Fuzzy Temporal Clustering Approach for E-Commerce Websites

Sudhamathy G. [*1], Jothi Venkateswaran C. [#2]

[*1] Department of Computer Applications, Velammal College of Engineering and Technology, Viraganoor, Madurai – 625 009, India
[#2] Department of Computer Science, Presidency College (Autonomous), Chennai – 600 005, India
[1] sudhamathi10@hotmail.com
[2] jothivenkateswaran@yahoo.co.in
[*] Department of Computer Applications, Velammal College of Engineering and Technology, Viraganoor, Madurai – 625 009, India
[1] sudhamathi10@hotmail.com

***Abstract*—In this paper a novel approach for clustering of web logs data and to predict intelligent recommendations on the E-Commerce web sites is proposed so as to improve the marketing strategy and to improve customer loyalty. Fuzzy Temporal Clustering Approach (FTCA) performs clustering of the web site visitors and the web site pages based on the frequency of visit and time spent. Time plays a crucial role in the analysis of web usage. Hence these clusters are studied over a period of time to study the migration behaviour of the users and the pages across periods. Such a study can provide intelligent recommendations for the E-Commerce web sites that focus on specific product recommendations and behavioural targeting. Experimental evaluation of the method has proved that this approach FTCA is most efficient, easy to use and a useful clustering approach.**

***Keyword*-Web Usage Mining, Web Logs, Clustering, Fuzzy Logic, Temporal**

I. INTRODUCTION

Web server records interaction information between the users and the web server in the web log files. This information in the web log files hides user's access patterns and interests and is of great significance for analysis of the user requirements, providing users with personalized services, assisting web personnel and optimizing web sites. Therefore, the web log mining attracts increasingly attention in the fields of science and business. To analyse user's interest on the web pages, clustering techniques are often used in Web log mining. This interest of the users on the web pages is dynamic and they change over a period of time [4]. Hence the clusters of users and the pages also change over a period of time. The clustering algorithms can be divided into partition method, hierarchical method, density based method, grid-based method, model based method and etc. The assessments on the clustering algorithm mainly uses two measurement indexes, namely intra class distance and inter class distance. An algorithm that can produce high-quality clustering effect must meet the following two conditions, namely the intra class data or object similarity is the strongest, while inter-class data or object similarity is the weakest. Clustering is a basic understanding activity of human beings. Only through appropriate clustering, the things can be easily researched, and the internal laws of things can be mastered by human beings. The so called clustering is to put things together into a class based on some attributes of things, so that the intra-class similarity is weak as possible and inter-class similarity is big as possible [13].

Having said that a novel clustering approach is proposed to find the clusters of users and pages of commercial web sites and this approach is called as Fuzzy Temporal Clustering Approach FTCA. This approach not only does the clustering of users and pages, but also studies the cluster migration of the users and the pages over a period of time. In E-commerce, companies want to analyse the user's preferences to place advertisements, to decide their market strategy, and to provide customized guide to web customers [15]. This can be achieved using the FTCA. Analysis of web access logs of different web sites helps to understand user behaviours and web site structures. The outcome is to improve the design of web site structures. There are two main applications of Web Usage Mining – General Access Pattern Tracking, Customized Usage Tracking [18]. Analysing web logs can help to identify potential customers and to trace service quality etc in the E-commerce environment.

The rest of the paper is organized as various sections. Section II will detail on the related work that forms the basis for the proposed clustering approach. Section III elaborates on the background information upon which this current work is based. It explains in detail the Temporal Web Usage Mining and Fuzzy Clustering methods. Section IV discusses on the proposed approach FTCA in detail. Finally Section V demonstrates the experimental results and Section VI conclusion with future work.

II. Related Work

The proposed approach FTCA is based on two important works viz., "Temporal Cluster Migration Matrices for Web Usage Mining." by Lingras, Hogo and Snorek [3] and "Study on Web Mining Algorithm Based on Usage Mining" by Han, Gao and Wu [2]. The first approach is on temporal web logs clustering. Temporal web usage mining is the analysis of cluster behaviour over time and it can reveal additional information on the web site usage. There can be two different temporal changes in cluster analysis - change in cluster compositions and change in cluster memberships. TCMM – Temporal Cluster Migration Matrices is a framework useful for the analysis of changes in nature of the web site usage and loyalty of web site users. TCMM also serves as a visualization tool for analysis of results of temporal data mining. The second approach discussed is a Fuzzy Clustering Algorithm that produces the design mentality of the electronic commerce websites. This algorithm is simple, effective and easy to realize, it is suitable to the web usage mining demand of constructing a low cost B2C website. As a continuation of the previous work, "Web Logs Clustering Approaches – A Survey" [1], the use of fuzzy logic is combined along with temporal clustering to explore the interesting dimension of the change in web usage behaviours.

III. Background

Before going into the proposed approach, the two basic approaches considered for the proposed work are narrated. Temporal web usage mining is the analysis of cluster behaviour over time and it can reveal additional information on the web site usage. There can be two different temporal changes in cluster analysis namely, changes in cluster compositions and changes in cluster memberships. TCMM – Temporal cluster Migration Matrices is a framework useful for the analysis of changes in nature of the web site usage and loyalty of web site users. TCMM also serves as a visualization tool for analysis of results of temporal data mining. TCMM is constructed by repetitive application of clustering process for a sequence of time periods. The matrix stores the time sequences, the clustering analysis results like the changing nature of web site usage and the changing nature of individual users. TCMM can be considered as a relational table and can use SQL for analysis. Mass customization means, vendors can customize his business or product for individual customers. For web personalization we have to concentrate on individual visitors. The company may want to encourage loyal visitors by giving special offers. Another important marketing strategy is to detect the changes in customer loyalty. Business is worried about the attrition rate of their best customers [14].

