Novel Modified Fuzzy Possibilistic C Means (FPCM) for Web Log Mining by Removing Global Noise and Web Robots

[1]**Nithya Palani Sami;** and [2]**Sumathi Palani Aban**

[1]Dept. of Computer Science, Manonmanaiam Sundaranar University, Tirunelveli
[2]Research Dept. of Computer Science, Govt. Arts College, Coimbatore, Tamilnadu, India.

ID # (2795)
Received: 05/03/2014
In-revised: 28/02/2015
Correspondent Author:
Nithya Palani Sami
E-mail: nithi.selva@gmail.com

KEYWORDS

Preprocessing, Data Cleaning. Modified Fuzzy Possibilistic C Means, Fuzzy C Means, Hit Rate, Execution Time.

ABSTRACT

Nowadays, internet is a useful source of information in everyone's daily activity. Hence, this made a huge development of world wide web in its quantity of interchange and its size and difficulty of websites. Web Usage Mining WUM is one of the main applications of data mining, artificial intelligence and so on to the web data and forecast the user's visiting behaviors and obtains their interests by investigating the samples. SinceWUM directly involves in large range of applications, such as, e-commerce, e-learning, Web analytics, information retrieval etc. Web log data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link and this information can be used in several applications like adaptive web sites, modified services, customer summary, pre-fetching, generate attractive web sites etc. There are varieties of problems related with the existing web usage mining approaches. Existing web usage mining algorithms suffer from difficulty of practical applicability. So, a novel research is very much necessary for the accurate prediction of future performance of web users with rapid execution time. The main aim of this paper to remove the noise and web robots by novel approach and provide faster and easier data processing and it also helps in saving time and it resource. In this paper, a novel pre-processing technique is proposed by removing local and global noise and web robots. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset are used for evaluating the proposed preprocessing technique. An Effective Web User Analysis and Clustering are analyzed using Modified Fuzzy Possibilistic C Means (FPCM). Then results are evaluated using Hit Rate and Execution time.

أحدث طرق تحليل المعلومات بواسطة إزالة الضوضاء وروبوتات الشبكة العنكبوتية

[1]نيثيا بالاني سامي، و [2]سوماذي بالاني أبان

[1]قسم علوم الحاسوب، جامعة مانونمانايام سوندارانار تيرونيلفيلي
[2]قسم بحوث علوم الحاسوب، كلية الفنون، جامعة كويمباتوري، تاميلنادو، الهند

رقم المسودة: (2795)
تاريخ استلام المسودة: 2014/03/05
تاريخ المسودة المُعَدَلة: 2015/02/28
الباحث المُرَاسِل: نيثيا بالاني ساميبريد
بريد الكتروني:
nithi.selva@gmail.com

الكلمات الدالة

تجهيز، إعداد، تنظيف البيانات، تعديل تاوسائل الغامضة، معدل الإصابة، وقت التنفيذ

المُستلخص

يعتبر الانترنت مصدراً مهماً وأساسياً للمعلومات في كافة مناحي الحياة. وبناءً على ذلك فقد حدث تطور مُذهِل في شبكات التواصل والشبكة العنكبوتية بشكل عام كماً وكيفاً. وعليه يعتبراستخدام البحث في الشبكة أحد الأدوات المهمة للحصول على المعلومات وكذلك في مجال الذكاء الاصطناعي. لذلك يمكن وبتحليل هذه المعلومات وتحليل اهتمامات الزائرين والمراجعين لشبكات المعلومات المختلفة، استقراء الكثيرمن البيانات. هذا ويشمل تحليل معلومات الشبكة العديد من التطبيقات المختلفة منها التجارة الالكترونية والتعليم عن بعد وتحليل الشبكات واسترجاع المعلومات وغيرها. عليه يعتبرمعامل معلومات الشبكة أحد المصادرالأساسية لتحليل المعلومات لمستخدمي الشبكة العنكبوتية والمتعاملين مع محركات البحث المختلفة. كما يمكن أيضاً توظف بيانات الوقت المستهلك في البحث في تطبيقات عديدة وخدمات مختلفة مثل خدمات ما قبل البيع، إلى جانب إنشاء مواقع جذابة على الشبكة العنكبوتية. ووفقاً لتبان الاستخدامات والمستخدمين توجد بعض المشكلات الناتجة والمترتبة على طريقة تحليل المعلومات أوالبحث عنها، لعل من أهمها طبيعة القواعد المنظمة لهذا البحث وصعوبات تطبيقه عملياً. من هنا تبرزأهمية وضرورة اكتشاف طرق جديدة ومُحَسَنة وغير تقليدية وغيرمستهلكة للوقت. وعليه فإن الهدف الأساس من هذا البحث هو إزالة الضوضاء الالكترونية بطرق حديثة وغيرمسبوقة فضلاًعن إيجاد طرق سريعة وسهلة لاستخدامها في تحليل المعلومات. هذا ومن الجديربالذكر أنه قد تم استخدام قاعدة بيانات شركة مايكروسوفت (Microsoft Web Dataset & MSNBC) لتقييم الطريقة المطروحة في هذا البحث كما قد تم تقييم النتائج الخاصة بها باستخدام الهيتريت (Hit Rate & Execution time) وغيرها من الطرق المستخدمة في هذا المجال.

