IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-2, Issue-4 E-ISSN: 2347-2693

An Implementation of Hybrid Genetic Algorithm for Clustering based Data for Web Recommendation System

Animesh Shrivastava[1*] and Singh Rajawat[2]

[1*,2]*Department of CSE, Shri Vaishnav Institute of Technology & Science, Indore, India*

www.ijcaonline.org

Received: 23/03/2014 Revised: 8/04/2014 Accepted: 24/04/2014 Published: 30/04/2014

Abstract - Web Mining is an interesting domain in information processing that includes a large variety of applications i.e. recommendation system design, next user web page prediction, navigational pattern analysis and others. In this paper a new hybrid clustering algorithm is proposed and implemented using Genetic algorithm and K-NN algorithm and the implementation of desired algorithm is given using a web recommendation system which analyze user navigational pattern from web server access log file and recommends the next user web page. The performance of the designed system is evaluated and listed in this paper. According to the results, the proposed hybrid approach is efficient and effective for the given application domain.

***Index Term*—**Recommendation Systems; k-NN; Genetic Algorithm; Clustering

I. INTRODUCTION

Web is a rich source of information and knowledge, and this data source is scalable. However, the growth of web data found in unstructured manner. Web data is found in three different places, over server access log, in content of web pages, and in organization of web pages. The Web data is the collection of data related to the web addresses, time of the request and response, type of the operating system and browser, agents used, status codes and methods used as get or post. As the web log contains variety of information, which may include noisy data, redundant data and some unuseful information. This is the huge collection of Web data[1].Therefore, there must be some technique, which retrieves useful information in short span of time. In order to overcome this problem, the data mining techniques are available which extract useful information from the huge web log. Data mining techniques like association rules, sequential pattern mining, classification and clustering. In the past, classification techniques were used in web usage mining. However, due to the difficulty of labeling large quantities of data for supervised learning, clustering has been adopted. Clustering is unsupervised classification technique and domain independent so it can be applied to many different domains. Most of the methods used during pattern discovery phase are same as those applied to other data mining tasks.

There is immense of information available on the web so the user finds it difficult to get relevant information in short span of time. The recommendation system helps the user in their activity by suggesting some items of user's interest. These systems require analysis of the web log in order to find the web items of user's interest. There are many techniques available to analyze the user's behavior or web log[2].Among them clustering is most widely used to discover patterns. Clustering is needed in web log in order to distinguish similar items and dissimilar items in the web log. Clustering divides the whole web data into clusters. The items in one cluster resemble some similarity among them but are different to the items in the other clusters[3].The rest of paper is organized as follows. Section II presents a review on existing clustering techniques. Section III provides description about K-NN algorithm. Section IV describes genetic algorithm. Section V, presents the proposed methodology, Section VI, and provides the result analysis of proposed methodology. Section VII concludes the paper with future work.

II. LITERATURE SURVEY

Recommendation systems are the systems that guide the web users in their browsing activity by suggesting some web items that may be of their relevance. These systems analyze the web user browsing history in order to find the most accessed web site in their previous sessions[4].These systems use this information to provide suitable recommendations.

Clustering is the process of dividing the entire data set into small groups of data items .These group contain data items that resembles some common property among items of same group but dissimilar to the items in the other groups. These groups are referred as clusters.

A. K-Means Clustering Algorithm

In order to solve clustering problems K-means algorithm introduced. It is an unsupervised partition algorithm. It initially assigns k cluster centers to the data set, one for each cluster. The distance of every point in the data set to the cluster centers is calculated. The points that have smaller distances to a cluster centre are grouped into a cluster containing that cluster centre. Now the mean of all the points in the cluster is calculated to obtain new cluster centers for

Corresponding Author: *Animesh Shrivastava*
Dept. of CSE, Shri Vaishnav Institute of Technology & Science, Indore, India

each cluster. The process of forming new cluster centre continues until k cluster centre position becomes stable. This algorithm aim to minimize the sum of the squared distances to the cluster centres. This algorithm has problem with different cluster size and density [5].