It is assumed that the cluster labels for the users represent their desire to the business. Then customer's attrition rate will be evident from their increasing cluster labels during repetitive application of clustering. Such a customer may be a potential target for promotional material.

TCMM can be defined as $< H, C, n, \Omega: H \rightarrow C^n >$, where: H is the set of users, C is a set of cluster labels and n represents the number of time periods. $\Omega: H \rightarrow C^n$ is a mapping that describes the sequence of cluster memberships for each user over n time periods. Web site users are identified by a seven digit number: $H = \{h \mid 1000000 <= h <= 9999999\}$. The users are grouped into five clusters based on their visiting patterns in a month. These clusters can be named as Loyal big spenders (C = 1), Loyal moderate spenders (C = 2), Semi-loyal big spenders (C = 3), Semi loyal moderate spenders (C = 4) and Infrequent visitors (C = 5). Hence $C = \{1, 2, 3, 4, 5\}$ represents the cluster labels n = 6; that is cluster behaviour analysed for six months.

The steps in this approach are, to collect the web log data of a web site for six months, clean the web log data, format the web log data, identify users and find the number of pages visited by each user in each month. Based on the count of pages visited by each user, the users are categorized into different clusters identified above. This clustering is done for each month. From this the TCMM matrix is constructed. After this perform analysis on the TCMM table data using SQL to find the number of loyal big spenders in each month, to view the overall clustering for a given period, to list the customers who where throughout the study period and to detect the changes in the customer loyalty over the periods. The results can be shown pictorially using bar and pie charts for better visualization. To execute any analysis and find the result for any set of data, it will be efficient to embed the SQLs in a programming language such as Java. This will help the analysers to execute more generic and complex queries. TCMM can be applied for a retail ecommerce site / marketing databases.

In Fuzzy Clustering, as a result of applying web mining techniques, mainly clustering and classification, the user's interests or user community's interests are discovered and constructed an interest model. From the store house of interesting patterns, deleted the less significant patterns and then perform comprehensive analysis of the interesting patterns. Finally the discovered knowledge is applied to improve the service of E-commerce websites. Application Strategy of Web Usage Mining technology in E-Commerce, provides different marketing strategy for different customer communities in which each customer community has similar interests. It reduces the customers development cost by identifying the customer category which can be low value customers,

valuable customers and potential valuable customers [16]. This will help an organization to understand their customer's behaviour and analyse the main drawback in their business which drains the customers when compared to its competitors. It also maximizes the possibility of using the existing customers, make full use of the retrieval function module of a website which consists of the retrieval types, precise retrieval and fuzzy retrieval so as to achieve effective layout of the correlated products [21].

In Fuzzy Clustering Algorithm, web logs are collected and pre-treated and establish topology of the web site by finding V = {URL_1, URL_2, URL_n}, the set of all URLs and R = {<URL_1, URL_2>, <URL_2, URL_3>, …}, the ordered hyperlink set of pages. Then establish the matrix of users visiting pages as below.

$$B_{m\times n} = \begin{pmatrix} URL_1 \\ URL_2 \\ \cdots \\ URL_m \end{pmatrix} (User_1 \quad User_2 \quad \cdots \quad User_n) = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & & A_{21} \\ & \vdots & & \\ A_{21} & A_{22} & & A_{21} \end{pmatrix} \qquad (1)$$

This matrix can be considered as a relational table and SQL can be used to find the useful information like, the first N pages that are mostly, first N users that has the most visiting time, pages which the users are mostly interested in and the visit characteristics of specific users.

Calculate the time spent by every user visiting all the pages of the web site using the below formula (2).

$$S_i = \sum_{j=1}^{n} A_{ij} \qquad (2)$$

Calculate the rate of a user visiting all the pages of this website as per the below formula (3).

$$r_{ij} = A_{ij} \Big/ \sum_{j=1}^{n} S_j \qquad (3)$$

Using this compose a matrix R_{mxn} as below.

$$R_{m\times n} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ & \vdots & & \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix} \qquad (4)$$

Calculate the time the user spends in each page of this web site as per the formula - Settling time = End time – Begin time. Using this compose a matrix T_{mxn} as below.

$$T_{m\times n} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ & \vdots & & \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{pmatrix} \qquad (5)$$

Take the values from both the above matrices, R & T and arrange them in descending order of the values of rate and time. From this the pages which the users are mostly interested in is obtained. That is the visiting rate of these user interested pages are high and the settling time is also longer. The visit characteristics of specific users can be each users frequent access path and each user's frequent access time. This algorithm can be used for E-commerce websites. The advantages of this algorithm is it is simple, easy to apply and very effective. This algorithm can be used to construct an efficient, low cost E-commerce web site.

IV. APPROACH ELABORATION

First collect the web log files of the web site for a set of periods, say six months. Pre-process the web logs by cleaning and removing the unwanted requests. Split the field in the web logs and extract the data under the columns Date, Time, User Id and URL. A web log file is generally chronologically ordered and the User Id here refers to the individual Client IP address. After the splitting step a database of records with the above said fields is formed. This database is called as D1.

