A Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm

Gajendra Singh Chandel, Kailash Patidar,.Man Singh Mali

Faculty of Computer Science & Engineering, Sri Satya Sai Institute of Science & Technology Sehore, M.P
Sri Satya Sai Institute of Science and Technology Sehore M.P.

Summary
Web usage mining has become very critical for effective Web site management, business and support services, personalization, and network traffic flow analysis and so on. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, and network traffic flow analysis and so on. Previous study on Web usage mining using a concurrent Clustering has shown that the usage trend analysis very much depends on the performance of the clustering of the number of requests. In this paper, a novel approach FCM Algorithm is introduced kind of Clustering Technique, in the process of Web Usage Mining to detect user's patterns. The process details the transformations necessaries to modify the data storage in the Web Servers Log files to an input of FCM Algorithm. The Experiment show that the efficiency of FCM Algorithm is better than k-mean algorithm for web log data.
Key words:
Web Usage Mining, Clustering, Web Server Log File

1. Introduction

Web mining has been developed into an autonomous research area. Web mining involves a wide range of applications that aim at discovering and extracting hidden information in data stored on the Web. Web mining can be categorized into three different classes based on which part of the Web is to be mined [2] [3].These three categories are Web content mining, Web structure mining and Web usage mining. Web content mining [7] [6] is the task of discovering useful information available on-line. There are different kinds of Web content which can provide useful information to users, for example multimedia data, structured (i.e. XML documents), semi structured (i.e. HTML documents) and unstructured data (i.e. plain text). The aim of Web content mining is to provide an efficient mechanism to help the users to find the information they seek. Web content mining includes the task of organizing and clustering the documents and providing search engines for accessing the different documents by keywords, categories, contents. Web structure mining is the process of discovering the structure of hyperlinks within the Web. Practically, while Web content mining focuses on the inner-document information, Web structure mining discovers the link structures at the inter-document level. The aim is to identify the authoritative and the hub pages for a given subject. Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web.

The aim of Web usage mining is to discover patterns of user activities in order to better serve the needs of the users for example by dynamic link handling, by page recommendation etc. The aim of a Web site or Web portal is to supply the user the information which is useful for him. There is a great competition between the different commercial portals and Web sites because every user means eventually money (through advertisements, etc.). Thus the goal of each owner of a portal is to make his site more attractive for the user. For this reason the response time of each single site have to be kept below 2s. Moreover some extras have to be provided such as supplying dynamic content or links or recommending pages for the user that are possible of interest of the given user. Clustering of the user activities stored in different types of log files is a key issue in the Web community. There are three types of log files that can be used for Web usage mining [4]. Log files are stored on The server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional Information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server Side data. Web usage mining consists of three main steps:

1) preprocessing,
2) pattern discovery

Manuscript received January 5, 2016
Manuscript revised January 20, 2016

3) Pattern analysis [15]

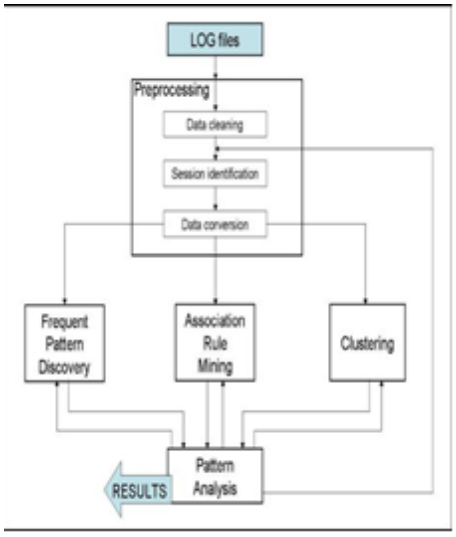

Fig. 1 shows the block diagram

In the preprocessing phase the data have to be Collected from the different places it is stored (client side, server side, proxy servers). After identifying the users, the click-streams of each user has to be split into sessions. In general the timeout for determining a session is set to 30 minute [5]. The pattern discovery phase means applying data mining techniques on the preprocessed log data.

