An Overview of Implementing the Genetic Algorithm towards Clustered Data

Pankaj B. Dhumane[1], S. R. Pande[2]
[1]Department of Computer Science, Sardar Patel Mahavidyalaya, Chandrapur
[2]Department of Computer Science, SSES Science College, Nagpur

Abstract

Web Mining is a fascinating area in data preparing that incorporates a vast assortment of utilizations i.e. recommendation framework outline, next client website page forecast, navigational example examination and others. In this paper another half and half clustering algorithm is proposed and actualized utilizing Genetic algorithm and K-NN algorithm and the execution of wanted algorithm is given utilizing a web recommendation framework which investigate client navigational example from web server get to log document and suggests the following client site page. The execution of the composed framework is assessed and recorded in this paper. As per the outcomes, the proposed half breed approach is productive and compelling for the given application space.

Keywords : Clustering, k-NN, Recommendation Systems, Genetic Algorithm.

I. INTRODUCTION

Web is a rich wellspring of data and knowledge, and this information source is adaptable. Be that as it may, the development of web information found in unstructured manner. Web information is found in three better places, over server get to log, in substance of website pages, and in association of site pages. The Web information is the gathering of information identified with the web addresses, time of the demand and reaction, sort of the working framework and program, operators utilized, status codes and techniques utilized as get or post. As the web log contains assortment of data, which may incorporate uproarious information, repetitive information and some unuseful data. This is the tremendous accumulation of Web data[1].Therefore, there must be some method, which recovers valuable data in limited ability to focus time. With a specific end goal to beat this issue, the information mining systems are accessible which extricate valuable data from the colossal web log. Information mining systems like affiliation rules, successive example mining, characterization and clustering. Before, arrangement methods were utilized as a part of web utilization mining. Be that as it may, because of the trouble of marking vast amounts of information for administered getting the hang of, clustering has been received. Clustering is unsupervised grouping procedure and domain autonomous so it can be connected to various domains. The majority of the strategies utilized amid design revelation stage are same as those connected to other information mining tasks.

There is tremendous of data accessible on the web so the client thinks that its hard to get important data in limited ability to focus time. The recommendation framework helps the client in their movement by proposing a few things of client's advantage. These systems require investigation of the web sign with a specific end goal to discover the web things of client's advantage. There are numerous systems accessible to examine the client's conduct or web log[2].Among them clustering is most broadly used to find designs. Clustering is required in web sign so as to recognize comparable things and disparate things in the web log. Clustering isolates the entire web information into bunches. The things in a single group look like some comparability among them however are distinctive to the things in the other clusters[3].The rest of paper is composed as takes after. Segment II shows an audit on existing clustering procedures. Area III gives depiction about K-NN algorithm. Segment IV depicts genetic algorithm. Area V, introduces the proposed philosophy, Section VI, and gives the outcome examination of proposed approach. Segment VII finishes up the paper with future work.

II. LITERATURE SURVEY

Recommendation systems are the systems that guide the web clients in their perusing action by proposing some web things that might be of their significance. These systems break down the web client perusing history so as to locate the most got to site in their past sessions[4].These systems utilize this data to give appropriate recommendations.

Clustering is the way toward isolating the whole informational index into little gatherings of information things .These gathering contain information things that takes after some basic property among things of same gathering yet not at all

like the things in alternate gatherings. These gatherings are alluded as bunches.

K-Means Clustering Algorithm

With a specific end goal to tackle clustering issues K-implies algorithm presented. It is an unsupervised segment algorithm. It at first doles out k bunch focuses to the informational index, one for each group. The separation of each point in the informational collection to the bunch focuses is computed. The focuses that have littler separations to a bunch focus are gathered into a group containing that group focus. Presently the mean of the considerable number of focuses in the group is figured to get new bunch places for each bunch. The way toward framing new group focus proceeds until the point when k bunch focus position ends up stable. This algorithm plan to limit the aggregate of the squared separations to the group focuses. This algorithm has issue with various bunch size and thickness [5].