Now order this database records in the order of User Id, Date and then Time. To find the time spent by each user in a page on a given date and time, add another field named Time-Spent in the database records. This Time-Spent field takes the value by the below calculation. That is the Time-Spent field of a record takes the value that is equal to the difference of the Time fields of this record and its subsequent record. This calculation stands valid only when the User id's in the subsequent records are same and the Date's in the subsequent records are same and the time difference between the two records does not exceed five minutes (maximum idle time to recognize an users request) and the record is not the last record. In case of failing on any of the above criteria, the Time-Spent field takes the value 60 seconds. The pseudo code for calculating the Time-Spent field is given as below.

```
If [(Record1.Userid = Record2.Userid) and
    (Record1.Date = Record2.Date) and
    ((Record1.Time + 5 minutes) > Record2.Time)] Then
            {
            Record1.Time-spent = Record2.Time - Record1.Time;
            }
Else If [(Record1.Userid = Record2.Userid) and
          (Record1.Date = Record2.Date) and
          ((Record1.Time + 5 minutes) <= Record2.Time)] Then
            {
            Record1.Time-spent = 60;
            }
Else If [(Record1.Userid = Record2.Userid) and
          (Record1.Date <> Record2.Date)] Then
            {
            Record1.Time_spent = 60;
            }
Else If [(Record1.Userid <> Record2.Userid)] Then
            {
            Record1.Time-Spent = 60;
            }
Else If Record1 is the last record then
            {
            Record1.Time-Spent = 60;
            }
End If;
```

Note, generally the Time-Spent field is expressed in seconds. Now these database records can be grouped and stored in another database D2, which has the records made of the fields User-Id, URL, Visit-Frequency, Total-Time-Spent and Visit-Rate. Of these the fields User-Id and URL are taken from the previous database D1. But the field Visit-Frequency is the total number of records for a specific User-Id and URL in the database D1. Similarly, the field Total-Time-Spent is the total Time-Spent of records for a specific User-Id and URL in the database D1. The field Visit-Rate in each record of D2 is obtained by calculating the Visit-Frequency of the same record divided by the total visit frequencies of all the records of D2 for a specific User-Id. The pseudo code for the above said logic is given as below.

```
Create table D2 (User-Id, URL, Visit-Frequency, Time-Spent, Visit-Rate);
Insert into D2
        Select User-Id, URL, Count(*) Visit-Frequency, Sum(Time-Spent) Total-Time-Spent
        From D1
        Group By User-Id, URL;

Update D2 a
Set Visit-Rate =
        (Select a.Visit-Frequency / b.Total-Visit-Frequency
         From   (Select User-Id, Count(*) Total-Visit-Frequency
                 From D1
                 Group By User-Id) b
        Where b.User-Id = a.User-Id);
```

From the database D2, the visit frequency of a particular page by all users, the visit frequency of a particular user on all pages, the time spent on a particular page by all users, the time spent by a particular user on all pages are found out.

By analysing the two databases D1 and D2 and posting SQL queries on these databases, the below useful information that can help any E-Commerce Web site is found out.

- First N pages that are mostly accessed - This is obtained by sorting the URL's in the descending order of their Total Visit Frequency.
- First N pages that have been browsed for more time – This is obtained by sorting the URL's in the descending order of their Total Time Spent.
- First N pages that have been browsed for more time and that is mostly accessed – This is obtained by sorting the URL's in the descending order of their Total Time Spent and then by descending order of their Total Visit Frequency.
- First N users who access more pages - This is obtained by sorting the User ID's in the descending order of their Total Visit Frequency.
- First N users who browsed for more time – This is obtained by sorting the User ID's in the descending order of their Total Time Spent.
- First N users who have been browsing for more time and who access more pages – This is obtained by sorting the User ID's in the descending order of their Total Time Spent and then by descending order of their Total Visit Frequency.
- First N user, page combination who have the most visiting time (Time the User Spends in a Page) and the most visit rate (Time the User Spends in a Page) – This is obtained by sorting the User ID and URL pairs in the descending order of their Total Time Spent and then by descending order of their Total Visit Rate.

From the above it is also possible to arrive at the frequently accessed pages for each user. This is obtained by considering the User Page combination that has the highest "Time the User Spends in a Page" and "Rate of Users Visiting Pages". Note that these SQL queries can be embedded in a programming language like JAVA for less processing time and for handling more complex queries.

So, far this approach has been discussing about data for a specific period, say a month. Now coming to the temporal part of this new approach FTCA, it is possible to collect data of the similar type over a period of time, say six months. The users and pages can be clustered into the different cluster categories in the different months and their migration among the different clusters can be studied. The User Clusters are defined as the Most Loyal User (C = 1), Loyal User (C = 2), Most Frequent User (C = 3), Frequent User (C = 4) and Least Frequent User (C = 5). Similarly, let the clusters for the Pages be Most Popular Page (C = 1), Popular Page (C = 2), Most Favourite Page (C = 3), Favourite Page (C = 4) and Least Favourite Page (C = 5).

TABLE I
User Clusters

User Clusters	
Cluster Name	**Cluster Number**
Most Loyal User	1
Loyal User	2
Most Frequent User	3
Frequent User	4
Least Frequent User	5

TABLE II
Page Clusters

Page Clusters	
Cluster Name	**Cluster Number**
Most Popular Page	1
Popular Page	2
Most Favourite Page	3
Favourite Page	4
Least Favourite Page	5

The clusters are formed for each month based on the data for that particular month. Create the TCMM for the User Clusters and the Page Clusters. The TCMM can be used to study the changing nature of web usage. Convert this TCMM into a database of records and can perform several analysis by posting SQL statements. The result of these queries can be represented pictorially using bar graph or pie graph.

Some of the example SQL statements that can be posted on these TCMM that provides useful knowledge for web usage mining of ecommerce web sites are listed below.