Introduction and Related Works

1. Introduction

Millions of electronic data are included on hundreds of millions data that are previously on-line today. With this significant increase of existing data on the Internet and because of its fast and disordered growth, the World Wide Web has evolved into a network of data with no proper organizational structure. In addition, survival of plentiful data in the network and the varying and heterogeneous nature of the web, web searching has become a tricky procedure for the majority of the users. This makes the users feel confused and at times lost in overloaded data that persist to enlarge. Moreover, e-business and web marketing are quickly developing and significance of anticipate the requirement of their customers is obvious particularly. As a result, guessing the users' interests for improving the usability of web or so called personalization has turn out to be very essential. Web personalization can be depicted as some action that builds the web experience of a user personalized according to the user's interest.

Generally, three kinds of information have to be handled in a web site: content, structure, and log data. Content data contains of anything is in a web page, structure data is nothing but the organization of the content and usage data is nothing but the usage patterns of web sites. The usage of the data mining process to these dissimilar data sets is based on the three different research directions in the area of web mining: web content mining, web structure mining and web usage mining (Jalali, *et al.*, 2008) (Maratea, and Petrosino, 2009).

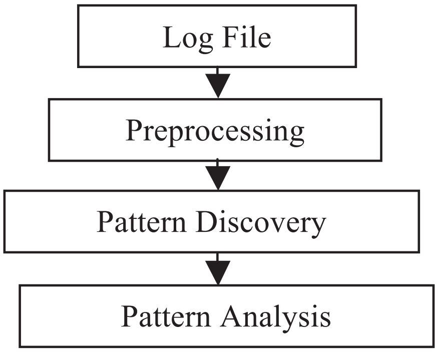

Figure 1: Steps in Web Usage Mining

Web usage mining (Jian Chen, *et al.*, 2004) (Wu, Yu, and Ball**,** 1998) consists of three main steps. Web log file is usually given as input. Figure 1 shows steps in web usage mining.

Preprocessing is an important step because of the complex nature of the Web architecture which takes 80% in mining process. The raw data is pretreated to get reliable sessions for efficient mining. It includes the domain dependent tasks of data cleaning, user identification, session identification, and path completion and construction of transactions. Data cleaning is the task of removing irrelevant records that are not necessary for mining. Data cleaning includes; elimination of local and global noise, removal of records of graphics, videos and the format information, removal of records with the failed HTTP status code, and robots cleaning.

User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user sessions. Path completion is used to fill missing page references in a session. Classifications of transactions are used to know the users interest ad navigational behavior. The second step in web usage mining (Labroche, *et al.*, 2007) is knowledge extraction in which data mining algorithms like association rule mining techniques, clustering, classification etc. are applied in preprocessed data. The third step is pattern analysis in which tools are provided to facilitate the transformation of information into knowledge. Knowledge query mechanism such as SQL is the most common method of pattern analysis. This paper focuses on path completion process which is used to append lost pages and construction of transactions in preprocessing stage. In this study a referrer-based method is proposed to efficiently construct the reliable transactions in data preprocessing.

