Volume 3, Issue 6, June 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Effective Web User Analysis and Clustering Using Fuzzy Possibilistic C Means Algorithm

P.Nithya
Doctoral student ,
Manonmaniam Sundaranar University,
Tirunelveli ,Tamil Nadu, India

Dr. P.Sumathi
Asst. Professor, PG & Research
Dept.of Computer Science,Govt.Arts College, Coimbatore,
Tamil Nadu, India

Abstract: ***In the internet era websites on the internet are helpful source of information for nearly every activity. Thus there's a fast development of World Wide Web in its volume of traffic and the size and complexional of websites. There are varieties of issues connected with the existing web usage mining approaches. Existing web usage mining algorithms suffer from problem of practical applicability. So, a completely unique analysis is extremely abundant necessary for the accurate prediction of future performance of web users with fast execution time. This paper consists of preprocessing and clustering of web users. Log data is routinely noisy and unclear, so preprocessing is an essential process for effective mining process. We present a novel approach to novel pre-processing of removing local and global noise and web robots and clustering Web site users into different groups and generating common user profiles. These profiles can be used to make recommendation, personalize Web sites, and for other uses such as targeting users for advertising. This FPCM (Fuzzy Possibilistic C Means) algorithm is relatively simple to use and gives comparable results with FCM (Fuzzy C Means) reported in the literature of web mining. Anonymous Microsoft Web Dataset is used for evaluating the proposed preprocessing technique and clustering process***

Keywords: ***WWW, Preprocessing, Data Cleaning, FPCM, FCM***

I. INTRODUCTION

The World Wide Web has become increasingly necessary as a medium for commerce likewise as for dissemination of information. In E-commerce, companies wish to analyze the user's preferences to place advertisements, to come back to a call their market strategy, and to produce customized guide to web customers. In today's information based society, there is associate urge for net surfers to search out the desired information from the overwhelming resources on the internet[1].An important attribute conducive to the recognition of a web site is that the degree of personalization it offers when presenting its services to users. However, improving the level of user personalization by reorganizing the entire Web site structure according to the interests of each user increases the number of computations at the Web server hosting the Web site. One solution to avoid this problem is to group users based on their Web interests, and then organize the structure of the Web site in a manner suitable to the Web needs of different groups. It's difficult to cluster users according to their web interests mainly attributable to two reasons: (1) users' interests are diverse and, (2) users' interests change with time.web access logs serve as a substantial supply of information about users' web access patterns. Properly exploited, the web access logs can be used to analyze and discover useful information regarding users' interests with the site [2]. The problem of web log mining consists in automated analyzing of web access logs so as to discover trends and regularities (patterns) in users' behavior. The discovered patterns typically sometimes used for improvement of web site organization and presentation. The term of adjectives web sites has been proposed to denote such automatically reworked web sites [10]. One of the most interesting web log mining methods is *web users clustering* [11]. The problem of web users clustering (or segmentation) is to use web access log files to partition a set of users into clusters such that the users within a cluster are more similar to each other than users from different clusters. The discovered clusters can then help in on-the-fly transformation of the web site content. In particular, web pages can be automatically linked by artificial hyperlinks. The idea is to try to match an active user's access pattern with one or more of the clusters discovered from the web log files [3].

Consider a web access log file, wherever every url request is hold on along with a timestamp and a user ip address. An example of such dataset is given in Figure 1.1 with its transformed type given in Figure 1.2, wherever an ordered list of url requests is hold on for every web user. Assume that the problem is to cluster the users' access sequences into two clusters, containing the users having similar access histories. The basic question here is: how to measure the similarity of two user's access sequences? It seems that e.g. the users 150.254.32.101 and 150.254.32.105 are similar since they both contain the same subsequence *'url7 → url8'*.In general, assumption is taken that two sequences are similar if either they contain the identical subsequences, or there exists a *connecting path.* Through a set of other sequences. In our example, the best solution would be to put the users 150.254.32.101 and 150.254.32.105 into one cluster and the users 150.254.32.102, 150.254.32.103, 150.254.32.104 into the other. Notice that the similarity between two user access sequences always depends on the presence of other sequences contributing to connecting paths [3].