- Number of Most Loyal Users in each Month
- Number of Most Popular Page in each Month
- Snapshot of overall User Clustering for a given Month
- Snapshot of overall Page Clustering of a given Month
- Users who are Loyal throughout the study period
- Pages that are Popular throughout the study period
- User who's Loyalty declined from one month to next month

For analysing more complex queries, the approach can be automated and the required SQL's can be embedded in a programming language like JAVA and the results can be visually presented using graphs and TCMM tables.

V. Experimental Evaluation

To understand the proposed approach FTCA, consider the example web site with the below structure. Say there are nine pages, namely, V = {P1, P2, P3, P4, P5, P6, P7, P8, P9}. The topology of the web site can be said as below.

R = { P1 → P2, P1 → P3, P1 → P4, P2 → P7, P1 → P2 → P7, P3 → P5, P1 → P3 → P5, P3 → P6, P1 → P3 → P6, P4 → P8, P8 → P9, P1 → P4 → P8, P1 → P4 → P8 → P9, P4 → P8 → P9}

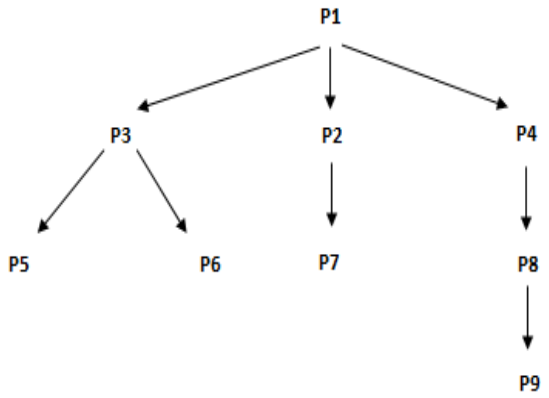

Fig. 1. A sample web site structure

This example web site is a simulation of the eretailstore website which was taken as the based for studying our proposed approach. Web logs for this website were collected for a six months period, via, 01/07/2010 to 31/12/2010. Now, perform pre-processing of these web logs and create the databases D1 and D2 as specified in the previous section. The sample records of the database D1 can be seen as below.

TABLE III
Sample Records of the Database D1

Date	Time	User-Id	URL	Time-Spent
13/07/10	16:44:30	U1	P1	10
13/07/10	16:44:40	U1	P2	23
13/07/10	16:45:03	U1	P7	56
…				
21/07/10	08:15:32	U2	P1	214
21/07/10	08:19:06	U2	P3	48
21/07/10	08:19:54	U2	P6	18
…				
08/07/10	11:02:32	U3	P1	46
08/07/10	11:03:18	U3	P2	39
08/07/10	11:03:57	U3	P7	154
…				
…				
23/07/10	00:05:15	U23	P1	46
23/07/10	00:06:01	U23	P2	39
23/07/10	00:06:40	U23	P7	154

From the database D2, the tables of data as shown in the below samples are obtained.

	Visit Frequency of Users on the Pages																							
	Users																							Visit Frequency of the Page by all
Web Pages / URLs	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15	U16	U17	U18	U19	U20	U21	U22	U23	
P1	5	3	2	1	3	2	3	3	3	6	8	5	3	4	7	7	1	1	2	2	7	5	5	88
P2	4	0	2	0	2	1	4	2	2	4	6	4	0	2	2	3	0	3	0	0	2	3	4	50
P3	4	5	0	1	0	5	0	3	3	6	4	0	10	1	5	2	0	0	3	3	5	1	0	61
P4	2	2	2	1	1	0	4	1	1	2	3	3	0	2	2	2	2	0	2	2	2	2	3	41
P5	1	1	0	0	0	2	0	1	1	2	1	0	5	0	2	0	0	0	1	1	2	0	0	20
P6	1	2	0	2	0	2	0	1	1	2	1	0	3	0	0	0	0	0	1	1	0	0	0	17
P7	2	0	1	4	1	1	2	1	1	2	3	2	0	1	0	1	0	2	0	0	0	1	2	27
P8	2	1	2	2	1	0	7	1	1	2	3	3	0	1	0	0	4	0	2	3	0	0	3	38
P9	1	0	0	2	0	0	3	0	1	1	1	0	0	1	0	0	2	0	1	2	0	0	0	15
Visit Frequency of the User on all Pages	22	14	9	13	8	13	23	13	14	27	30	17	21	12	18	15	9	6	12	14	18	12	17	

Fig. 2. A sample table that shows visit frequency of users on the pages

	Time the User Spends on the Pages																							
Web Pages / URLs	Users																							Time Spent on the Page
	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15	U16	U17	U18	U19	U20	U21	U22	U23	
P1	64	257	59	54	66	153	657	220	220	440	130	125	114	182	175	259	71	111	80	90	175	220	125	4047
P2	67	0	62	0	0	31	858	154	154	308	67	62	62	135	50	118	0	423	0	0	50	126	62	2789
P3	377	188	0	36	207	576	0	99	99	198	584	207	690	78	148	78	0	0	120	135	148	48	207	4223
P4	300	214	340	54	0	0	268	77	77	154	300	340	0	172	54	70	102	0	110	210	54	96	340	3332
P5	102	351	0	0	191	109	0	22	22	44	293	191	300	0	68	0	0	0	30	35	68	0	191	2017
P6	89	392	0	112	110	192	0	44	44	88	199	110	220	0	0	0	0	0	50	55	0	0	110	1815
P7	151	0	154	234	0	22	456	77	77	154	151	154	0	29	0	45	0	282	0	0	0	98	154	2238
P8	99	93	123	78	0	0	512	88	88	176	99	123	0	91	0	0	284	0	110	345	0	0	123	2432
P9	321	0	0	69	0	0	366	0	46	46	321	0	0	50	0	0	102	0	10	250	0	0	0	1581
Time Spent by the User on all Pages	1570	1495	738	637	574	1083	3117	781	827	1608	2144	1312	1386	737	495	570	559	816	510	1120	495	588	1312	