2. Related Works

The discovery of the users' navigational patterns using SOM is proposed by (Labroche, *et al.*, 2007). In (Etminani. *et al.*, 2009) presented a Web usage mining technique based on fuzzy clustering in Identifying Target Group. In (Jianxi Zhang, *et al.,* 2009) suggests a complete idea for the pattern discovery of Web usage mining. In (Nina, *et al.,* 2009) given a Web Usage Mining technique based on the sequences of clicking patterns in a grid computing environment. The author discovers the usage of MSCP in a distributed grid computing surroundings and expresses its effectiveness by empirical cases. In (Chih-Hung Wu, *et al.,* 2010) proposed the usage of incremental fuzzy clustering to Web Usage Mining. Rough set based feature selection for web usage mining is proposed by (Aghabozorgi, and Wah, 2009). In (Inbarani, 2007) put forth a web usage mining technique based on LCS algorithm for online predicting recommendation systems. For providing the online prediction effectively, in (Shinde, and Kulkarni,2008) provides a architecture for online recommendation for predicting in Web Usage Mining System. In (Zhang Huiying and Liang Wei, 2004) given an intelligent algorithm of data pre-processing in Web usage mining. In (Nasraoui, *et al.,* 2008) provides a whole framework and findings in mining Web usage navigation from Web log files of a genuine Web site which has every challenging characteristics of real-life Web usage mining, together with evolving user profiles and external data describing an ontology of the Web content. In (Hogo, *et al.,* 2003) proposed the temporal Web usage mining of Web users on single educational Web site with the help of the adapted Kohonen SOM based on rough set properties. A development of data preprocessing technique for Web usage mining and the information's of algorithm for path completion are provided by(Yan Li, *et al.,* 2008). (Baraglia, and Palmerini, 2002) proposed a Web Usage Mining (WUM) system, called (Suggest) which continuously creates the suggested connections to Web pages of probable importance for a user. In (Chu-Hui Lee and Yu-Hsiang Fu, 2008) put forth a Web Usage Mining technique based on clustering of browsing characteristics.

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.

1. Data Cleaning

The process of data cleaning 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:

1.1. Elimination of Local and Global Noise

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

i. 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*.

ii. 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 remove for better results.

1.2. 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.

1.3. 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.

1.4. 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.

1.5. 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 second or 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.

2.The Fuzzy Clustering Algorithm

The basic theory about Fuzzy Clustering analysis based on Fuzzy Equivalent Relation is that Fuzzy Corresponding R (fuzzy relation) is a subset of U×U (universe of fuzzy set), when we are using λ-threshold, the subset of U×U will be just an ordinary equivalent relation and the objects in the U will be classified. When λ is reduced from 1 to 0, gradually the classification will merge forming a dynamic clustering dendritic diagram consequently. Thus, the establishment of fuzzy relation R is a key step of fuzzy analysis method.

Fuzzy Clustering is an iterative algorithm. The aim of Fuzzy Cluster is to find cluster centers (centroid) that minimize a dissimilarity function. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point (Bezdek, 1981).

2.1. 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.

2.2. The Record Identification

One record in web access log is written as: 219.144.222.253 [16/Aug/2004:15:36:11 +0800] "GET /images/1 r3 c2.jpg HTTP/1.1" 200 418 "http://202.117.16.119:8089/index.html" "Mozilla /4.0 (compatible; MSIE 6.0; Windows NT 5.1)". The meaning of each field is explained in table 1

Table 1: Format of Web Access Log

Field	Meaning
219.144.222.253	User's IP address (UIP)
[16/ Aug/2004:15:36 :11+0800]	The date and time of the request
GET	The method of the request
/images/1….r3…	The URL of the current request (URI)
Http/1.1	The version of the transport protocol (Version)
200	The HTTP status code returned to the client (Byte)
418	The content length of the page transferred (Bytes)
http://202.11...	The URL requested just before (Refer URI)
Mozilla/4.0(…	Browser & Os (BrowserOS)

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 $S = \{s_1, s_2, \ldots \ldots s_{ns}\}$. Each session contains information about accesses to pages during the session time. Precisely, a user session is formally described as a triple $S_i = \{u_i, t_i, p_i\}$ 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: $P^i = \{(p_{i1}, t_{i1}, N_{i1}), (p_{i2}, t_{i2}, N_{i2}), \ldots \ldots (p_{i,n_i}, t_{in_i}, N_{in_i})\}$ 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.

3. 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 and Modified by Bezdek in 1981, and this is frequently used in pattern recognition (Dunn, 1973) (Bezdek , 1981). Fuzzy Possibilistic C Means (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).