It can be frequent pattern mining, association rule mining or clustering. In this paper we are dealing only with the task of clustering web usage log. In web usage mining there are two types of clusters to be discovered: usage clusters and page clusters. The aim of clustering users is to establish groups of users having similar browsing behavior. The users can be clustered based on several information. In the one hand, the user can be requested filling out a form regarding their interests, for example when registration on the web portal. The clustering of the users can be accomplished based on the forms. On the other hand, the clustering can be made based on the information gained from the log data collected during the user was navigating through the portal. Different types of user data can be collected using these methods, for example (I) characteristics of the user (age, gender, etc.), (ii) preferences and interests of the user, (iii) user's behavior pattern. The aim of clustering web pages is to have groups of pages that have similar content. This information can be useful for search engines or for applications that create dynamic index pages. The last step of the whole web usage mining process is to analyze the patterns found during the pattern discovery step. Web Usage Mining try to understand the patterns detected in before step.

2. Related Work on Web Usage Mining

The analysis of web user behaviors is known as web Usage Mining (WUM) that is to say, the application of data mining techniques to the problem of learning web usage patterns. The WUM is a relatively new research field that mainly focus on the study of the stream of users requests (or sessions) but that deals more generally with any user interactions with web sites (inserting or editing text in a web page (Kay, 2006), printing document). Masseglia et al. (Masseglia 1999a, 1999b) proposed webTool, an expert system that relies on sequential patterns extraction and uses the incremental method ISEWUM. The system aims at reorganizing web sites or at predicting web pages like Davison (Davison 2002, 1999). Perkowitz and Etzioni (Perkowitz 1999) reorganize web sites via the generation of a thematic index page using a conceptual clustering algorithm named PageGather. Some other works (Labroche 2003, Baraglia 2002, Heer 2001, Fu 1999, Yan 1996) apply clustering algorithms to discover homogeneous groups of (web) sessions. The underlying idea of these methods is that the algorithm should group in the same cluster the sessions that correspond to the users that navigate similarly, that is to say, that have the same motivation and interests.

Yan et al. (Yan 1996) use the First Leader clustering algorithm to create groups of sessions. The sessions are described as hits vectors in which each component corresponds to a web page and indicates the number of times it has been accessed during the session.

Estivill-Castro et al. (Estivill-Castro 2001) uses a k-Means-like algorithm that relies on a median rather than a mean to estimate the cluster centers. In (Nasraoui 2002), the authors use a fuzzy C Medoids approach, derived from the fuzzy C Medoids algorithm FCMdd described in (Krishnapuram 2001), to deal with the uncertainty and inaccuracy of the web sessions.

To automatically evaluate this number of expected clusters Heer and Chi (Heer 2002) propose a simple heuristic that computes the stability of the partitions for different number of clusters. The method works well but is extremely time consuming and thus may not be applicable in a real study context. In (Nasraoui 1999), the authors propose the CARD relational clustering algorithm (Competive Agglomeration for Relational Data) that is able to determine the appropriate number of clusters starting from a large number of small clusters. In (Suryavanshi 2005), the authors propose an incremental variant of a subtractive algorithm that allow to update partitions from previous analysis. This method determines automatically the number of clusters according to an estimation of the potential of each object to be a cluster center based on the density of the other objects in its neighborhood. Labroche et al. (Labroche 2003) describe a

relational clustering algorithm inspired by the chemical recognition system of ants named AntClust. In this model, each artificial ant possesses an odor representative of its nest membership called "label" and a genome, which is associated with a unique object of the data set. The algorithm simulates meetings between artificial ants according to behavioral rules to allow each ant to find the label (or nest) that best fits its genome. AntClust does not need to be initialized with the expected number of clusters and runs in linear time with the number of objects. AntClust has also been successfully applied to the web sessions clustering problem with various representations of sessions. More recently, Labroche (Labroche 2006) introduces the Leader Ant algorithm that relies on the same biological model and that is between 5 to 8 times faster than AntClust with similar clustering error values on benchmark data sets. Recently, Mobasher (Mobasher 2006) proposed an overview of the main data mining approaches that have been used to mine user profiles in a web personalization perspective (association rules, sequential patterns and clustering methods).