Hierarchical Clustering Algorithm

This algorithm makes a progression by either isolating or consolidating groups. It utilizes top down or base up way to deal with shape a chain of command. It takes set of articles and discovers remove among these items each match of articles with littler separation is joined to shape one bunch and this procedure proceeds until the point when a solitary gathering containing all articles is framed. Presently every one of the separations between combine of bunches is kept up in a metric and is refreshed with each converging of groups. There are different types of agglomerative various leveled clustering. On the off chance that the separation between the individuals from two groups is most extreme, normal and least then linkage is finished, normal and single separately. This algorithm has a few favorable circumstances as any type of closeness or separations are effortlessly dealt with, relevant to any quality sort, installed adaptability in regards to a level of granularity and effectively adjust to any capacity [6].

K-NN ALGORITHM

K-NN is a regulated grouping procedure. It orders unlabeled articles in light of their closeness with objects in the preparation set. It processes remove between the unlabeled information focuses and the preparation information focuses to locate the nearest focuses. As it is non-parametric, it doesn't make presumptions on the information conveyance. This algorithm isn't intrigued to utilize preparing information focuses to sum up. It utilizes preparing information focuses for testing reason. This system has utilizes as a part of different zones like example acknowledgment, closest neighbor based substance recovery, quality articulation, protein– protein collaboration and 3D structure forecast [7].It is likewise used to quantify the separation between two focuses utilizing some separation capacities, for example, Euclidian separation and Absolute separation.

Assume given an informational collection containing n situations and every situation having m highlights. Presently for each element, highlight deviation is computed. The element whose deviation is negligible is the best component in the information set.K-NN is additionally used to quantify the uniqueness and similitude among the people. The uniqueness is estimated utilizing either Euclidian separation or supreme separation. The closeness is estimated utilizing relationship coefficient.

III. GENETIC ALGORITHMS

Genetic Algorithm is a subset of developmental algorithm. Its working rule looks like the advancement of species. It can perform seeking on intricate and substantial informational index. Genetic Algorithm can be connected to clustering and optimation issues [8]. It acknowledges a substantial populace of every conceivable arrangement and arrangement with the best wellness esteem proliferate to the people to come. The means associated with this procedure are as per the following.

A. Generate initial population

The extensive and complex informational collection constitute the population. It contains all the conceivable answers for the given issue. Every arrangement is spoken to as a string. Every individual string of population is encoded in the paired bits 0 and 1.

B. Calculate Fitness Value

The fitness value of every individual string is calculated. The client characterizes the fitness work as per the predefined issue. It assesses the capacity of every person to deliver the best people for the people to come.

C. Selection

Among every one of the people of a population, the people are arbitrarily chosen to go through the procedure of crossover and mutation. The people with great fitness values are specifically sent to the people to come and this procedure is called elitism.

D. Crossover

The chose singular strings trade some piece of their string to deliver new individual strings. In this way, the new individual strings are the most ideal blend of their parent strings. Be that as it may, a few mixes won't not create great individual strings. In this way, these are passed to the mutation administrator.

E. Mutation

A portion of the arbitrary changes made in the new individual strings by flipping parallel bits from 0 to 1 and 1 to 0.This process guarantee assorted variety among the individual strings

F.New generation

The new people delivered from crossover and mutation tasks are joined with first class people to shape the new generation. This procedure proceeds until the point when the required arrangement is acquired [9].

IV. HYBRID APPROACH

Web Recommendation framework predicts the following website pages for the client by investigating the web get to log. Genetic algorithm investigations the navigational conduct of client and finds the most got to URL. Genetic algorithm takes the web get to log as the initial population. It forms all the conceivable arrangement accessible in the population and after progressive number of generations, the algorithm restores the fittest or the coveted arrangement [10].As the individual achieves the coveted fitness value, the algorithm gets ended. The execution of this algorithm expends more assets and influences the effectiveness of the algorithm. So as to enhance the execution of genetic algorithm, the population estimate is required to decrease. The initial population of genetic algorithm can be diminished by expelling a portion of the groupings which are unutilized and incomprehensible arrangement. KNN algorithm is utilized for this reason.

