IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 7, Issue 2 (Nov-Dec. 2012), PP 09-15
www.iosrjournals.org

Cloud Mining: Web usage mining and user behavior analysis using fuzzy C-means clustering

M.Rathamani[1], Dr. P.Sivaprakasam[2]
[1](Assistant Professor, Nallamuthu Gounder Mahalingam College, India)
[2](Reader, Sri Vasavi College, Erode, India)

Abstract:*There is a rapid development of World Wide Web in its volume of traffic and the size and complexity of web sites. In this paper, a new approach is presented based on hybrid clustering methods for Web Usage Mining (WUM). The WUM process contains three steps: pre-processing, data mining and result analysis. First, it gives a brief description of the WUM process and Web data, then the presentation of the pre-processing step and the data warehouse that were employed. The hybrid clustering methods based on Fuzzy C-means clustering are used for analyzing and the Web logs taken from the real world Web servers. The results obtained after applying these methods and the corresponding interpretations are also presented. Furthermore, this paper also described web usage mining through cloud computing i.e. cloud mining. The Future work of web mining is to introduce a hierarchy on the information about the website.*

Keywords: *Clustering, Web Usage Mining, WWW, Web logs, pre-processing, Data Warehouse, Cloud Computing.*

I. Introduction

The development of the Web that occurred in the recent years generated a boom of data related to its activities. To analyze (or rather excavate) these new types of data new methods appeared and were grouped under the generic term of "Web Mining".

Web usage mining is an active, technique used in this field of research. It is also called web log mining in which data mining techniques are applied to web access log. Web mining [2] is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user's visiting behaviors and extracts their interests using patterns. Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval etc., web mining [1] has become one of the important areas in computer and information science. Web Usage Mining [3] uses mining methods in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, prefetching , creating attractive web sites etc.,

a. Web Usage Mining

This study (the WUM) of the users' navigations extracted from the web server's log files or proprietary traces may help the webmaster to understand the user behavior and then to rethink the structure and design of his/her website or to detect users' problems and improve the navigability. The WUM analysis, allows the webmaster to optimize the response of the Web server (Web caching) and to make recommendations to the user.

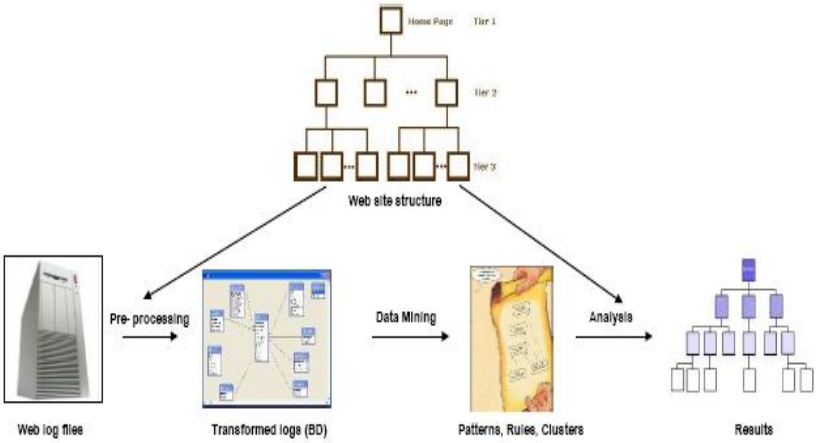

Fig. 1. The WUM schema

A Web Usage Mining process is commonly split in three phases: preprocessing, data mining and results analysis (Fig. 1).

b. Web Usage Data

The Web log file is the input data in the Web Usage Mining process. The Web site structure (hyperlinks graph) and the users' profiles may constitute supplementary data for such a process. To solve the problem of organizing these various and bulky data a database structure is necessary. More precisely, this should be a data warehouse as its characteristics (orientation, subject, integration, history and non-volatility) are useful (Kimball, 2001)[4]. The data warehouse is developed for decision purposes. The data warehouse is feed by mechanisms of extracting, transforming and loading data from the log files.

c. HTTP Log Files

According to the HTTP client-server protocol, the client accessing a resource will send a request to the server containing this resource. At the other side, Web server interprets the HTTP request, accesses the requested resource and delivers it to the client. As most of the software programs, these operations are recorded in a log file. The log file allows having a detailed trace of the Web server activity. By using the ECLF log file format for the HTTP log files as described in (Luotonen, 1995) [5].