B. *Hierarchical Clustering Algorithm*

This algorithm creates a hierarchy by either dividing or combining clusters. It uses top down or bottom up approach to form a hierarchy. It takes set of objects and finds distance among these objects each pair of objects with smaller distance is combined to form one cluster and this process continues until a single group containing all objects is formed. Now all the distances between pair of clusters is maintained in a metric and is updated with every merging of clusters. There are various forms of agglomerative hierarchical clustering. If the distance between the members of two clusters is maximum, average and minimum then linkage is complete, average and single respectively. This algorithm has some advantages as any form of similarity or distances are easily handled, applicable to any attribute type, embedded flexibility regarding a level of granularity and easily adapt to any function [6].

III. K-NN ALGORITHM

K-NN is a supervised classification technique. It classifies unlabeled objects based on their similarity with objects in the training set. It computes distance between the unlabeled data points and the training data points to find the closest points. As it is non-parametric, it does not make assumptions on the data distribution. This algorithm is not interested to use training data points to generalize. It uses training data points for testing purpose. This technique has uses in various areas like pattern recognition, nearest neighbour based content retrieval, gene expression, protein–protein interaction and 3D structure prediction [7].It is also used to measure the distance between two points using some distance functions such as Euclidian distance and Absolute distance.

Suppose given a data set containing n scenarios and each scenario having m features. Now for each feature, feature deviation is calculated. The feature whose deviation is minimal is the best feature in the data set.K-NN is also used to measure the dissimilarity and similarity among the individuals. The dissimilarity is measured using either Euclidian distance or absolute distance. The similarity is measured using correlation coefficient.

IV. GENETIC ALGORITHMS

Genetic Algorithm is a subset of evolutionary algorithm. Its working principle resembles the evolution of species. It can perform searching on complex and large data set. Genetic Algorithm can be applied to clustering and optimation problems [8]. It accepts a large population of all possible solutions and solution with the best fitness value propagate to the next generation. The steps involved in this process are as follows.

A. *Generate initial population*

The large and complex data set constitute the population. It contains all the possible solutions to the given problem. Each solution is represented in the form of a string. Each individual string of population is encoded in the binary bits 0 & 1.

B. Calculate Fitness Value

The fitness value of each individual string is calculated. The user defines the fitness function according to the specified problem. It evaluates the ability of each individual to produce the best individuals for the next generation.

C. *Selection*

Among all the individuals of a population, the individuals are randomly selected to pass through the process of crossover and mutation. The individuals with good fitness values are directly forwarded to the next generation and this process is called elitism.

D. *Crossover*

The selected individual strings exchange some part of their string to produce new individual strings. Therefore, the new individual strings are the best possible combination of their parent strings. However, some combinations might not produce good individual strings. Therefore, these are passed to the mutation operator.

E. *Mutation*

Some of the random changes made in the new individual strings by flipping binary bits from 0 to 1 & 1 to 0.This process ensure diversity among the individual strings

F. New generation

The new individuals produced from crossover and mutation operations are combined with elite individuals to form the new generation. This process continues until the required solution is obtained [9].

V. HYBRID APPROACH

Web Recommendation system predicts the next web pages for the user by analyzing the web access log. Genetic algorithm analyses the navigational behavior of user and finds the most accessed URL. Genetic algorithm takes the web access log as the initial population. It processes all the possible solution available in the population and after successive number of generations, the algorithm returns the fittest or the desired solution [10].As the individual reaches the desired fitness value, the algorithm gets terminated. The execution of this algorithm consumes more resources and affects the efficiency of the algorithm. In order to improve the performance of genetic algorithm, the population size is required to reduce. The initial population of genetic algorithm can be reduced by removing some of the sequences which are unutilized and impossible solution. KNN algorithm is employed for this purpose.