Fuzzy Possibilistic C Means (FPCM) solves the noise sensitivity defect of Fuzzy C-Means (FCM), overcomes the coincident clusters problem of Possibilistic C-Means (PCM). 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: Fuzzy Possibilistic C Means (FPCM) algorithm merges the advantages of both Fuzzy and Possibilistic c-means (FCM) & (PCM) techniques. Memberships and typicality's are essential for the accurate characteristic of data substructure in clustering technique.

$$J_{\downarrow}(FPCM\ )(U,T,C) = \sum_{\downarrow}(i=1)^{\uparrow}c \equiv \sum_{\downarrow}(j=1)^{\uparrow}n \equiv \llbracket(\mu\rrbracket_{\downarrow}ij^{\uparrow}m + t^{\uparrow}(n))d^{\uparrow}(2)(X_{\downarrow}(j)\dashv, v_{\downarrow}i) \quad (1)$$

With the following constraints:

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

$$\sum_{j=1}^{n} t_{ij} = 1, \forall i \in \{1, \dots, c\} \quad (3)$$

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.

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2/(m-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n. \quad (4)$$

$$t_{ij} = \left[\sum_{k=1}^{n}\left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2/(\eta-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n. \quad (5)$$

$$v_i = \frac{\sum_{k=1}^{n}\llbracket(u\rrbracket_{ik}^{m} + t_{ik}^{\eta})X_K}{\sum_{k=1}^{n}\llbracket(u\rrbracket_{ik}^{m} + t_{ik}^{\eta})}, 1 \le i \le c. \quad (6)$$

Fuzzy Possibilistic C Means (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).

Wen-Liang Hung presented a new approach called Modified Suppressed Fuzzy c-means (MS-FCM), which drastically improves the performance of (FCM) owing to the use of prototype-driven learning of parameter α (Ming Yang and Hong Li, 2009). Exponential separation strength between clusters is the base for the learning process of α and is updated at each of the iteration. The parameter α can be calculated as:

$$- \min_{i \ne k} \llbracket \frac{\llbracket \| v \rrbracket_i - v_k \| \square \rrbracket^2}{\beta} \rrbracket \quad (7)$$

In the above equation β is a normalized term hence β is chosen as a sample variance. It is to be noted that the common value used for this parameter by all nodes at each iteration may induce an error. A new parameter is integrated which suppresses this common value of α and substitutes it by a new parameter like a weight to each vector. Or every point of the data set possesses a weight in association with every cluster. Accordingly this weight allows having a better selection. The following equation is used to calculate the weight.

$$w_{ji} = exp\left[-\frac{\|x_j - v_i\|^2}{\left[\sum_{j=1}^{n}\|x_j - \overline{v}\|^2\right] * {}^{c}/_{n}}\right] \quad (8)$$

In the previous equation $\mathbf{w_{ji}}$ denotes the weight of the point $\mathbf{j}$ in relation to the class $\mathbf{j}$. In order to modify the fuzzy and typical partition, this weight is exploited. The objective function is composed of two expressions: the first is fuzzy function which uses fuzziness weighting exponent and the second is possibilistic function which uses a typical weighting exponent; however the two coefficients in the objective function are only used as exhibitor of membership and typicality. A new relation, lightly different, enabling a faster decrease in the function and enhancement in the membership and the typicality when they tend toward 1 and decrease this degree when they tend toward 0. This relation is to add weighting exponent as exhibitor of distance in the two under objective functions. The objective function of the (MFPCM) can be given as follows:

$$J_{MFPCM} = \sum_{i=1}^{c}\sum_{j=1}^{n} \llbracket(\mu\rrbracket_{ij}^{m} w_{ij}^{m} d^{2m}(x_j, v) + t_{ij}^{\eta} w_{ij}^{\eta} d^{2\eta} \llbracket(x\rrbracket_j, v_i)) \quad (9)$$

U = {$\mathbf{\mu_{ij}}$} represents a fuzzy partition matrix, is defined as:

$$\mu_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d?X_j, v_i}{d?X_j, v_k}\right)^{2m/(m-1)}\right]^{-1} \quad (10)$$

T = {$\mathbf{t_{ij}}$} represents a typical partition matrix, is defined as:

$$t_{ij} = \left[\sum_{k=1}^{n} \left(\frac{d?X_j, v_i}{d?X_j, v_k} \right)^{2\eta/(\eta-1)} \right]^{-1} \quad (11)$$

V = {$\mathbf{v_i}$} represents c centers of the clusters, is defined as:

$$v_i = \frac{\sum_{j=1}^{n} [(\mu]_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta}) * X_j}{\sum_{j=1}^{n} [(\mu]_{ik}^{m} w_{ji}^{m} + t_{ik}^{\eta} w_{ij}^{\eta})} \quad (12)$$