El-Shishiny, H.Sobhy Deraz,S.Badreddin (2008) use the artificial Neural Network to predict software aging phenomenon and analyze resource data collected on a typical running software system using Multilyaer Perceptron algoirhtm.Santhi, S.Shrivasan.P (2009) use Back progation algorithm (BPA) has been applied to learn the navigated web pages by different users at different session. the performance of the BPA in prediting the next possible web pages is about 90%

3. Traditional Work used in web usage Mining

3.1 Clustering Used In Web Usage Mining

Clustering is an unsupervised learning technique which aim is to find structure in a collection of unlabeled data. It is being used in many fields such as data mining, Intrusion Detection System[7].

3.2 K-means algorithm:

The K-means clustering[16][19] is a classical clustering algorithm. After an initial random assignment of example to K clusters, the centres of clusters are computed and the examples are assigned to the clusters with the closest centres. The process is repeated until the cluster centres do not significantly change.Once the cluster assignment is fixed, the mean distance of an example to cluster centres is used as the score. Using the K-means clustering algorithm, different clusters were specified and generated for each output class

Input: The number of clusters K and a dataset for intrusion detection

Output: A set of K-clusters that minimizes the squared – error criterion.

Algorithm:

1. Initialize K clusters (randomly select k elements from the data)
2. While cluster structure changes, repeat from 2.
3. Determine the cluster to which source data belongs Use Euclidean distance formula. Add element to cluster with min (Distance (xi, yj)).
4. Calculate the means of the clusters.
5. Change cluster centroids to means obtained Using Step 3.

The Main Disadvantage of K-Mean algorithm is that algorithm may take a large number of iterations through dense data sets before it can converge to produce the optimal set of centroids. This can be inefficient on large data sets due to its Unbounded convergence of cluster centroid.

4. Proposed Approach

4. 1 Fuzzy C - M ean algo rit hm

The Fuzzy C-Means algorithm was developed by Bezdek and his colleagues and is one of the most popular clustering algorithms. In the Fuzzy C-Means algorithm we start by initializing the value of C (the number of groups or clusters that we assume that the training data-points belong to) and the C cluster centers. Then, the Fuzzy C-Means algorithm redefines these cluster centers and membership values of the training points in a way that takes into consideration the distances of the points in the training set from these cluster centers. The membership value of a training point with respect to a cluster center indicates how strongly we believe that this training point belongs to the cluster center under consideration. This redefinition of cluster centers and membership values continues until satisfactory performance is attained.

Algorithm
Steps of Fuzzy C-Mean Algorithm
The algorithm is composed of the following steps:This algorithm determines the following steps [4].

Step1. Randomly initialize the membership matrix (U) that has constraints
Step2. Calculate centroids(ci)

Step3. Compute dissimilarity between centroids and data points Stop if its improvement over previous iteration is below a threshold.
Step4. Compute a new U Go to Step 2.

By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the "right" location within a data set.

Advantages OF fuzzy C-Mean Algorithm

Step1. Gives best result for overlapped data set and comparatively better then k-means algorithm.
Step2. Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more then one cluster center

5. Experimental Results

We can broadly categorize Web data clustering into (i) users' sessions-based and (ii) link-based. The former uses the Web log data and tries to group together a set of users' navigation sessions having similar characteristics. In this framework, Web-log data provide information about activities performed by a user from the moment the user enters a Web site to the moment the same user leaves it [8]. The records of users' actions within a Web site are stored in a log file. Each record in the log file contains the client's IP address, the date and time the request is received, the requested object and some additional information -such as protocol of request, size of the object etc. Figure 1 presents a sample of a Web access log file from a Web server.