time	User IP	item
03.18	150.254.32.101	url_1
03.19	150.254.32.104	url_2
03.20	150.254.32.102	url_3
03.21	150.254.32.103	url_4
03.25	150.254.32.105	url_1
03.27	150.254.32.102	url_5
03.31	150.254.32.104	url_2
03.34	150.254.32.101	url_{11}
03.36	150.254.32.104	url_2
03.40	150.254.32.102	url_{12}
04.05	150.254.32.103	url_4
04.25	150.254.32.105	url_{11}
04.27	150.254.32.102	url_5
04.34	150.254.32.104	url_{12}
04.38	150.254.32.101	url_1
04.45	150.254.32.104	url_{12}

Figure 1.1: Example Dataset

User_IP	Sequence
150.254.32.101	$url_1 \rightarrow url_4 \rightarrow url_5 \rightarrow url_7$
150.254.32.102	$url_3 \rightarrow url_2 \rightarrow url_9$
150.254.32.103	$url_8 \rightarrow url_{11} \rightarrow url_{12} \rightarrow url_{14}$
150.254.32.104	$url_2 \rightarrow url_8 \rightarrow url_{10} \rightarrow url_{17}$
150.254.32.105	$url_1 \rightarrow url_4 \rightarrow url_5 \rightarrow url_7 \rightarrow url_{10} \rightarrow url_{17}$

Figure 1.2: Transformed form of the example dataset

II. Related Works

Data mining, which is referred to as knowledge discovery in database, has become an important research area as a consequence of the maturity of very large databases. It uses techniques from areas such as machine learning, statistics, neural networks, and genetic algorithms to extract implicit information from very large amounts of data. The goals of data mining are prediction, identification, classification, and optimization. The knowledge discovered by data mining includes association rules, sequential patterns, clusters, and classification. Garofalakis [4] gives a review of popular data mining techniques and the algorithms for discovering the Web. Cooley *et al* [5] proposes taxonomy of Web mining and identified further research issues in this field. Yu [15] examines new developments in data mining and its application to personalization in E-commerce. Mobasher, Cooley and Srivastava [8] propose a technique for capturing common user profiles based on associationrule discovery and usage-based clustering. This technique directly computes overlapping clusters of URL references based on their co-occurrence patterns across user transactions. A *hypergraph* is built whose hyperedges are frequent itemsets that are found by the *a priori* algorithm. The weight of a hyperedge is calculated by averaging all the confidences of association rules in this frequent itemset. Clusters are obtained by applying the hypergraph partitioning algorithm to this hypergraph. The fuzzy c-means (FCM) algorithm [6] is one of the best known methods in fuzzy clustering. FCM introduces the concepts of fuzzy logic to classic K-means. Based on FCM, the fuzzy clustering multiple prototype (FCMP) framework [7] proposes a model of how the data are generated from a cluster structure to be identified. Nascimento et al. [11] extend the FCMP framework to some clustering criteria, and study the FCMP properties on fitting the underlying proposed model from which data is generated.

Tjhi and Chen [12] propose an algorithm called FCC-STF for clustering standard text documents, in order to expand the existing fuzzy clustering algorithms such as Fuzzy CoDoK and FCCM. The FCC-STF algorithm is different from the above two algorithms in that CC-STF uses a different fuzzifier. Ahn et al. [13] apply open user models to adaptive news systems, in order that the adaptive system becomes more transparent and controllable to the user. They explore the role of open and editable user profiles so that the users can see view and edit their interest profiles right after changes are made.

Chunhui et al., [16] presented a similarity based fuzzy and possibilistic c-means algorithm called SFPCM. It is derived from original fuzzy and possibilistic-means algorithm (FPCM) which was proposed by Bezdek. The difference between the two algorithms is that the proposed SFPCM algorithm processes relational data, and the original FPCM algorithm processes propositional data. Experiments are performed on 22 data sets from the UCI repository to compare SFPCM with FPCM. The results show that these two algorithms can generate similar results on the same data sets. SFPCM performs a little better than FPCM in the sense of classification accuracy, and it also converges more quickly than FPCM on these data sets.

