologies (NGCT), 2nd International Conference ,IEEE, 2016,pp. 317-322.
16. Castro D.,and Timmis L.N..: Artificial Immune Systems: A New Computational Intelligent Approach. Springer, Berlin (2002).
17. M. Hossin and M.N. Sulaiman, A Review On Evaluation Metrics For Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2,(2015).
18. B. I Sadegh., and B.Mohammad, Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events, Theoretical Background. International Journal of Engineering Research and Applications Vol. 3, Issue 5, (2013), pp. 605-610.

AUTHORS PROFILE

Karunendra Verma is a research scholar from Department of CSE, Sir Padampat Singhania University, Udaipur ,Rajasthan, he is having 10+ years of teaching experience in the field of computer science & engineering. His area of research is data mining and information retrieval. He obtained his Master of Engineering from University of Pune in 2008 and his Bachelor of Engineering from RGPV,Bhopal in 2004 in computer Science and Engineering branch. He has published many International and national level papers in journals and conferences. He qualified GATE, UGC NET and Microsoft MTA certification. He Received Active Participation Award (Youth) 2016-17, Region III from Computer Society of India, Kolkata to actively organized CSI activities

Prateek Srivastava received the B.Tech and M. Tech degree from Uttar Pradesh Technical University. He received PhD degree in computer Science and Engineering from Sir Padampat Singhania University,Rajasthan. He had been worked as an Assistant professor in computer science department at Hindustan College of Science and Technology from 2005 to 2011.Presently he is associated with Sir Padampat Singhania University. His research interests includes System modelling, refinement of distributed systems, verification and reasoning of critical properties using formal techniques.

Dr. Amit Jain is presently working as Assistant Professor in Computer Science and Engineering Department, Sir PadampatSinghania University, Udaipur, India. He has completed his Doctoral Degree in Computer Engineering in year 2016. He is having teaching and administrative experience of 22 years of teaching experience. He has taught to post-graduate and graduate students of engineering. He has about 30 research publications in International Journals and Conferences. He holds the post of Associate Editor in many International Research Journal. He is reviewer in IEEE,Inderscience and Elsevier and Scopus Indexed Journals.