5.1 Data Set Description

Web data clustering is the process of grouping Web data into "clusters" so that similar objects are in the same class and dissimilar objects are in different classes [2, 7]. Its goal is to organize data circulated over the Web into groups / collections in order to facilitate data availability and accessing.

We can broadly categorize Web data clustering into
(i) users' sessions-based and
(ii) link-based.

The former uses the Web log data and tries to group together a set of users' navigation sessions having similar characteristics. In this framework, Web-log data provide information about activities performed by a user from the moment the user enters a Web site to the moment the same user leaves it [8].

The records of users' actions within a Web site are stored in a log file. Each record in the log file contains the client's IP address, the date and time the request is received, the requested object and some additional information -such as protocol of request, size of the object etc.

Figure presents a sample of a Web access log file from a Web server.

Fig. 2 Web Log Data Description

5.3 Data Preprocessing

We need to do some data processing, such as invalid data cleaning and session identification. Data cleaning removes log entries (e.g. images, java scripts etc) that are not needed for the mining process. In order to identify unique users' sessions, heuristic methods are (mainly) used, based on IP and session time-outs. In this context, it is considered that a new session is created when a new IP address is encountered or if the visiting page time exceeds a time threshold (e.g. 30 minutes) for the same IP-address. Then, the original Web logs are transferred into user access session datasets for analysis.

5.4 User registration data

In addition to web access logs, our given input includes personal data on a subset of users, namely those who are registered to the website (registration is not mandatory). For a registered user, the system records the following information: sex, city, and province, civil status, born date. This information is provided by the user in a web form at the time of registration and, as one could expect, the quality of data is up to the user fairness.

5.5 Web URL

Resources in the World Wide Web are uniformly identified by means of URLs (Uniform Resource Locators). The syntax of an http URL is: '

http:// host.domain [':'port] [abs path ['?' query]]

Where { host.domain[:port] is the name of the server site. The TCP/IP port is optional (the default port is 80),{ abs path is the absolute path of the requested resource in the server file system. We further consider abs path of the form path '/' filename ['.' extension], i.e. consisting of the file system path, filename and ¯le extension. {query is an optional collection of parameters, to be passed as an input to a resource that is actually an executable program, e.g. a CGI script.

5.6 Empirical Setting

The K-Means and FCM algorithms are written in visual basic 6.0 as front-end and MS-Access used as Backend and compiled into mix files. K-Mean algorithms are relatively efficient due to vectored programming and active optimization. All experiments are run on a PC with a 3.06GHz Pentium-4 CPU with 1GB DRAM and running Windows XP.
In order to study the effect of the total number of clusters on the web mining results, we performed empirical studies with 80 total numbers of clusters.

5.7 Experimental Results

Now, we compare the aforementioned clustering algorithms on the whole data set with 5000 data set The Computation time results for the clustering algorithms with 80clusters are shown in Table 6.6 respectively

Table 1: A comparison between K-Mean and FCM algorithm

Cluster	K-Mean Algorithm (Time)	FCM Algorithm (Time)
20	393	260
40	581	338
60	806	425
80	997	586

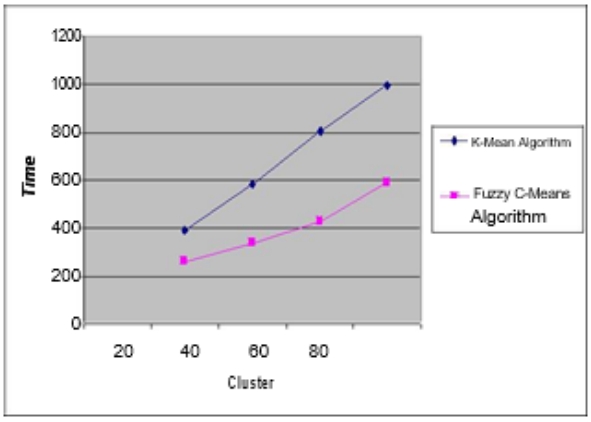

Fig. 3 Comparison Graph of K-Mean and FCM Algorithm for Timing Efficiency

Now we calculate the accuracy of cluster in term of time

Table 2: Comparison table of K-Mean and FCM Algorithm in term of accuracy

Cluster	**K-Mean Algorithm (Accuracy%)**	**FCM Algorithm (Accuracy%)**
20	46.07	47.40
40	44.19	46.62
60	41.96	45.75
80	40.03	44.14