Fig. 3. A sample table that shows the time the user spends on the pages

	Rate of Users Visiting all the Pages																						
Web Pages / URLs	Users																						
	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15	U16	U17	U18	U19	U20	U21	U22	U23
P1	0.23	0.21	0.22	0.08	0.38	0.15	0.13	0.23	0.21	0.22	0.27	0.29	0.14	0.33	0.39	0.47	0.11	0.17	0.17	0.14	0.39	0.42	0.29
P2	0.18	0.00	0.22	0.00	0.25	0.08	0.17	0.15	0.14	0.15	0.20	0.24	0.00	0.17	0.11	0.20	0.00	0.50	0.00	0.00	0.11	0.25	0.24
P3	0.18	0.36	0.00	0.08	0.00	0.38	0.00	0.23	0.21	0.22	0.13	0.00	0.48	0.08	0.28	0.13	0.00	0.00	0.25	0.21	0.28	0.08	0.00
P4	0.09	0.14	0.22	0.08	0.13	0.00	0.17	0.08	0.07	0.07	0.10	0.18	0.00	0.17	0.11	0.13	0.22	0.00	0.17	0.14	0.11	0.17	0.18
P5	0.05	0.07	0.00	0.00	0.00	0.15	0.00	0.08	0.07	0.07	0.03	0.00	0.24	0.00	0.11	0.00	0.00	0.00	0.08	0.07	0.11	0.00	0.00
P6	0.05	0.14	0.00	0.15	0.00	0.15	0.00	0.08	0.07	0.07	0.03	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.08	0.07	0.00	0.00	0.00
P7	0.09	0.00	0.11	0.31	0.13	0.08	0.09	0.08	0.07	0.07	0.10	0.12	0.00	0.08	0.00	0.07	0.00	0.33	0.00	0.00	0.00	0.08	0.12
P8	0.09	0.07	0.22	0.15	0.13	0.00	0.30	0.08	0.07	0.07	0.10	0.18	0.00	0.08	0.00	0.00	0.44	0.00	0.17	0.21	0.00	0.00	0.18
P9	0.05	0.00	0.00	0.15	0.00	0.00	0.13	0.00	0.07	0.04	0.03	0.00	0.00	0.08	0.00	0.00	0.22	0.00	0.08	0.14	0.00	0.00	0.00

Fig. 4. A sample table that shows the rate of users visiting all the pages

By analysing the two databases and posting SQL queries on these databases, the below useful information that can help any E-Commerce Web site can be obtained.

- First N pages that are mostly accessed.

TABLE III
Result of first N mostly accessed pages

URL	Visit Frequency
P1	88
P3	61
P2	50
P4	41
P8	38
P7	27
P5	20
P6	17
P9	15

- First N pages that have been browsed for more time.

TABLE V
Result of first N pages browsed for more time

URL	Total Time Spent
P3	4223
P1	4047
P4	3332
P2	2789
P8	2432
P7	2238
P5	2017
P6	1815
P9	1581

- First N pages that have been browsed for more time and that is mostly accessed.

TABLE VI
Result of first N pages browsed for more time and most frequently accessed

URL	Total Visit Frequency	Total Time Spent
P3	61	4223
P1	88	4047
P4	41	3332
P2	50	2789
P8	38	2432
P7	27	2238
P5	20	2017
P6	17	1815
P9	15	1581

- First N users who access more pages.

TABLE VII
Result of first N users who access more pages

User Id	Total Visit Frequency
U11	30
U10	27
U7	23
U1	22
...	...
...	...
U3	9
U17	9
U5	8
U18	6

- First N users who browsed for more time.

TABLE VIII
Result of first N users who browsed for more time

User Id	Total Time Spent
U7	3117
U11	2144
U10	1608
U1	1570
...	...
...	...
U17	559
U19	510
U15	495
U21	495

- First N users who have been browsing for more time and who access more pages.

TABLE IX
Result of first N users who browsed for more time and who accessed more pages

User Id	Total Visit Frequency	Total Time Spent
U7	23	3117
U11	30	2144
U10	27	1608
U1	22	1570
...	...	...
...	...	...
U17	9	559
U19	12	510
U15	18	495
U21	18	495

- First N user, page combination who have the most visiting time (Time the User Spends in a Page) and the most visit rate (Time the User Spends in a Page).

TABLE X
Result of first N users, pages combination with most visiting time and most visiting rate

User Id	URL	Rate of Users Visiting Pages	Time the User Spends in a Page
U10	P8	0.19	421
U16	P3	0.33	411
U10	P2	0.19	411
U20	P4	0.36	382
…	…	…	…
U9	P1	0.07	18
U20	P9	0.07	18
U17	P2	0.11	17
U16	P1	0.07	16
U1	P4	0.00	0
U1	P6	0.00	0
…	…	…	…
U23	P5	0.00	0
U23	P7	0.00	0

From the above the frequently accessed pages for each user can also be arrived at. Data of similar type is collected for a period of say six months. The sample TCMM for the User Clusters and Page Clusters for six months can be obtained as below. This clustering is based on the User and Page Clustering definition specified in the previous section.