Experimental Results

In order to evaluate the proposed preprocessing phase with robots cleaning, experiments were carried out using University of California Irvine (UCI) Machine Learning Repository. This repository contains 211 datasets. Three standard datasets from the University of California Irvine (UCI) Machine Learning Repository datasets and a real dataset is collected from reputed college were selected for the evaluation purpose. Following are the data sets used for evaluating the proposed preprocessing phase with robots cleaning.

i. Anonymous Microsoft Web Dataset:- (http://www.archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data)
ii. MSNBC.com Anonymous Web Dataset :- (http://www.archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data)
The web user analysis are evaluated using: Hit Rate, and Execution time

1. Preprocessing

Evaluate the preprocessing phase by using Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset. In this Anonymous Microsoft Web dataset consists of 37711 records in the log file. Whereas MSNBC.com Anonymous Web dataset contains 989818 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 and 865412 records are obtained in this two type of datasets. Preprocessing phase is briefly described in (Nithya, 2012).

2. Hit Rate

The proposed web user clustering uses the Modified Fuzzy Possibilistic C Means (MFPCM) algorithm. The accuracy of the proposed web user clustering system is compared with the previous web user clustering systems which uses the Modified Fuzzy C Means (MFCM) for clustering. Figure 2 shows the clustering accuracy comparison of the proposed and the existing approaches.

The Hit Rate of FPCM cluster view is from 75.3% on Top-1 relevance to 85 % on Top-5 relevance which is shown in figure 2. Table 2 gives the comparison table of Hit Rate for proposed approaches against existing approaches. The Hit Rate of FCM cluster view is from 62% on Top-1 relevance to 72% on Top-5 relevance. Then the Hit Rate of MFCM cluster view is from 65% on Top- 1 relevance to 75 on Top- 5 relevance. From the four methods MFPCM, FPCM, MFCM, FCM, proposed approaches of MFPCM are proved better results against FPCM, MFCM, and FCM.

Figure 2 shows the comparison of Hit Rate for proposed approaches against existing approaches. From the figure it is clearly noticed that the proposed method of MFPCM gives better result and shows higher performance than the FPCM, MFCM, and FCM.

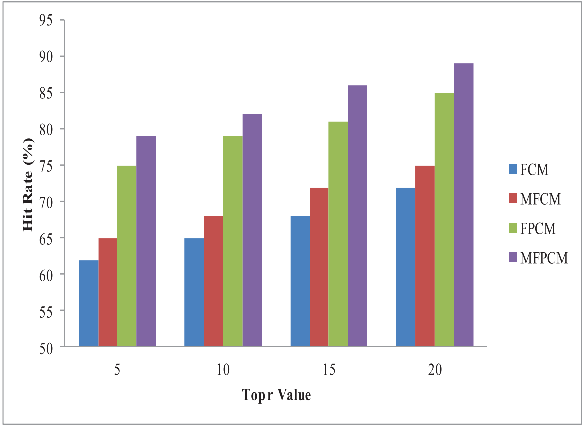

Figure 2: Comparsion of Hit Rate

Table 2: Comparison of Hit Rate for Proposed Method with Existing Method

Top r value	FCM	MFCM	FPCM	MFPCM
05	62	65	75	79
10	65	68	79	82
15	68	72	81	86
20	72	75	85	89

3. Execution Time

Execution time shows the time taken by the standard methods to execute the process in all the datasets. The method that takes less time to execute is considered as the best one and it helps in increasing the growth of clustering results. Table 3 gives the comparison table for the execution time for proposed method against existing method. That the Proposed method of MFPCM have 0.25 seconds is very low compared to other approaches followed by FPCM have 0.28, MFCM, 0.32 seconds, and FCM 0.37 seconds.

Table **3:** Comparison of Execution Time

Methods	Execution Time
FCM	0.37
MFCM	0.32
FPCM	0.28
MFPCM	0.25

In the Figure 3, it shows graphical representation of the comparison of the time taken for the clustering classification proposed method against existing methods. It is observed from the graph that the time taken by the MFPCM is very less when compared to the FPCM, MFCM, FCM.