III. Methodology

Web log data preprocessing is a complex process and takes 80% of total mining process. Log data is pretreated to get reliable data. The aim of data preprocessing is to select essential features clean data by removing irrelevant records and finally transform raw data into sessions.

A. Data Cleaning

The data cleaning process is removal of outliers or irrelevant data. Analyzing the huge amounts of records in server logs is a cumbersome activity. So initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, etc., are also downloaded which are not useful for further analysis are eliminated. The records with failed status

code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be removed from log files. Thus removal process in the experiment includes

1. Elimination of Local and Global Noise

Web noise can be normally categorized into two groups depending on their granularities:

Global Noise: It corresponds to the unnecessary objects with huge granularities, which are no smaller than individual pages. This noise includes mirror sites, duplicated Web pages and previous versioned Web pages, etc.

Local (Intra-Page) Noise: It corresponds to the irrelevant items inside a Web page. Local noise is typically incoherent with the major content of the page. This noise includes banner ads, navigational guides, decoration pictures, etc. These noises have to removed for better results.

2. The records of graphics, videos and the format information

The records have filename extension of GIF, JPEG, CSS, and so on, which can be found in the URI field of the every record, can be removed. This extension files are not actually the user interested web page, rather it is just the documents embedded in the web page. So it is not necessary to include in identifying the user interested web pages. This cleaning process helps in discarding unnecessary evaluation and also helps in fast identification of user interested patterns.

3. The records with the failed HTTP status code

The HTTP status code is then considered in the next process for cleaning. By examining the status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. This cleaning process will further reduce the evaluation time for determining the used interested patterns.

4. Method field

It should be pointed out that different from most other researches, records having value of POST or HEAD in Method field are reserved in present study for acquiring more accurate referrer information.

5. Robots Cleaning

Web robot (WR) (also called spider or bot) is a software tool that periodically scans a web site to extract its content. Web robots automatically follow all the hyperlinks from a web page. Search engines, such as Google, periodically use WRs to gather all the pages from a web site in order to update their search indexes. The number of requests from one WR may be equal to the number of the web site's URIs. If the web site does not attract many visitors, the number of requests coming from all the WRs that have visited the site might exceed that of human-generated requests.

Eliminating WR-generated log entries not only simplifies the mining task that will follow, but it also removes uninteresting sessions from the log file. Usually, a WR has a breadth (or depth) first search strategy and follows all the links from a web page. Therefore, a WR will generate a huge number of requests on a web site. Moreover, the requests of a WR are out of the analysis scope, as the analyst is interested in discovering knowledge about users' behavior.Most of the Web robots identify themselves by using the user agent field from the log file. Several databases referencing the known robots are maintained [Kos, ABC]. However, these databases are not exhaustive as each day new WRs appear or are being renamed, making the WR identification task more difficult.

To identify web robots' requests, the data cleaning module implements two different techniques.

- In the first technique, all records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed.
- The next technique is based on the fact that the crawlers retrieve pages in an automatic and exhaustive manner, so they are distinguished by a very high browsing speed. Therefore, for each different IP address, the browsing speed is calculated and all requests with this value more than a threshold are regarded as made by robots and are consequently removed. The value of the threshold is set up by analyzing the browser behavior arising from the considered log files.

This helps in accurate detection of user interested patterns by providing only the relevant web logs. Only the patterns that are much interested by the user will be resulted in the final phase of identification if this cleaning process is performed before start identifying the user interested patterns.

B. The session identification

User session identification is the process of analyzing usage data in order to extract useful information concerning user navigational behavior by structuring the requests contained into the Web log files. Access log files of a Web site consist in text files where the server stores all the accesses made by the users in chronological order. According to the Common Log Format, each log entry includes: the user's IP address, the request's date and time, the request method, the URL of the accessed page, the data transmission protocol, the return code indicating the status of the request, the size of the visited page in terms of number of bytes transmitted. Based on such information, we can determine the user sessions, i.e. the sequence of URLs that each user has accessed during his/her visit.