TABLE XI
Sample TCMM for user clusters for six months

User Id	Month1	Month2	Month3	Month4	Month5	Month6
U1	3	1	2	4	1	3
U2	4	3	5	1	3	4
U3	5	5	2	3	2	3
…	…	…	…	…	…	…
U12	4	2	3	5	5	1
U13	4	1	3	2	5	4
U14	5	4	3	3	2	2
…	…	…	…	…	…	…
U21	5	4	3	3	3	2
U22	5	3	2	1	1	1
U23	4	2	1	2	3	4

TABLE XII
Sample TCMM for page clusters for six months

URL	Month1	Month2	Month3	Month4	Month5	Month6
P1	1	2	1	1	2	1
P2	3	1	2	1	3	1
P3	1	2	1	2	3	4
P4	2	5	5	1	1	3
P5	5	3	1	4	4	5
P6	5	4	5	1	1	2
P7	4	3	3	4	4	5
P8	4	5	2	3	1	2
P9	5	1	3	2	3	1

These User TCMM and Page TCMM are converted into database of records and SQL statements are posted. The result of these queries is represented pictorially using bar graph or pie graph.

Some of the example SQL statements that are posted on these TCMM that provides useful knowledge for web usage mining of ecommerce web sites are listed below.

- Number of Most Loyal Users in each Month.
  Select Count(User-Id) From User-TCMM Where Month1 = 1;

Select Count(User-Id) From User-TCMM Where Month2 = 1;
Select Count(User-Id) From User-TCMM Where Month3 = 1;
Select Count(User-Id) From User-TCMM Where Month4 = 1;
Select Count(User-Id) From User-TCMM Where Month5 = 1;
Select Count(User-Id) From User-TCMM Where Month6 = 1;

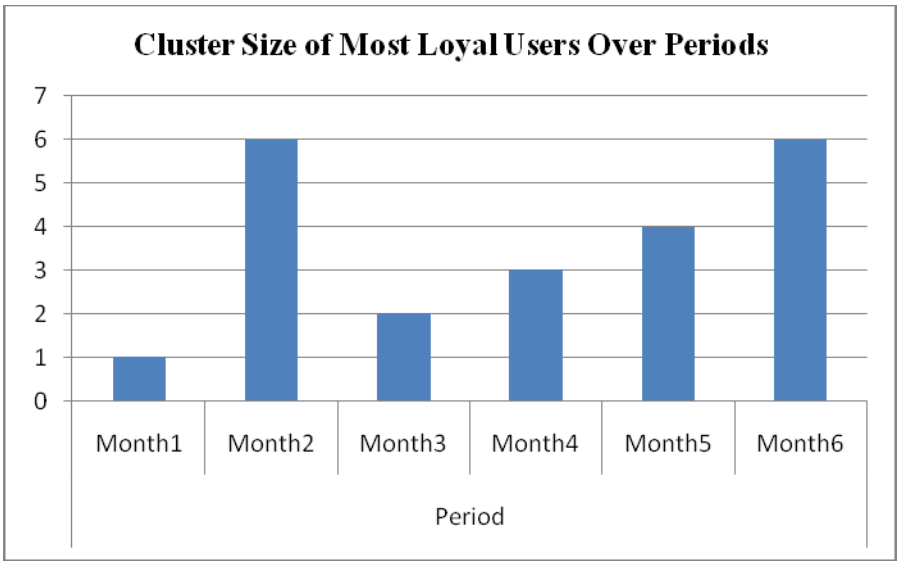

Fig. 5. Graph showing the cluster size of most loyal users over periods

- Number of Most Popular Page in each Month.
  Select Count(URL) From Page-TCMM Where Month1 = 1;
  Select Count(URL) From Page-TCMM Where Month2 = 1;
  Select Count(URL) From Page-TCMM Where Month3 = 1;
  Select Count(URL) From Page-TCMM Where Month4 = 1;
  Select Count(URL) From Page-TCMM Where Month5 = 1;
  Select Count(URL) From Page-TCMM Where Month6 = 1;

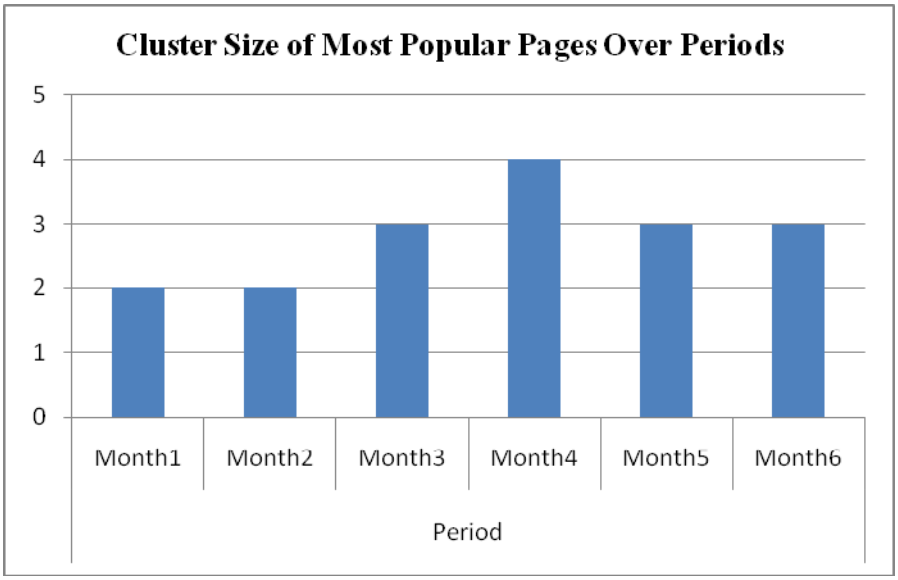

Fig. 6. Graph showing the cluster size of most popular pages over periods

- Snapshot of overall User Clustering for a given Month (Month4).
  Select Count(User-Id) From User-TCMM Where Month4 = 1;
  Select Count(User-Id) From User-TCMM Where Month4 = 2;
  Select Count(User-Id) From User-TCMM Where Month4 = 3;
  Select Count(User-Id) From User-TCMM Where Month4 = 4;