A. Pre-processing

The objective of the pre-processing step is to identify and structure user navigations. This step is based on two main processes: data cleaning and data transformation. At the end of this phase, the Web logs will be placed in a database warehouse.

The preprocessing process consist of following steps

- Pre-processing of the Web log files Data Cleaning
- Data Transformation
- User/Session Identification

a. Pre-processing of the Web log files Data Cleaning

Log files have useful information about access of all users to a specific website. Extracting the information, reformatted log file which contains useful information such as “time, date, accessed URL and IP address” is formed and useless requests such as accesses to images are removed from log file in data cleaning process. Identifying Web robots and deleting the requests coming from these robots is another task of this process.

b. Data Transformation

To perform complex analysis, including clustering, grouping together several requests. All the requests made by a single user during the analyzed period constitute his/her session. A session is further split in several navigations, each navigation representing a single visit to the Web site. A navigation ends when a time threshold of at least 30 minutes exists between two consecutive requests.

c. User/Session Identification

Identifying users/sessions from the log file is difficult task because of several factors like: proxy servers, dynamic addresses, and the case of two or more users using the same computer or the same user that uses more than one browser or computer. In fact, by using the log file user can know only the computer's address and the User Agent of the user. This is not sufficient in most of the cases that is why there are other methods that can provide more information. The most used are: cookies, dynamic Web pages (with a session ID in the URL), registered users, modified browsers etc. In (Cooley, 2000) [6] the author differentiated users by their navigation.

B. Building a WUM Data Warehouse

The Web Usage data can be used to populate the database (DB). The DB is a lasting data warehouse where all the viewpoints are kept. It is different form all other DBs presented in the literature as they are more short-lived or more specialized**.**

a. The facts

The facts came from the log files filtered, ordered and enriched with data calculated as we have presented in the previous sections. The study is centered on the user, therefore the fields are mainly the duration and the size read when he/she accesses a Web page.

b. The dimensions

The selected dimensions can be split up into:

Dimension related to the URL and the page viewed. It is the view Content. Many classifications are possible, just from the file extension but also from other information from the Web site.

Dimension related to the date. It is the view Access regularity. The usual hierarchy second, minute, hour, day, month, and year can be used as well as any other hierarchy with more specialized periods. Dimension related to the session and the user. It is the view User. A session is issued from the user (when he/she stops requesting for a relatively long time). In our context the user belongs to the following hierarchy: domain → research unit → research team/service. But others classifications are allowed based on different criterions as we shall see further. The description may include the relevant class with this meaning. Dimension related to the referrer and the navigation. It is the view Navigation. The referrer allows mapping out the user navigation which may be linear or with many returns. To help the user in a collaborative view on his/her search in the web site it is essential to classify these navigations. Dimension related to the transaction status. It is the view Access efficiency. The server log files give the result as status success, failure, redirection, forbidden access. Here, once again, also define another status type within the context of the request emission rather than the request treatment.

II. Literature Survey

Data mining contains a number of clustering, but only few are used in the Web Usage Mining: BIRCH in [7], CLIQUE in [8], EM in [9]. A reason is that there is the difficulty of adapting these methods to the particularities of Web data: the big size of the data or the large number of variables (Web pages).In (Mobasher, 2002) [10], the authors compare clustering methods, but they don't take the order of the requests into account. The sessions (transactions in this paper) are represented by binary vectors balanced by page views. The weight of a page can be the value of time function or a function that considers the type of the page. PACT, the clustering method, groups the sessions of the users. First, a similarity function is applied.

