OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

Integrated Web Usage Mining Algorithm for User's Navigation

Ms. Khushboo Saxena
Asst. Porofessor
Corporate Institute of Science and Technology, Bhopal

Dr. Akash Saxena
Associate Professor
Compucom Institute of IT & Management, Jaipur

Abstract: In an era of technology, human collects huge amount of structural and unstructured data, in this scenario World Wide Web (WWW) performs vital role to find the solution in a single place. Various implementers give different types of data model to implement this. Web mining can be understand by three ways i.e. web structure mining, web usage mining, and web content mining. This paper focus on web usage mining, aims to describe pre-fetching system, extract useful information among huge amount of data, define user's navigation, customer future prediction on the basis of navigation, Web personalization, web caching.

Various work of web usage mining has been proposed based on their threshold, time taken, and error rate; so in this paper, we proposed an integrated efficient web usage mining algorithm using fuzzy clustering and genetic algorithm which is able to generate better results as compared to previous results.

Keyword: World Wide Web (WWW), Fuzzy Clustering, Genetic algorithm, Web Usage Mining

Introduction

As exponential growth of data on to the internet, makes difficult to find useful content among huge documents. A conventional search engine provides hundreds of search results which is more time consuming. So to overcome these problem researchers move towards automated efficient algorithms which has been designed to give better results at minimum time. Clustering and classification are the two most popular methods that ensure to provide better results in for large voulme of data data on internet and genetic algorithm use to overcome the problem of clustering .

Fuzzy C-means Algorithm (FCM)

The goal of traditional clustering techniques is to assign each data point to only one cluster. In contrast, fuzzy

ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

clustering assigns different degrees of membership to each point where the membership of a point is shared among various clusters. Fuzzy clustering method has been chosen from the overlapping clustering method to be compared to non overlapped clustering method. The fuzzy was expected to perform better, in cases where there are a significant number of outliers. FCM is one of the fuzzy clustering techniques.

Let R be the set of real, R^p the set of p tuples of real, R^+ the set of nonnegative real, and W_{cn} the set of real c × n matrices. R^p will be called feature space, and elements $x \in R^p$ feature vectors; The fuzzy c-means algorithm uses iterative optimization to approximate minima of an objective function which is a member of a family of fuzzy c-means functional using a particular inner product norm metric as a similarity measure on $R^p \times R^p$. The distinction between family members is the result of the application of a weighting exponent m to the membership values used in the definition of the functional .

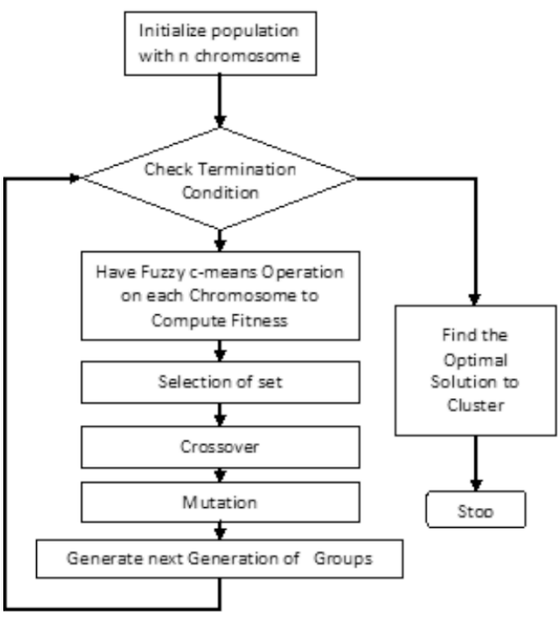


Figure 1. Proposed weblog mining algorithm based on FCM and GA

Genetic Algorithm

Conventional techniques of searching and optimization are very slow to find the solution; soft computing gives an efficient algorithm i.e. Genetic Algorithm (GA). This is a robust search algorithm which requires some information helps to search effectively in a large or poorly-understood search space. Genetic algorithm have distinguished operators such as crossover, mutation for optimize solution. Selection of proper crossover and mutation technique depends upon the encoding method and as per the requirement of the problem.

Understand the genetic algorithm with following steps:

Step 1: According to fitness value Select two parent chromosomes from a population.

ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

Step 2: With a crossover probability cross over the parents to form a new offspring. If no crossover was performed, offspring is an exact copy of parents.

Step 3: With a mutation probability mutate new offspring at each.