We construct the graph in term of accuracy

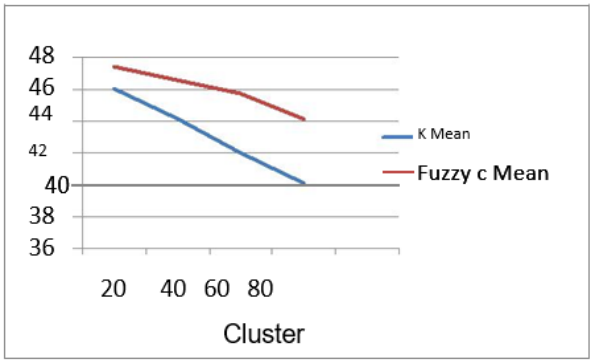

Fig. 4 Comparison Graph of K-Mean and FCM Algorithm for accuracy Efficiency

6. Conclusion

The K-Means and Fuzzy C-Means algorithms are one of the important clustering algorithms in data mining domain. Researchers have explored the usage of these two algorithms and reported satisfactory results in the literature. Generally, the time taken will vary from processor to processor. Both the algorithms have been implemented in DOTNET Language and their performance analyzed using normal and uniform distributions for different input data points. It is very evident from experiment result that the performance of the K-Means algorithm and FCM is better for both normal and uniform distributions. FCM produces close results to K-means clustering, yet it requires more computation time than K-means because of the fuzzy measures calculations involved in the algorithm. Thus for the data points generated using statistical distributions, the K-Means algorithm seems to be superior to Fuzzy C-Means. However, among the distributions Uniform distribution yields the best results. In the future, this research program will continue to investigate both K-means and FCM clustering algorithms. In particular, we are investigating methods to enable the optimal number of clusters to be automatically and consistently identified. Further tissue sections will be collected and used to evaluate our findings in this paper and future research.

References

[1] Abraham, A., Ramos, V. (2003). Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming.

[2] R. Kosala, H. Blockeel, Web Mining Research: A Survey, SIGKKD Explorations, vol. 2(1), July 2000.

[3] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKKD Explorations, vol.1, Jan 2000.

[4] Cernuzzi, L., Molas, M.L. (2004). Integrando diferentes técnicas de Data Mining en procesos de Web Usage Mining. Universidad Católica "Nuestra Señora de la Asunción". Asunción. Paraguay.

[5] Chau, M.; Chen, H., "Incorporating Web Analysis Into Neural Networks: An Example in Hopfield Net Searching", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume 37, Issue 3, May 2007 Page(s):352 – 358

[6] Raju, G.T.; Satyanarayana, P. S. "Knowledge Discovery from Web Usage Data: Extraction of Sequential Patterns through ART1 Neural Network Based Clustering Algorithm", International Conference on Computational Intelligence and Multimedia Applications, 2007, Volume 2, Issue , 13-15 Dec. 2007 Pages :88 -92

[7] Jalali, Mehrdad Mustapha, Norwati Mamat, Ali Sulaiman, Md. Nasir B. , " A new classification model for online predicting users' future movements", in International Symposium on Information Technology, 2008. ITSim 2008 26-28 Aug. 2008, Volume: 4, On page(s): 1-7, Kuala Lumpur, Malaysia

[8] Wang X., Abraham A. and Smith K.A, Soft Computing Paradigms for Web Access attern Analysis, Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery, pp. 631-635, 2002.