The proposed algorithm is a crossover algorithm, which is planned utilizing genetic algorithm and KNN algorithm. Where the preparing steps are acquired from genetic algorithm and the separation estimation and arrangement disposal process is inferred utilizing KNN algorithm. The half breed algorithm takes the web get to log as the initial population. The measure of the population is lessened utilizing KNN algorithm. It calculates remove among every one of the people of population utilizing separation work.

The individual groupings whose separation is more prominent than 0.5 are expelled from the population. As these people get unutilized and cannot shape the coveted arrangement. Presently the pursuit space is diminished so the half and half algorithm can acquire the coveted arrangement in lesser time than the conventional genetic algorithm. Along these lines, the proposed framework can be compressed utilizing the given underneath steps.

Process:
Randomly generate initial population
In generated population
Pick two sequences randomly
Evaluate distance using sequencea − sequenceb
if distance > 0.5 then
remove a sequence
end if
Termination condition is checked.

Apply genetic operators as selection, crossover, and mutation.
New generation is obtained.
emove impossible set of sequences.

V. 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.

Accuracy

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

Accuracy = Total correctly classified samples X100 (2)
Total samples to classify

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.

Error rate = 100 – accuracy% (3)

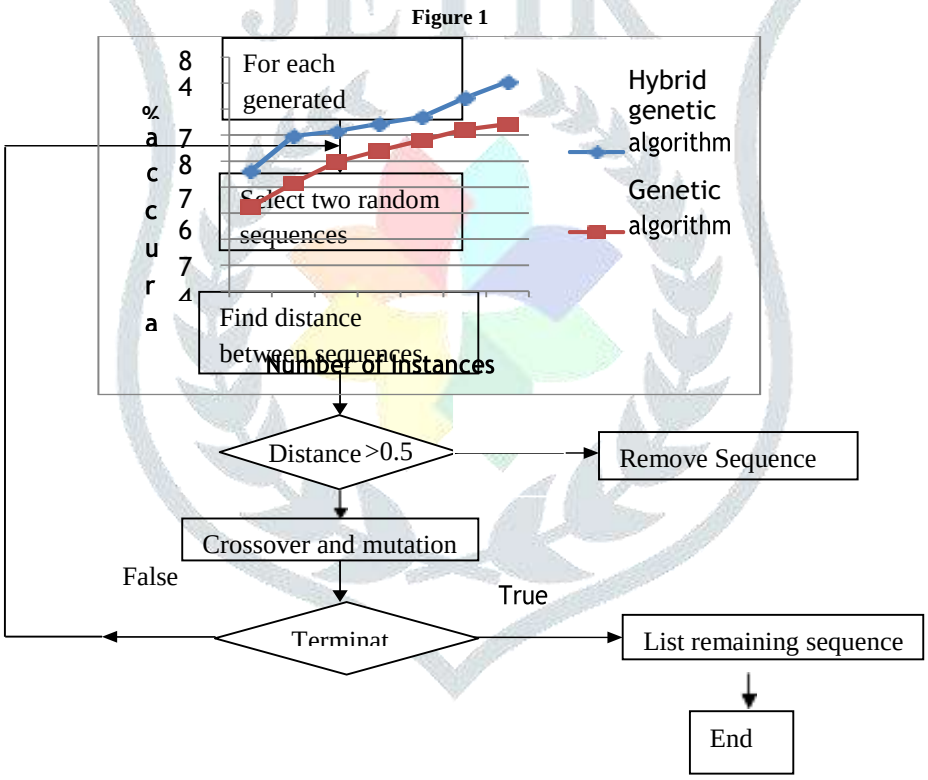
Memory Used

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

Search Time

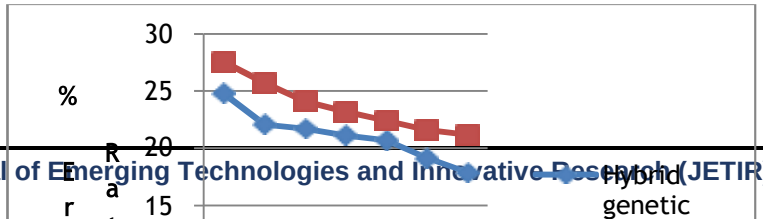
Search time is another performance parameter that indicates that for finding any suitable code block how much time is consumed by the system.