In [7], the sessions are generalized by means of an induction based on the attributes. This induction reduces the dimensions of the data. For instance, www-sop/axis/teaching/stid-projet2.html is organized as a hierarchy such as the following: www-sop! axis! teaching ! stidprojet2. html. The authors use BIRCH (Zhang, 1996) to cluster the generalized data. The method has been tested on the log of the UMR's server. The log contained 2.5 millions lines, but the tests were conducted on 500 000 lines. The number of users was 26 107 and the number of pages was 21 203. However, upon the authors, BIRCH gets less efficient when data dimension is increased, so the generalization is limited to a few levels.

III. Comparison With Existing Works

Web Usage data are often used for Web site access statistics or for forecast of requested pages. In order to do this, the data are filtered and then organized and stored according to two essential ways: Graphs and trees are used when complex navigation models must be processed. For example WUM [7] (Web Utilization Miner) uses weighted aggregation trees to represent the navigation traffic along roads corresponding to the logical structure of the Web site. WUM proposes a language named MINT with a syntax close to SQL in order to make requests about the routes in the navigation tree. N-dimensional vectors are also used when the space of navigation is well known. WEB Miner [10] represents a transaction as a vector in the space of the reachable pages. On the other hand, in this work, some other information about users and documents are used for the analysis. There is a general request language to access the data but then different structures are used according to the goal of the analysis.

Data are only stored in files or relational tables. This is the case for many widespread software for web traffic analysis. Files and tables are often at and basic and they are hidden from the scene. Such representations can be better developed and can federate all the aspects of the web usage data in the same structure avoiding the many representations of the precedent method (one structure for navigation, one for user profile, etc.).

In [11], the authors implement a data warehouse as a relational database with Microsoft SQL Server. The goal is to optimise memory caching thanks to data analysis. Facts are from log files and with calculated fields as Last Access and Next Access, the distances (numbers of lines in the log file) between the current access to the same URL and the last or the next one. The field Page Delay is the number of seconds from the last request for a page to the current request for a page issued from the same user. Thanks to this field the pages are divided into links and contents, the first ones being read quickly and the others slowly. For us this is related to a visit and it is different from our notion of session which is larger. The dimensions are not in specific tables; only to the dimension URL is given an ID in addition to its character string because strings are very slow to process. Other information from the Web site are added; for example the depth of the requested file in the logical structure of the site. At last, they point out two interesting things: the transactions temporal ordering and the field Class (in the table Web Log) which can be later valued to classify the transactions. It also uses temporal

ordering and the results issued of our classification can be XML coded and placed in a field of a dimension table.

Web Log Miner [12] is a web usage mining software using a data and building a multi-dimensional cube to apply OLAP techniques. Among these dimensions, there are the ones found in the log files (URL, type and size transmitted, date, duration, user, agent, domain, status) and two more dimensions Field Size and Event. Field Size is related to the structure of the site and Event is related to a typical action of the user (use a V-group add a message, read a message). It is very close to our concerns to describe the site and to have a better tracking of the user actions.

IV. Methodology

After the pre-processing step, the data from the log file is structured in sessions and navigations and stored in a DB. This step objectify is to discover different types of user behaviors or categories of user behaviors using different approaches [13], sequences analysis algorithms, cluster analysis, predictive models, neural networks and automatic clustering. The objective is to develop a strategy which analyses the relations between the structure of the Web site and the log file. To reach it, applying hybrid clustering methods on different types of Web data (continuous and qualitative).

Document Clustering

In this work, content mining is used approach for document clustering. Assume G = { g_1, g_2,..,g_n }is the set of n website's pages. Applying the clustering algorithm shown in following Document clustering steps, pages were grouped in content based clusters.