Select Count(User-Id) From User-TCMM Where Month4 = 5;

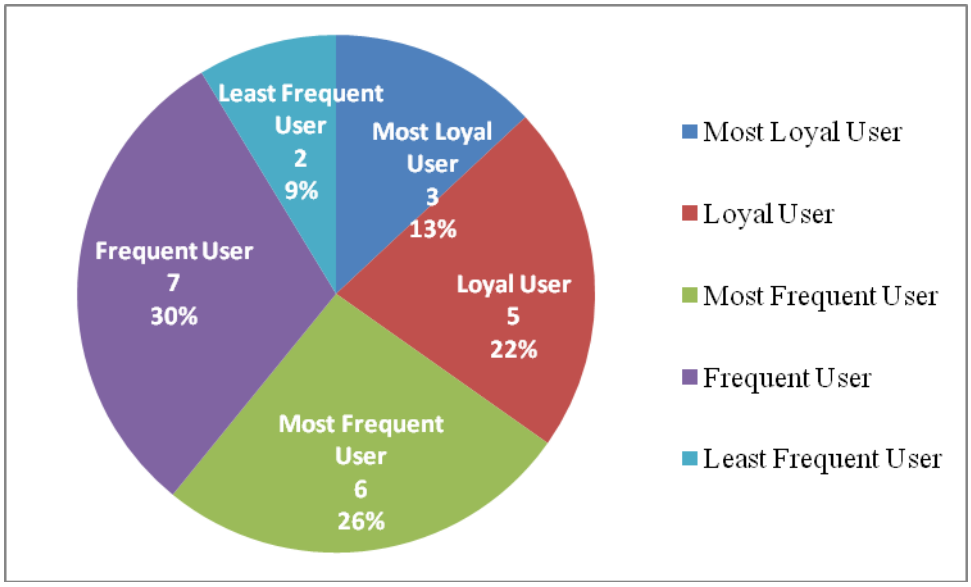

Fig. 7. Graph showing various users clustering for a particular month

- Snapshot of overall Page Clustering of a given Month (Month4).
  Select Count(URL) From Page-TCMM Where Month4 = 1;
  Select Count(URL) From Page-TCMM Where Month4 = 2;
  Select Count(URL) From Page-TCMM Where Month4 = 3;
  Select Count(URL) From Page-TCMM Where Month4 = 4;
  Select Count(URL) From Page-TCMM Where Month4 = 5;

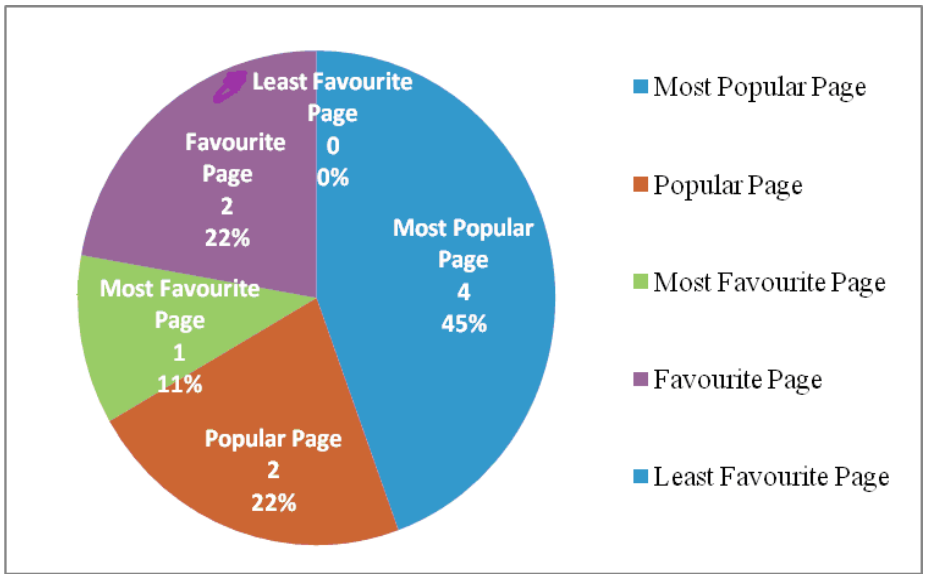

Fig. 8. Graph showing various pages clustering for a particular month

- Users who are Loyal throughout the study period.
  Select User-Id
  From User-TCMM
  Where (Month1 = 1 OR Month1 = 2) AND (Month2 = 1 OR Month2 = 2)
  AND (Month3 = 1 OR Month3 = 2) AND (Month4 = 1 OR Month4 = 2)
  AND (Month5 = 1 OR Month5 = 2) AND (Month6 = 1 OR Month6 = 2);
  The result is the users U7 and U11.

- Pages that are Popular throughout the study period.
  Select URL
  From Page-TCMM
  Where (Month1 = 1 OR Month1 = 2) AND (Month2 = 1 OR Month2 = 2)
  AND (Month3 = 1 OR Month3 = 2) AND (Month4 = 1 OR Month4 = 2)
  AND (Month5 = 1 OR Month5 = 2) AND (Month6 = 1 OR Month6 = 2);

  The result is the page P1.