Step 4: Place new offspring in a new population.

Proposed Work

Proposed integrated web usage mining algorithm

Step-1: Make initial population randomly which have n chromosomes.

Step-2: Apply fuzzy C-means operation to calculate the fitness value of each chromosome

Step-3: Generate new population through randomly generated chromosomes and respective fitness value. To create a new population repeats the following steps: Initially select two parent chromosomes from a population which have better fitness.

(a) Apply crossover function on this better fitness chromosome to form a new offspring (children).

(b) Now apply mutation probability to mutate newly generated

(c) Finally fix this newly generated children in a new population

(d) Now replace the old population to new population.

Experiment Results

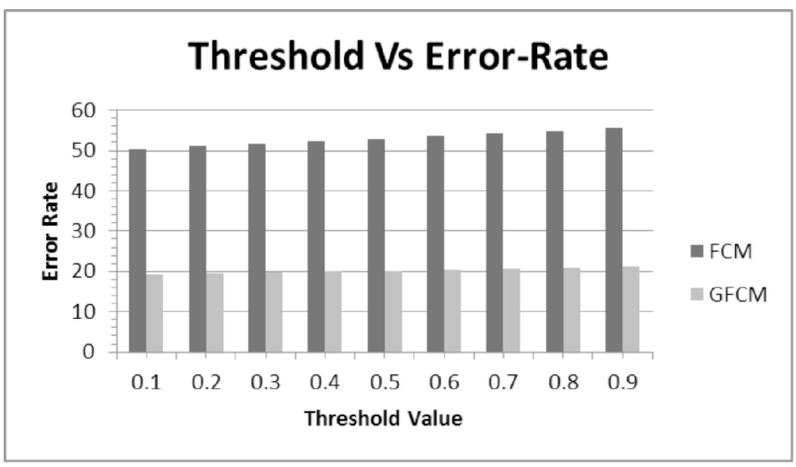
Measuring the quality of FCM algorithm and proposed weblog mining algorithm used Web log dataset in our experiment, which is extracted through Microsoft Server log file. This weblog file has 18 attributes. Table 1 shows the analysis of proposed algorithm and FCM algorithm based on some parameters such as threshold, error rate and time.

Table 1: Analysis of FCM and proposed algorithm for weblog dataset

Method	Threshold	Error Rate	Time
FCM	0.1	50.191	59.191740
GFCM	0.1	19.321	61.104803
FCM	0.2	51.212	62.939058
GFCM	0.2	19.252	63.751756
FCM	0.3	51.623	62.104490
GFCM	0.3	19.521	64.026603
FCM	0.4	52.522	62.372756
GFCM	0.4	18.989	63.583065
FCM	0.5	53.973	62.084189
GFCM	0.5	20.201	64.041465
FCM	0.6	53.618	62.155665
GFCM	0.6	20.427	63.856052
FCM	0.7	53.624	61.654710
GFCM	0.7	20.566	62.092531
FCM	0.8	54.909	60.510172
GFCM	0.8	20.872	63.207093
FCM	0.9	55.565	60.662062
GFCM	0.9	21.211	62.714621

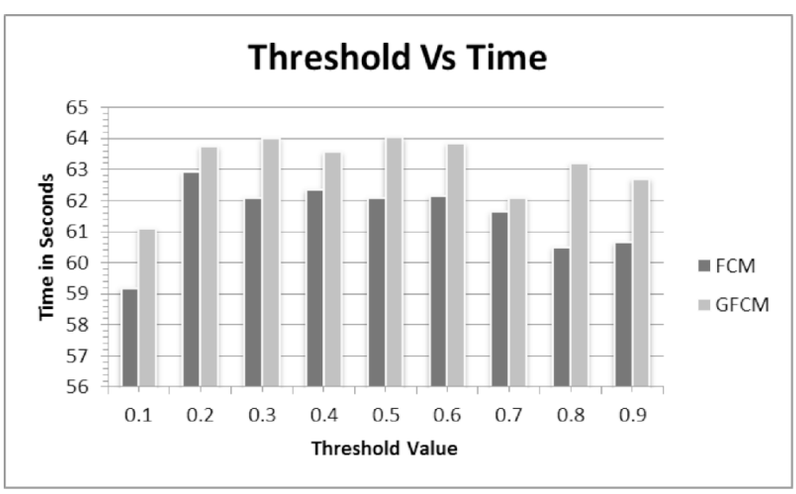
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

Graph 1 is plotted between threshold value and error-rate according to table 1. This graph plotted the generated error result of FCM and proposed weblog mining algorithm on the basis of threshold value. Graph 1 shows that as the threshold value increases the error-rate also increases in both the algorithms but the error rate of FCM is higher than the error rate of proposed weblog mining algorithm.