1. Clear each document from stop words such as: about, all , am, almost, as, be, by, but, do and any other word which haven't any key role in determining the content of document.
2. Identify document keywords by TF-IDF technique.
3. Assign each document keyword list as a document to a single cluster.
4. Merge primary clusters based on the Jaccard coefficient similarity measure.

Defined as:

$$sim\left(g_x, g_y\right) = \frac{|g_x \cap g_y|}{|g_x \cup g_y|} \qquad (1)$$

$|g_x \cap g_y|$ Represents the number of common words and $|g_x \cup g_y|$ represents total number of words between two basic clusters.

5. The second step repeated until all documents being clustered into a pre defined number of clusters. DC = {DC1, DC2,…, DC_n} is the result set. Each DC_i represents a set of URLs with similar content.

FCM

The fuzzy C-means algorithm requires good initialization. These initial values are provided by the ant based algorithm. The result will be small homogenous heaps that will be merged by repeating the steps. By increasing the number of iterations the number of heaps decreases. The User clustering algorithm shows the hybrid algorithm used in this study to cluster users in appropriate groups:

6. Scatter the users randomly on the board
7. Use the cluster centers obtained in step 3 to initialize cluster centers for the fuzzy C-means algorithm
8. Cluster the data using the fuzzy C-means algorithm
9. Harden the data obtained from the Fuzzy C-means algorithm, using the maximum membership criterion, to form new heaps
10. Repeat step 1-6 by considering each heap as a single object

When a new user starts a transaction, our model matches the new user with the most similar user clusters and provides suitable recommendations for him/her. Support value which was calculated through following steps [2]:

Step1. Assign active user to a new cluster. UC_{new}= {U_{new}}.

Step2. Calculate support value of UC_{new} to existing user clusters UC_i using Equation 2

$$support(UC_{new}, UC_i) = \frac{\sum|user(val(DC, UC_{new}) - val(DC, UC_i))|}{UC_i \cup UC_{new}} \qquad (2)$$

Val (DC, UC_i) shows the interesting value of users in UC_i to the documents in selected DC.

Match Score Identification

The match score calculation defines highest match user cluster for active user. This parameter between UCnew and UCi is defined as following.

$$match(UC_{new}, UC_i) = 1 - support\ (UC_{new}, UC_i) \quad (3)$$

This give us a list of corresponding user clusters from the highest match score down to lowest match score.

V. Experimental Results

A. Application on real web server Dataset

To test proposed approach, data is chosen from the real world reputed web server. There are many links between these Web servers and one of the purposes of our analysis was to detect groups of topics consulted together.

B. Raw Data

The Web logs were collected for a period of two weeks, between the 1st and 15th of January 2009. The data summary is presented in Table. 1.

Table 1.Data Summary for Proposed Experiment

Web servers	Reputed web server
Period	1 - 15 January 2009
Number of requests	6 040 312
Requests after pre-processing	673 389
Number of sessions	115 825
Number of navigation	174 015

After data cleaning in preprocessing step, the number of requests was 52322 which were structured to 12332 sessions. The number of accessed URLs in this website was 200 pages. Employing content based document clustering algorithm we grouped the URLs to 5 clusters. Then user's behavior modeled as access matrix. Using equation 2, interest degrees are calculated as a
12332×5 matrix which is shown in Table 2, this is the input of user clustering algorithm.

Table2 Access table

IP	DC1	DC2	DC3	DC4	DC5
65.54	0.6	0	0.4	0	0
77.38	0.3	0	0.3	0	0.3
65.247	0.2	0.2	0	0.5	0.1

Applying FCM and compound ant based clustering algorithm on access table, 10 user clusters was gained. The center of the cluster is a vector, which is computed as the mean of user preferences in the user cluster. Table 3 indicates the center of clusters 1.