The proposed algorithm is a hybrid algorithm, which is designed using genetic algorithm and KNN algorithm. Where the processing steps are inherited from genetic algorithm and the distance measurement and sequence elimination process is derived using KNN algorithm. The hybrid algorithm takes the web access log as the initial population. The size of the population is reduced using KNN algorithm. It calculates distance among all the individuals of population using distance function.

$$D(x,y) = \sum_{k=0}^{n} |x_k - y_k| \qquad (1)$$

The individual sequences whose distance is greater than 0.5 are removed from the population. As these individuals get unutilized and cannot form the desired solution. Now the search space is reduced so the hybrid algorithm can obtain the desired solution in lesser time than the traditional genetic algorithm. Thus, the proposed system can be summarized using the given below steps.

Process:

```
1. Randomly generate initial population
2. In generated population
    a. Pick two sequences randomly
    b. Evaluate distance using $sequence_a - sequence_b$
    c. if distance > 0.5 then
    d. remove a sequence
    e. end if
3. Termination condition is checked.
4. Apply genetic operators as selection, crossover, and mutation.
5. New generation is obtained.
6. Remove impossible set of sequences.
7. Go to step 3
```

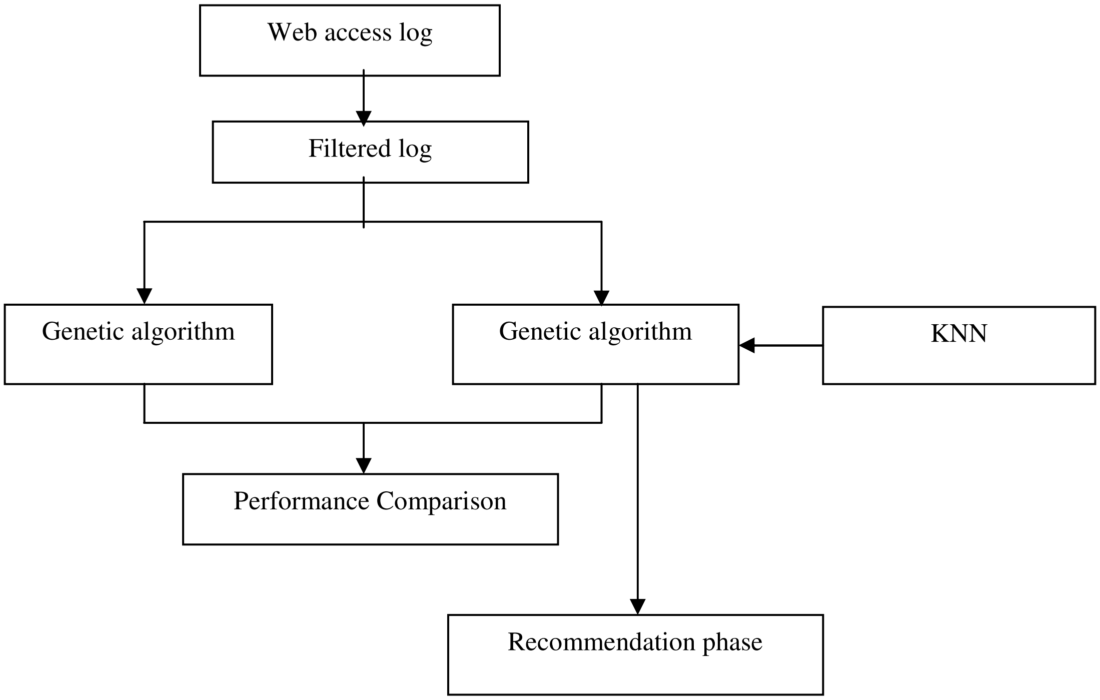

Figure 1

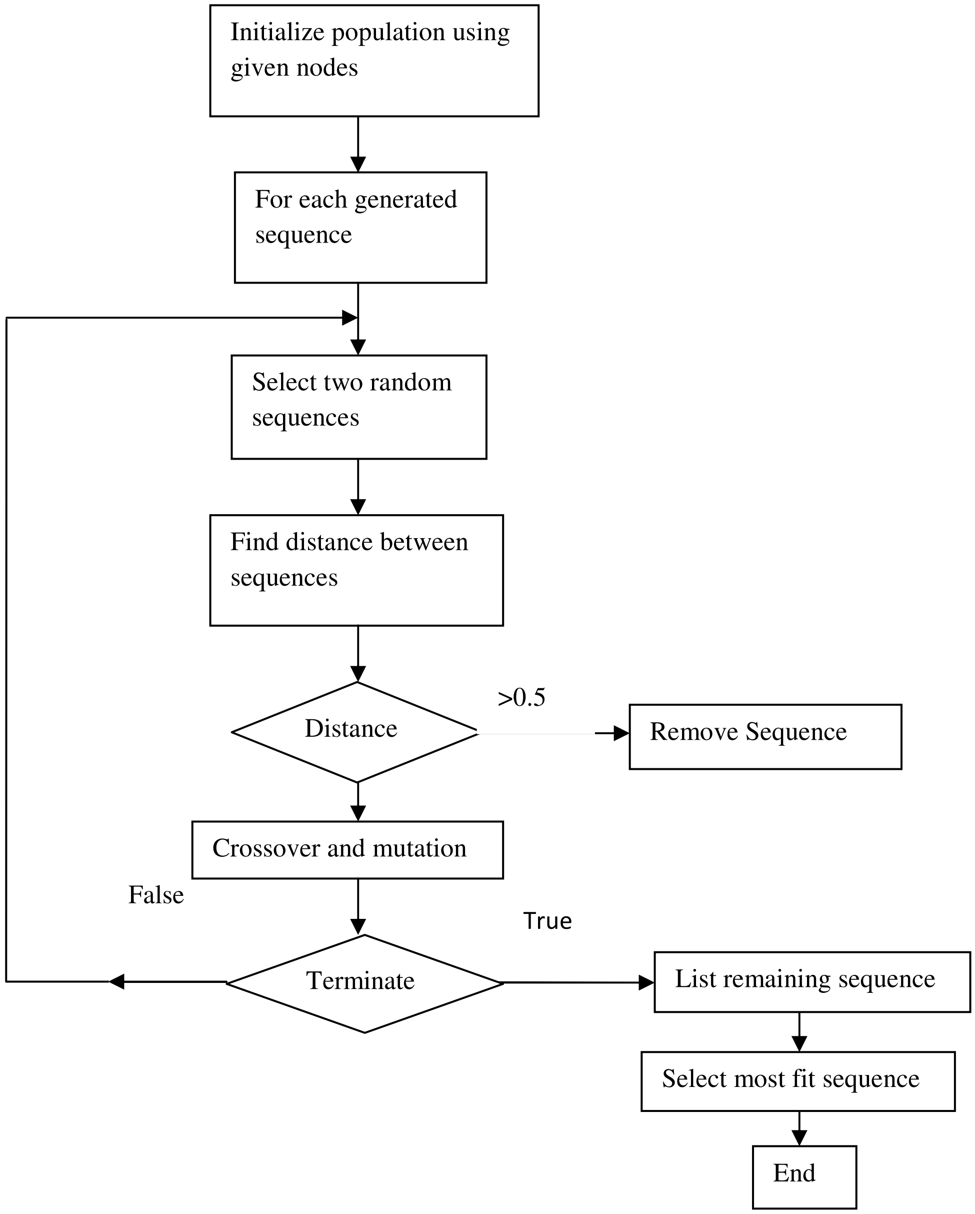
Initialize population using given nodes
For each generated sequence
Select two random sequences
Find distance between sequences
Distance
>0.5
Remove Sequence
Crossover and mutation
False
True
Terminate
List remaining sequence
Select most fit sequence
End

Figure 2 Algorithm Flow Chart

VI. RESULTS ANALYSIS

The performance parameters such as accuracy, error rate, memory used and time consumed are used to compare the performance of genetic algorithms and hybrid approach.

A. *Accuracy*

This parameter indicates the accuracy of the decision found by the system and evaluated using the following formula.

$$\text{Accuracy} = \frac{\text{Total correctly classified samples}}{\text{Total samples to classify}} \text{X} 100 \quad (2)$$

B. *Error Rate*

Error rate of the system indicates the error probability, which is found during the analysis by the system and evaluated using the given formula.

$$\text{Error rate} = 100 - \text{accuracy}\% \qquad (3)$$

C. *Memory Used*

The memory used provides the information how much memory is consumed during execution of the system.