Figure 1



249 590 719 1278 2378 3745 4826

Figure 2



249 590 719 1278 2378 3745 4826

Figure 3

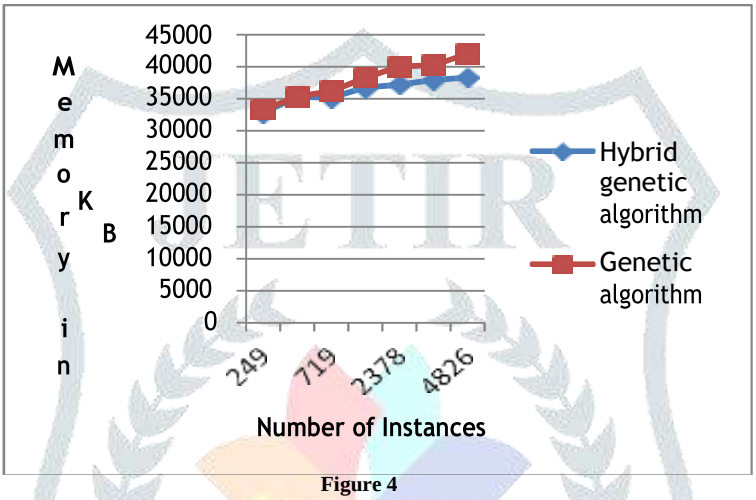


Figure 4

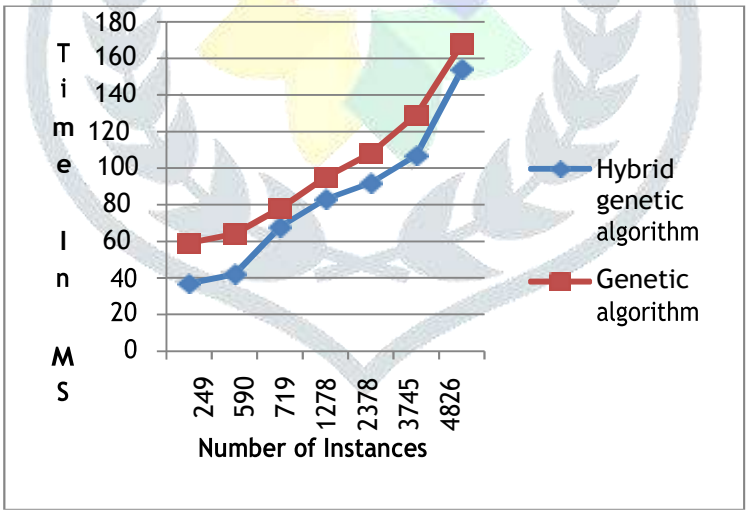


Figure 5

VI. CONCLUSION

In recommendation process, web utilization mining assumes a critical part and clustering is generally liked to find designs. Genetic algorithm is exceptionally proficient and successful look process for finding the ideal arrangement in a perplexing information domain. In this work, KNN is utilized to diminish the extent of web log information then genetic algorithm is connected to locate the most got to URL. The execution correlation of genetic algorithm and knn with genetic algorithm depends on the parameters, for example, exactness, mistake rate, memory utilized and time devoured. It is discovered that cross breed approach performs superior to genetic algorithm.

In not so distant future the arbitrary selection process utilized as a part of genetic algorithm can be coordinated by the

particular issue for the population generation and assessment. To demonstrate viability of the proposed approach, it can be actualized utilizing ongoing applications.

REFERENCES

Jaideep Shrivastava, Robert Cooley, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD Explorations,ACM SIGKDD Jan 2000 Volume1 Issue 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

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

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

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

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

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

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

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