Table3. Center of Cluster

UC1	DC1	DC2	DC3	DC4	DC5
U112	0.7	0	0.3	0	0
U21	0	0.2	0	0.3	0.6
U39	0.5	0.1	0.1	0	0.2
U14	0.2	0.5	0.3	0	0
center	037	0.20	0.17	0.07	0.20

Considering an entry of log file as a new user who is recently connected to the website; access table is calculated again for measuring the interest degree of new user to document clusters. Table 4 shows parts of this matrix.

Table 4. Reformatted access table for new user

Center	DC1	DC2	DC3	DC4	DC5
UC1	0.37	0.20	0.17	0.08	0.25
UC2	0.0	0.25	0.0	0.52	0.26
UC3	0.43	0.0	0.10	0.40	0.0
UC4	0.0	0.50	0.30	0.0	0.25
UC_{NEW}	003	0.0	0.0	0.26	0.45

According to equation 2, support value of active user for each user cluster is calculated and shown in Table 5 and using equation 3, match score tables are calculated as showed in Table 5.

Table 5. Support and Match score calculation for UC_{new}

Ucluster	Support	Match	Rate
UC1	0.7	0.3	L
UC2	0.26	0.74	VH
UC3	0.63	0.37	M
UC4	0.56	0.44	H

The new user has highest match score to the user cluster number 2. According to table 4 users in this user cluster have shown most interest to documents, the document cluster is chosen for making recommendations for new user. One common limitation in clustering algorithms is defining a suitable number of clusters which is solved with using a FCM. Evaluating the effectiveness of the purposed model, the top n recommendations are evaluated using precision, recall and F1 measures. Precision measure shows the accuracy of presented recommendations and recall is a measure of completeness and F1 measure combines Precision and Recall. In Figure 4, 5 and 6 these metrics are evaluated for FARS recommender system which uses hybrid clustering method with systems which uses FCM.

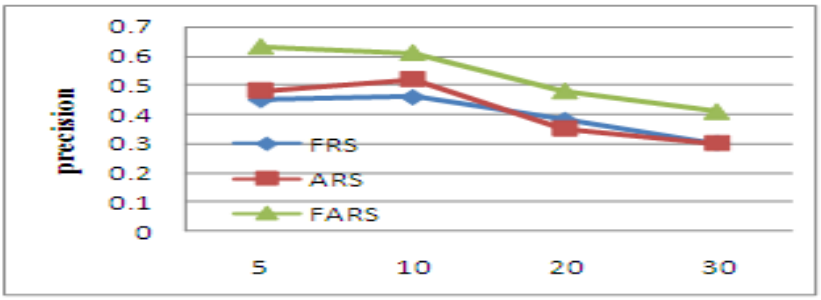

Figure 2. Precision measures for N top recommendations

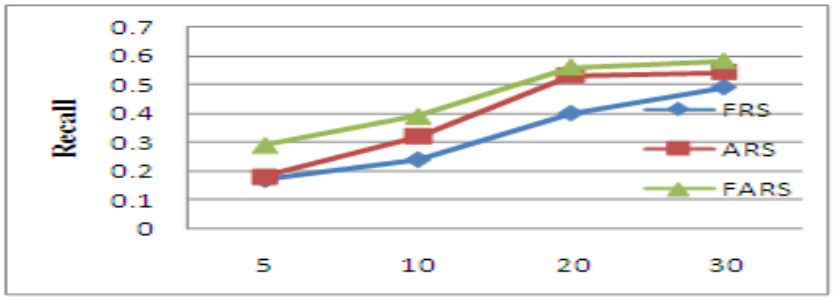

Figure 3. Recall measures for N top recommendations

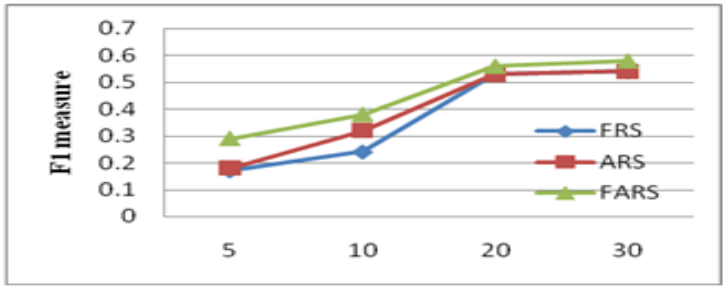

Figure 4. F1 measures for N top Recommendations

The Figures 2, 3, 4 indicates the proposed approach improves the performance of the models which uses FCM clustering methods. Therefore p proposed approach shows more qualified recommendations.