- User who's Loyalty declined from Month5 to Month6.
  Select User-Id from User-TCMM where (Month5 < Month6);

  The result is the users U1, U2, U3, U6, U17, U18, U19 and U23.

VI. CONCLUSION

Thus this proposed work is a hybrid combination of the Fuzzy Clustering and Temporal Clustering methods. But when these two methods are combined, the benefit of the resultant knowledge is multi-fold and provides almost all required knowledge for enhancing E-commerce websites. The importance of web usage mining is unquestionable with the rising importance of the web not only as an information portal but also as a business edge. Web access logs contain abundant raw data that can be mined for web access patterns, which in turn can be applied to improve the overall surfing experience of users. By taking this into consideration it is mainly focused on clustering the web users and the web pages so that the users preferences can be studied from time to time and enhance the web site and the business as per the changing trends of the end users. Experiments conducted on web logs show the viability of this approach. However, there is scope for future work in this to add more functionality to web mining services and to make web usage mining more useful in the electronic commerce domain.

REFERENCES

[1] Sudhamathy, G., & Jothi Venkateswaran, C. (2011). "Web Log Clustering Approaches – A Survey". *International Journal on Computer Science and Engineering, 3*(7), 2896-2903.

[2] Han, Q., Gao, X., & Wu, W. (2008). "Study on Web Mining Algorithm Based on Usage Mining." In 9th International Conference on *Computer-Aided Industrial Design and Conceptual Design, CAID/CD*.

[3] Lingras, P., Hogo, M., & Snorek, M. (2004). "Temporal Cluster Migration Matrices for Web Usage Mining." In Proceedings of *IEEE/WIC/ACM International Conference on Web Intelligence*.

[4] Hogo, M., Snorek, M., & Lingras, P. (2003). "Temporal Web Usage Mining." In Proceedings of International Conference on *Web Intelligence, IEEE/WIC* (pp. 450–453).

[5] Antunes, C., & Oliveira, A. (2001). "Temporal Data Mining: An Overview." In Proceedings of *KDD 2001 Workshop on Temporal Data Mining*, http://www.acm.org/sigkdd/kdd2001/Workshops/ano.pdf.

[6] Mobasher, B., Cooley, R., Srivastava, J. (2000). "Automatic Personalization Based on Web Usage Mining." *Communications of the ACM, 43*(8), 142-151.

[7] Perkowitz, M., & Etzioni, O,. (1999). "Adaptive web sites: Conceptual cluster mining." In Proceedings of the Sixteenth International *Joint Conference on Artificial Intelligence*.

[8] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). "Web Usage Mining, Discovery and Applications of Usage Patterns from Web Data", in *SIGKDD Explorations, 1*(2), 1-12.

[9] Cooley, R., Mobasher B., & Srivastava, J. (1997). "Web Mining: Information and Pattern Discovery on the World Wide Web." In Proceedings of the *9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*.

[10] Zaiane, O. R., Xin M., & Han, J. (1998). "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs". *Advances in Digital Libraries Con, Santa Barbara, CA*, (pp.19-29).

[11] Cooley, R., Mobasher, B., & Srivastava, J. (1999). "Data preparation for mining world wide web browsing patterns". *Knowledge and Information System*, (pp. 5-32).

[12] Xinlin, Z., & Xiangdong, Y. (2008). "Design of an Information Intelligent System based on Web Data Mining", *IEEE International Conference on Computer Science and Information Technology*. (pp. 88-91).

[13] Chu-Hui, L., & Yu-Hsiang, F. (2008). "Web Usage Mining Based on Clustering of Browsing Features", *IEEE Eighth International Conference on Intelligent Systems Design and Applications.* (pp. 281-286).

[14] Zhang, Y., & Jiao, J. (2007). "An associative classification-based recommendation system for personalization in b2c e-commerce applications." *Expert Systems with Applications, 33*. (pp. 357–367).

[15] Thorleuchter, D., Poel, D. V. D., & Prinzie, A. (2012). "Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing." *Expert Systems with Applications, 39*. (pp. 2597–2605).

[16] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). "Web usage mining: Discovery and applications of usage patterns from web data." *SIGKDD Explorations.* (pp. 12–23).

[17] Facca, F. M., & Lanzi, P. L. (2005). "Mining Interesting Knowledge from Weblogs: A Survey". *Data Knowledge Engineering, 53*(3). (pp. 225–241).

[18] Kim, J. K., Cho, Y. H., Kim, W. J., Kim, J. R., & Suh, J. H. (2002). "A personalized recommendation procedure for Internet shopping support." *Electronic Commerce Research and Applications, 1*. (pp. 301–313).

[19] Fenstermacher, K.D., & Ginsburg, M. (2002). "Mining client-side activity for personalization". In Fourth *IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS 02)*. (pp. 205–212).

[20] Berendt, B., Mobasher, B., Nakagawa, M., & Spiliopoulou, M. (2002). "The impact of site structure and user environment on session reconstruction in web usage analysis". In Proceedings of the *4th WebKDD 2002 Workshop, at the ACMSIGKDD Conference on Knowledge Discovery in Databases.*

[21] Ansari, S., Kohavi, R., Mason, L., & Zheng, Z. (2000). "Integrating e-commerce and data mining: Architecture and challenges". In *WEBKDD 2000—Web Mining for E-Commerce—Challenges and Opportunities, Second International Workshop.*