D. *Search Time*

Search time is another performance parameter that indicates that for finding any suitable code block how much time is consumed by the system.

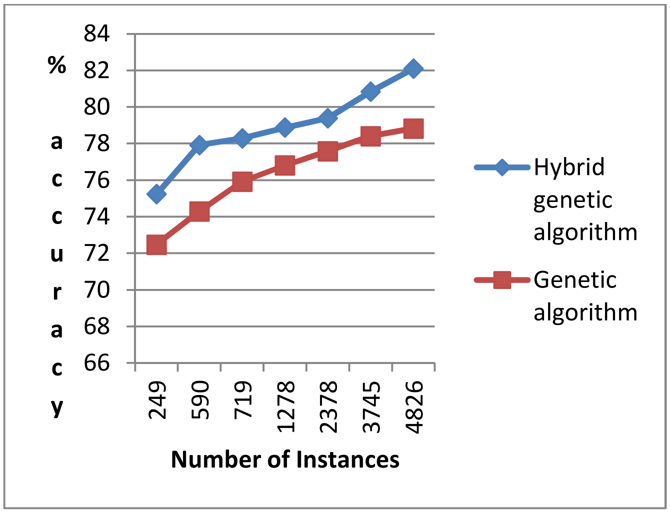

Figure 3

.