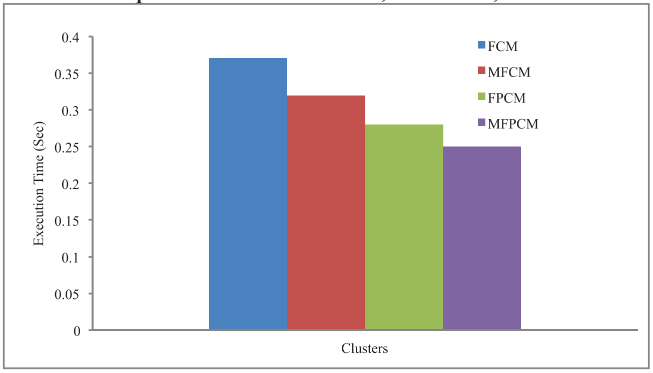

Figure 3: Comparsion of Execution Time

Conclusion

This paper mainly focuses on web usage mining with the help of clustering process. The break out the complexity in clustering the web data is initially preprocessed. It reduces the noises and web robots effectively. Here used Modified Fuzzy Possibilistic algorithm for the purpose of clustering. (FPCM) algorithm has advantages of both fuzzy and Possibilistic c-means techniques. The experimental results of time and hit rate of Prediction Model express these clustering effective suits for web usage mining. Data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the failed (HTTP) status code and finally robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries. This preprocessing step is used to give a reliable input for data mining tasks. Accurate input can be found if the byte rate of each and every record is found. The data cleaning phase implemented in this paper will helps in determining only the relevant logs that the user is interested in. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset are used for evaluating the proposed preprocessing technique and it reveals that number of records.

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 (MFPCM) and it is compared with (FCM) algorithm. And the experimental result shows that the proposed technique results in higher hit rate and less execution time. Thus, the proposed (MFPCM) approach is best suited for the web users clustering applications effectively.

References

Aghabozorgi SR (2009) Using Incremental Fuzzy Clustering to Web Usage Mining. *In:* **Wah TY**(eds), *SOCPAR'09 Proceedings of the 2009 International Conference of Soft Computing and Pattern Recognition,* IEEE Computer Society Washington, DC, USA, pp653-658. Available at: http://www.dl.acm.org/citation.cfm?id=1685695

Baraglia R (2002) SUGGEST: a Web usage Mining System. *In:* **Palmerini P** (eds), *Proceedings of International Conference on Informa-*

tion Technology: Coding and Computing 8-10 April 2002, Las Vegas, NV, USA, pp 282-287. Available at: https://www.ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7847

Bezdek J (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*: Advanced *Applications In Pattern Recognition.* Plenum Press. New York USA & London, UK, pp1-253.
Available at: http://www.download.springer.com/static/pdf/518/bfm%253A978-1-4757-0450-1%252F1.pdf?auth66=1395160039_62369c3cda1ecc2046d8840135551a5f&ext=.pdf

Chih-Hung Wu (2010) Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment. *In:* **Yen-Liang Wu; Yuan-Ming Chang;** and **Ming-Hung Hung** (eds), *Proceedings of International Conference on Machine Learning and Cybernetics (ICMLC) 11-14 July 2010,* Qingdao, China , vol. 6, pp2909-2914.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5580751

Chu-Hui Lee (2008) Web Usage Mining Based on Clustering of Browsing Features. *In:* **Yu-Hsiang Fu** (eds) *Proceedings of the 8th International Conference on Intelligent Systems Design and Applications*, *26-28 Nov. 2008,* Kaohsiung, Taiwan, vol. 1, pp281-286.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4696217

Dunn JC (1973) A Fuzzy Relative of the ISODATA Process and its use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics,* **3** (3): 32-57.
Available at: http://www.tandfonline.com/doi/abs/10.1080/01969727308546046#preview

Etminani K (2009) Web usage Mining: Discovery of the users' Navigational Patterns using SOM. *In:* **Delui AR; Yanehsari NR;** and **Rouhani M** (eds), *Proceedings of First International Conference on Networked Digital Technologies*, , *28-31 July 2009,* Ostrava, Czech Republic, pp224-249.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5272158

Havens Timothy C (2011) Speedup of Fuzzy and Possibilistic Kernel C-Means for Large Scale Clustering. *In:* **Chitta Radha; Jain Anil Kand Jin Rong** (eds), *Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ), 27- 30 June 2011,* Taipe, Taiwan, pp463–470.
Available at: http://www.ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6007618