Here, a user session is defined as the finite set of URLs accessed by a user within a predefined time period (in our work, 25 minutes). Since the information about the user login is not available, user sessions are identified by grouping the requests originating from the same IP address during the established time period. Finally, data are filtered in order to retain only the most relevant pages and user sessions. At the end of preprocessing, we obtain collection of n_s sessions denoted by the set $S = \langle s_1, s_2, \ldots \ldots s_{ns} \rangle$. Each session contains information about accesses to pages during the session time. Precisely, a user session is formally described as a triple $S_i = \langle u_i, t_i, p_i \rangle$ where u_i represents the user identifier, t_i is the access time of the whole session, p_i is the set of all pages (with corresponding access information) requested during the i-th session.

Namely: $P^i = \langle (p_{i1}, t_{i1}, N_{i1}), (p_{i2}, t_{i2}, N_{i2}), \ldots\ldots (p_{i,n_i}, t_{in_i}, N_{in_i}) \rangle$

With $P_{ij} \in P$, where N_{ij} is the number of accesses to page P_{ij} during the i-th session and t_{ij} is the total time spent by the user on that page during the i-th session.

C. Fuzzy Possibilistic C Means Algorithm

It is a method of clustering which allows one piece of data to belong to two or more clusters. This method was developed by Dunn in 1973 [18] and Modified by Bezdek in 1981 [17] [19] and this is frequently used in pattern recognition. FPCM produces memberships and possibilities simultaneously, along with the usual point prototypes or cluster centers for each cluster. FPCM is a hybridization of possibilistic c-means (PCM) and fuzzy c-means (FCM) that often avoids various problems of PCM and FCM.

PFCM solves the noise sensitivity defect of FCM, overcomes the coincident clusters problem of PCM. But the noise data have an influence on the estimation of centroids. The choice of an appropriate objective function is the key to the success of the cluster analysis and to obtain better quality clustering results; so the clustering optimization is based on objective function. To meet a suitable objective function, we started from the following set of requirements: FPCM algorithm merges the advantages of both fuzzy and Possibilistic c-means techniques. Memberships and typicalities are essential for the accurate characteristic of data substructure in clustering technique.

$$J_{FPCM}(U,T,C) = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij}^{m} + t^{n})d^2(X_j, v_i) \quad (1)$$

With the following constraints:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1, \ldots, n\}$$

$$\sum_{j=1}^{n} t_{ij} = 1, \forall i \in \{1, \ldots, c\} \quad (2)$$

A solution of the objective function can be obtained through an iterative process where the degrees of membership, typicality and the cluster centers are updated with the equations as follows.

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2/(m-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n.$$

$$t_{ij} = \left[\sum_{k=1}^{n}\left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2/(\eta-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n. \quad (3)$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik}^{m} + t_{ik}^{\eta})X_K}{\sum_{k=1}^{n}(u_{ik}^{m} + t_{ik}^{\eta})}, 1 \le i \le c. \quad (4)$$

FPCM constructs memberships and possibilities simultaneously, together with the normal point prototypes or cluster centers for every cluster. Hybridization of Possibilistic C-Means (PCM) and Fuzzy C-Means (FCM) is the FPCM that frequently rejects several drawbacks of PCM, FCM. The noise sensitivity fault of FCM is solved by FPCM, which conquers the concurrent clusters drawbacks of PCM.

IV. Experimental Results

In order to evaluate the proposed preprocessing phase with robots cleaning, experiments were carried out using UCI Machine Learning Repository (University of California, Irvine). This repository contains 211 datasets. For the purpose of evaluating the proposed robot cleaning preprocessing phase, it is evaluated against,

- Initial log file and
- Preprocessed log file without removing robots.

Three standard datasets from the UCI Machine Learning Repository datasets and a real dataset is collected from reputed college were selected for the evaluation purpose. Following is the data sets used for evaluating the proposed preprocessing phase with robots cleaning.

- Anonymous Microsoft Web Dataset [20],

A. Anonymous Microsoft Web Dataset

This dataset consists of 37711 records in the log file. Then the data cleaning process is carried out. Initially, after removing records with local and global noise, graphics and videos format such gif, JPEG, etc., 29862 records are obtained. Then by checking the status code and method field, the total of 26854 records is resulted. Finally, 18452 records are resulted after applying robot cleaning process and it is shown in table 4.1.