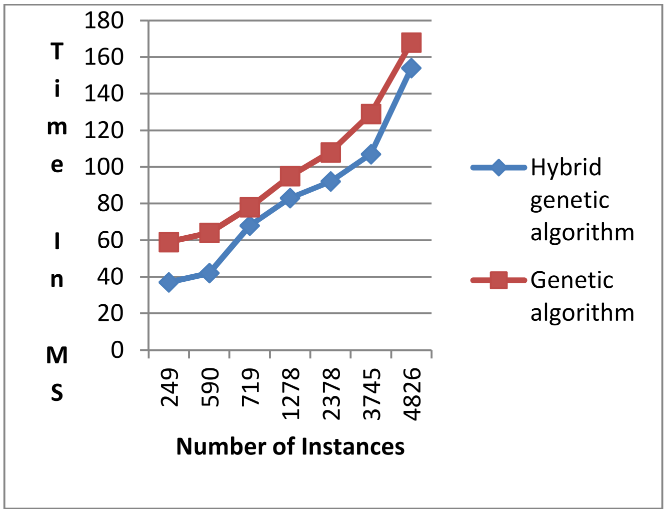

Figure 6

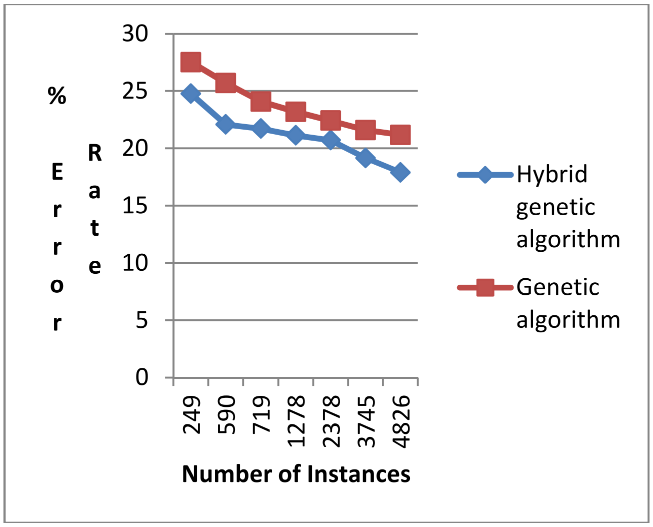

Figure 4

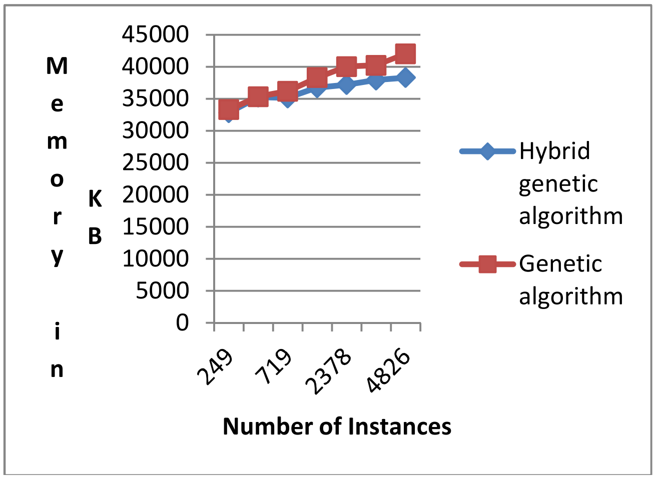

Figure 5

VII. CONCLUSION

In recommendation process, web usage mining plays an important role and clustering is mostly preferred to discover patterns. Genetic algorithm is highly efficient and effective search process for finding the optimum solution in a complex data domain. In this work, KNN is used to reduce the size of web log data then genetic algorithm is applied to find the most accessed URL. The performance comparison of genetic algorithm and knn with genetic algorithm is based upon the parameters such as accuracy, error rate, memory used and time consumed. It is found that hybrid approach performs better than genetic algorithm.

In near future the random selection process used in genetic algorithm can be directed according to the specific problem for the population generation and evaluation. In order to show effectiveness of the proposed approach, it can be implemented using real time applications.

ACKNOWLEDGEMENTS

It is a pleasure to express obligation and very thankful to all those people who have helped directly or indirectly in successful completion of this paper.

REFERENCES

[1] Jaideep Shrivastava, Robert Cooley, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD Explorations,ACM SIGKDD Jan **2000** Volume**1** Issue **2**.

[2] L.K. Joshila Grace,V.Maheswari, Dhinaharan Nagamalai,"Analysis of web logs and web user in web mining", International Journal of Network Security & Its Applications (IJNSA), Vol.**3**, No.**1**, January **2011**

[3] Osama Abu Abbas,"Comparision between Data Clustering Algorithms",The International Arab Journal of Information Technology, Vol **5**, No.**3**, July **2008**.

[4] C.P.Sumathi et. al,"Automatic Recommendation of Web Pages in Web usage mining" International Journal on Computer Science and Engineering Vol. **02**, No. **09**, **2010, 3046-3052**

[5] Tapas Kanungo, Nathan S. Netanyahu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE transactions on pattern analysis and machine intelligence, Vol**. 24**, no. **7**, July **2002**

[6] Hassan H. Malik, and John R. Kender, "Classification by Pattern-Based Hierarchical Clustering", Department of Computer Science, Columbia University, New York, NY 10027, USA{hhm2104, jrk}@cs.columbia.edu

[7] László Kozma Lkozma@cis.hut.fi,"k Nearest Neighbors algorithm" Helsinki University of Technology T-61.6020 Special Course in Computer and Information Science **20. 2. 2008**

[8] Olga Georgiou,Nicolas Tsapatsoulis, "Improving the Scalability of Recommender Systems by Clustering Using Genetic Algorithms", Volume 6352, 2010, pp **442-449** @Springer-Verlag Berlin Heidelberg ICANN **2010**

[9] Ujjwal Maulik,Sanghamitra Bandyopadhyay, "Genetic algorithm based clustering technique",PII: S 0 0 3 1 - 3 2 0 3 (9 9) 0 0 1 3 7 – 5@2000 Pattern Recognition Society. Published by Elsevier Science Ltd.

[10] Petra Kudov´a,"Clustering Genetic Algorithm",18th International Workshop on Database and Expert Systems Applications DOI 10.1109/DEXA.**2007**.65

AUTHOR PROFILE

Animesh Shrivastava completed his B.E. in Computer Science and Engineering from R.I.T.S. Bhopal affiliated to RGPV University Bhopal in 2010. He is pursuing M.E. in Computer Science and Engineering from S.V.I.T.S. Indore.