Hogo (2003) Temporal Web usage Mining. *In:* **Snorek M;** and **Lingras P** (eds), *Proceedings of International Conference on Web Intelligence*, *13- 17 Oct. 2003,* pp 450-453.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1241237

Inbarani HH (2007) Rough Set Based Feature Selection for Web Usage Mining. *In:* **Thangavel K;** and **Pethalakshmi A** (eds), *Proceedings of International Conference on Computational Intelligence and Multimedia Applications 13-15 Dec. 2007*, Sivakasi, Tamil Nadu, India, vol. 1 pp33-38.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4426549

Jalali M (2008) A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems. *In:* **Mustapha N; Sulaiman NB;** and **Mamat A** (eds), *Proceedings of 12th International Conference Information Visualisation, 9-11 July 2008,* London, UK, pp302-307.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4577963&url=

Jian Chen (2004) Discovering Web usage Patterns by Mining Cross Transaction Association Rules. *In:* **Jian Yin; Tung AKH;** and **Bin Liu** (eds) *Proceedings of International Conference on Machine Learning and Cybernetics,* 26-29 Augt. 2004, vol. 5, pp2655-2660.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1378232

Jianxi Zhang (2009) Web Usage Mining Based on Fuzzy Clustering in Identifying Target Group. *In:* **Peiying Zhao; Lin Shang;** and **Lunsheng Wang** (eds), *International Colloquium on Computing, Communication, Control, and Management*, 8-9 Augt. 2009, Sanya, China, vol. 4, pp209-212.

Labroche N (2007) A New Web Usage Mining and Visualization Tool. *In:* **Lesot MJ;** and **Yaffi L** (eds), *Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence*, 29-31 Oct. 2007, Patras, Greek, vol. 1, pp321-328.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4410301

Maratea A (2009) An Heuristic Approach to Page Recommendation in Web Usage Mining. *In:* **Petrosino A** (ed) *Proceedings of Ninth International Conference on Intelligent Systems Design and Applications*, 30 Nov.-2 Dec. 2009, Pisa, Tuscany, Central Italy, pp1043-1048.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5364464

Ming Yang (2009) User Analysis Based on Fuzzy Clustering. *In:* **Hong Li** (ed) *Proceedings of 2009 International Conference on Business Intelligence and Financial Engineering*, 24-26 July 2009, Beijing, China, pp164-169.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5208905

Nasraoui O (2008) A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites. *In:* **Soliman M; Saka E; Badia A;** and **Germain R** (ed)**,** *Proceedings of IEEE Transactions on Knowledge and Data Engineering,* Feb. 2008, *IEEE Computer Society*, vol.20, pp 202-215.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4358953

Nina SP(2009) Pattern Discovery of Web Usage Mining. In: **Rahman M; Bhuiyan KI;** and **Ahmed K** (eds), *Proceedings of International Conference on Computer Technology and Development,* 13-15 Nov. 2009, Kota Kinabalu, Malaysia, vol. 1, pp499-503.
Available at: http://www. ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5359726

Nithya P(2012) Novel Pre Processing Technique for Web Log Mining by Removing Global Noise and Web Robots. *In:* **Sumathi P** (ed), *Proceedings of National Conference on Computing and Communication Systems (NCCCS),* 21-22 Nov. 2012, Durgapur, Kolkata, india, pp.1-5 .
Available at: http://www.ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6412976

Shinde SK (2008) A New Approach for on Line Recommender System in Web Usage Mining. *In*: **Kulkarni UV**(ed), *Proceedings of International Conference on Advanced Computer Theory and Engineering,* 20-22 Dec. 2008, Phuket, Thailand, pp973- 977.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4737102

Wu KL; Yu PS; and **Ballman A** (1998) Speed Tracer: A Web Usage Mining and Analysis Tool. *IBM Systems Journal*, **37** (1): 89-105.
Available at: http://www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.4791

Yan Li (2008) Research on Path Completion Technique in Web Usage Mining. *In:* **Boqin Feng;** and **Qinjiao Mao** (eds), *Proceedings of International Symposium on Computer Science and Computational Technology, 20- 22 Dec. 2008,* Shanghai, China, vol. 1, pp554-559.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4731490

Zhang Huiying; and **Liang Wei** (2004) An Intelligent Algorithm of Data Pre-Processing in Web Usage Mining. *Fifth World Congress on Intelligent Control and Automation 15-19 June 2004,* vol. 4, pp 3119- 3123.
Available at: http://www.ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1343095&