Table 4.1: Number of Records Resulted After Prediction in Three Data Cleaning Phases In Anonymous Microsoft Web Dataset

Data Cleaning Phase	Number of Records
Initial Log	**37711**
After removing local and global noise, graphics and videos format records	**29862**
After checking status code and method field	**26854**
After robot cleaning process	**18452**

Figure 4.1: Time Taken for User Interested Pattern Anonymous Microsoft Web Dataset

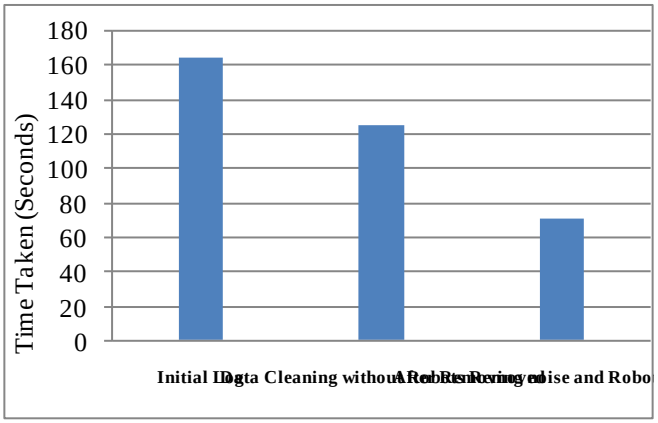

Figure 4.1 shows that the time required for the prediction of user interested pattern using initial log is 165 seconds, whereas, 126 seconds after cleaning by gif and status removal and it takes only 71 seconds after robots removal.

Clustering Accuracy:

The proposed web user clustering uses the modified fuzzy Possibilistic c means algorithm. The accuracy of the proposed web user clustering system is compared with the previous web user clustering systems which uses the fuzzy c means for clustering. Figure 4.2 shows the accuracy comparison of the proposed and the existing approaches. It is observed from the graph that the clustering accuracy of the modified fuzzy Possibilistic c means is very high when compared to the fuzzy c means.

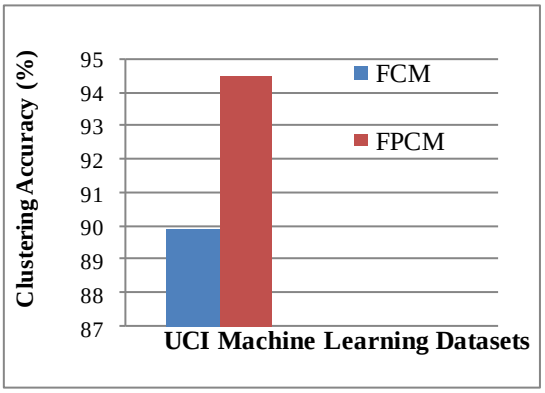

Figure 4.2: Clustering Accuracy Compared with FCM

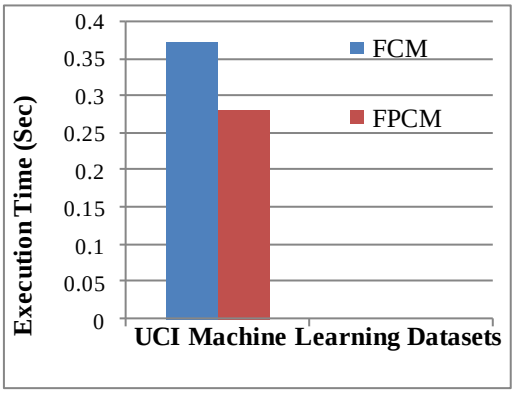

Figure 4.3: Execution Time Compared with FCM

Execution Time:

In the Figure 4.3, it shows graphical representation of the comparison of the time taken for the clustering classification of the fuzzy c means and modified fuzzy Possibilistic c means. It is observed from the graph that the time taken by the modified fuzzy Possibilistic c means is very less when compared to the fuzzy c means.

V. Conclusion

Data preprocessing dealing system for web usage mining has been analyzed and enforced for log data. Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the unsuccessful http status code and eventually robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries. Accurate input can be found if the byte rate of each and every record is found. The problem of web users clustering is to use web access log files to partition a set of users into clusters such that the users within a cluster are more similar to each other than users from different clusters are solved by using modified Fuzzy Possibilistic C Means algorithm and it is compared with FCM algorithm. An important feature of the algorithm is that it pre process the data and divide the web users into clusters that can be used to classify future web users. Thus, the proposed FPCM approach is best suited for the web users clustering applications effectively.