C. Web Mining Through Cloud Computing

Cloud Computing is most seductive technology areas due at least in part to its cost efficiency and flexibility. Despite increased activity and interest, there are significant, persistent concerns about cloud computing that are impeding momentum and will eventually compromise the vision of cloud computing as a new IT procurement model [15]. The term 'cloud' is a symbol for the Internet, used to mark the point at which responsibility moves from the user to an external provider. Basically Cloud Mining is new approach to faced search interface for the data. SaS (Software-as-a-Service) is used for reducing the cost of web mining and try to provide security that become with cloud mining technique. Now a day ready to modify the framework of web mining for demand cloud computing. In terms of "mining" clouds, the Hadoop and Map Reduce communities who have developed a powerful framework for doing predictive analytics against complex distributed information sources.

VI. Conclusion And Future Work

In this paper proposed methodologies used for classifying the user using Web Usage data. This model analysis the users behaviors and depend on the interests of similar patterns provides appropriate recommendations for active user. The model uses the benefits of both content based and collaborative based recommender systems. The results of evaluations shows that using more efficient algorithms for finding similar users lead to recommender system that provides more interesting recommendations for website users. Proposed work can be extended by considering the effect of users' feedback for increasing the quality of recommendation. This can be done, eventually, by introducing new parameters for the characterization of the Web Usage data. Future plan to introduce a hierarchy on the semantic topics (the information about the Web site).

References

[1] Bamshad Mobasher, "Data Mining for Web Personalization," LCNS, Springer-Verleg Berlin Heidelberg, 2007.
[2] Jaideep Srivastave, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations. ACM SIGKDD,2000.
[3] Pierrakos. D, "Web usage mining as a tool for personalization: a survey", User Modeling and User-Adapted Interaction, 13(4), pp. 311-372.
[4] Ralph Kimball. Entrepots de donnees. Editions Vuibert, 2001.
[5] Luotonen. The common log file format. http://www.w3.org/Daemon/User/Config/ Logging.html, 1995.
[6] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, May 2000
[7] M. Spiliopoulou, L. C. Faulstich, and K. Winkler. A data miner analyzing the navigational behaviour of web users. In Proc. of the Workshop on Machine Learning in User Modeling of the ACAI'99 Int. Conf., Creta, Greece, July 1999.
[8] M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically synthesizing web pages. In AAAI/IAAI, pages 727{732, 1998.
[9] V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using model-based clustering. In In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 280{284, Boston, Massachusetts, 2000.
[10] Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. Data Mining and Knowledge Discovery, 6(1):61{82, January 2002.
[11] F. Bonchi, F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C. Renso, and S. Ruggieri. Web log data warehousing and mining for intelligent web caching. Data Knowledge Engineering, 39(2):165{189, 2001.
[12] Osmar R. Zaiane, Man Xin, and Jiawei Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In Advances in Digital Libraries, pages 19-29, 1998.
[13] F. SÄauberlich and K.-P. Huber. A framework for web usage mining on anonymous logfile data. In Exploratory Data Analysis in Empirical Research, Proceedings of the 25th Annual Conference of the Gesellschaft fÄur Klassifikation e.V., March 2001, pages 229{239. Springer-Verlag, 2002.
[14] Diday La methode des nuees dynamiques. Revue de Statistique Appliquee, XIX (2):19{34, 1971.
[15] Robert. Cooley, Bamshed Mobasher and Jaideep Srinivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," journal of knowledge and Information Systems,1999.