Graph 1: Threshold Vs Error Rate of weblog Dataset.

Similarly, graph 2 is plotted between the result value of threshold and time taken. This graph shows no consistency among threshold values and time taken means, the processing time of machine is depends on other factors too.



Graph 2: Threshold Vs Time of Dataset 1.

Conclusion

This research paper aims to propose an efficient algorithm to give high degree of accuracy in web usage mining. To solve these issues researchers have given various solutions across the globe for last several years. This research paper also suggested a web usage mining technique which provides less error and minimum time taken to execute the algorithm.

ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

Genetic algorithm is a heuristic search method, used to solve optimization problem which gives better results in terms of error rate and time execution and fuzzy c-means generates the realistic clusters with some membership.

Web usage mining techniques used in various real life applications such as quality control, wireless sensor/ad-hoc network, vehicle routing problem, optimization of data compression system, neural network, image processing and many more. This paper shows the experimental result and comparison of previous work and proposed web usage mining algorithm on the basis of error rate, execution time and threshold value.

References

- Deepak Kumar Niware (2014). Web Usage Mining through Efficient Genetic Fuzzy C-Means. International Journal of Computer Science and Network Security, VOL.14 No.6,p.p 113-117
- Rupinder Kaur, Simarjeet Kaur(2014).A Review: Techniques for Clustering of Web Usage Mining. International Journal of Science and Research, Vol 3 Issue 5 pp 1541-1545
- Aparna N. Gupta (2014).A Review: Study of Various Clustering Techniques in Web Usage Mining. International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3,pp 5888-5891
- Ranu Singhal, NirupamaTiwari (2013), A Survey: Web Log Mining using Genetic Algorithm. International Journal Of Engineering Sciences & Research Technology pp 1284-1286
- Helo´ýna Alves Arnaldo and Benjam´ýn R. C. Bedregal (2013). A new way to obtain the initial centroid clusters in Fuzzy C-Means algorithm
- Shaily G. Langhnoja, Mehul P. Barot, Darshak B. Mehta(2013).Web Usage Mining to Discover Visitor Group with Common Behavior Using DBSCAN Clustering Algorithm. International Journal of Engineering and Innovative Technology ,Volume 2, Issue 7
- Soniya P. Chaudhari, Prof. Hitesh Gupta, Prof. S. J. Patil (2013).Web Log Clustering using FCM and Swarm Intelligence Based Algorithms. International Journal of Innovative Research in Science, Engineering and Technology
- Karunesh Gupta , Manish Shrivastava (2012). Web Usage Data Clustering Using Improved Genetic Fuzzy C-Means Algorithm. International Journal of Advanced Computer Research,Volume-2 Number-2 Issue-4
- Ranu Singhal, Nirupama Tiwari(2012). Web Log Mining Based on Improved FCM Algorithm using Multiobjective Genetic Algorithm. International Journal of Software and Web Sciences, pp-59-63

ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

- Suresh, K.; Mohana, R. Madana; RamaMohanReddy, A.(2011).Improved FCM algorithm for Clustering on Web Usage Mining. SInternational Journal of Computer Science Issues (IJCSI);Jan2011, Vol. 8 Issue 1, pp.42
- K.Poongothai, M.Parimala and Dr. S.Sathiyabama.(2011) Efficient Web Usage Mining with Clustering. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3,pp 203-209.
- Qingtian Han, Xiaoyan Gao , Wenguo Wu (2008). Study on Web Mining Algorithm Based on Usage Mining, Computer- Aided Industrial Design and Conceptual Design. 9th International Conference on 22-25 Nov. 2008
- X Wang, J M. Garibaldi (2005). Simulated Annealing Fuzzy Clustering in Cancer Diagnosis, Informatica, 29, pp. 61–70
- K. Krishna, M. Narasimha Murty(1999).Genetic K-Means Algorithm. IEEE Transactions on systems, man, and cybernetics-part b: cybernetics,Vol. 29, No. 3, pp: 433- 439.