References

[1] Yunjuan Xie andVir V. Phoha., "Web User Clustering from Access Log Using Belief Function", Victoria, British Columbia, Canada, K-CAP'01, October 22-23, 2001.

[2] Santosh K. Rangarajan$, VirV. Phoha$, Kiran Balagani$, S. S. Iyengar*, Rastko Selmic, "Web User Clustering and Its Application to Prefetching Using ART Neural Networks", *Department of Computer Science Louisiana State University,2003.

[3] Tadeusz Morzy, Marek Wojciechowski, and Maciej Zakrzewicz., "Web Users Clustering", Poznan University of Technolog, Institute of Computing Science ul. Piotrowo 3a, 60-965 Poznan, Poland,, 2001.
[4] Garofalakis, M.N., Rastogi, R., Seshadri, S., and Shim K. Data mining and the Web: past, present and future In Proceedings of the second international workshop on Web information and data management, ACM 1999
[5] Cooley, R., Mobasher, B., and Srivastava, J. Web Mining Information and Pattern Discovery on the World Wide Web, ICTAI'97, 1997
[6] Bezdek JC, Ehrlich R. FCM: the fuzzy c-means clustering algorithm. Comput Geosci 1984; 10:191–203..
[7] Nascimento S, Mirkin B, Moura-Pires F. Multiple prototypes model for fuzzy clustering. In: Kok J, Hand D, Berthold M, editors. Advances in intelligent data analysis. Third international symposium (IDA'99), lecture notes in computer science, vol. 1642. Springer-Verlag; 1999. p. 269–79.
[8] Mobasher, B., Cooley R., and Srivastava, J. Creating Adaptive Web Sites Through Usage-based Clustering of URLs, Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, 1999.
[9] Perkowitz, M., Etzioni, O.: Adaptive web sites: an AI challenge, Proc. of the 15th Int'l Joint Conf. on Artificial Intelligence (1997)
[10] Yan, T.W., Jacobsen, M., Garcia-Molina, H., Dayal, U.: From User Access Patterns to Dynamic Hypertext Linking, proc. Of Fifth International WWW Conference (1996)
[11] Nascimento S, Mirkin B, Moura-Pires F. A fuzzy clustering model of data and fuzzy c-means. In: The ninth IEEE international conference on fuzzy systems soft computing in the information age (FUZZ-IEEE 2000), IEEE neural networks council. San Antonio, Texas, USA, May 2000. p. 302–7..
[12] Tjhi W-C, Chen LH. Fuzzy co-clustering of web documents. In: Fourth international conference on cyberworlds (CW 2005), 23–25 November 2005,
[13] Ahn J-W, Brusilovsky P, Grady J, He DQ, Syn SY. Open user profiles for adaptive news systems: help or harm? In: The proceedings of the sixteenth international World Wide Web conference (WWW2007). Banff, Alberta, Canada, May 8–12; 2007.
[14] A. Vaishnavi, " Effective Web personalization system using Modified Fuzzy Possibilistic C Means", Bonfring International Journal of Software Engineering and Soft Computing, Vol. 1, Special Issue, December 2011.
[15] Yu, P. S. Data Mining and Personalization Technologies Proceedings of the 6th International Conference on Database Systems for Advanced Applications, 1998
[16] Long Yu; Jian Xiao and Gao Zheng, "Robust Interval Type-2 Possibilistic C-means Clustering and its Application for Fuzzy Modeling", FSKD '09. Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, Vol: 4, Publication Year: 2009 , Pp. 360 – 365
[17] Bezdek, J, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York: Plenum Press, 1981.
[18] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics Vol. 3, Pp. 32-57, 1973.
[19] Havens Timothy C, Chitta Radha, Jain Anil K and Jin Rong, "Speedup of fuzzy and possibilistic kernel c-means for large-scale clustering", IEEE International Conference on Fuzzy Systems (FUZZ), Pp. 463–470, 2011.
[20] http://archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data

