Management Science Letters 5 (2015) 377–386

Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

A fuzzy method for improving the functionality of search engines based on user's web interactions

Farzaneh Kabirbeyk[a*], Ali Harounabadi[b] and Mostafa Sabzekar[a]

[a]*Department of Computer, Science and Research Branch of South Khorasan, Islamic Azad University, Birjand, Iran*
[b]*Department of Computer, Central Tehran Branch, Islamic Azad University, Tehran, Iran*

C H R O N I C L E

Article history:
Received January 2, 2015
Received in revised format 6 February 2015
Accepted 7 February 2015
Available online
February 14 2015

Keywords:
Web Personalization
Recommender System
Web Usage Mining
Fuzzy clustering

A B S T R A C T

Web mining has been widely used to discover knowledge from various sources in the web. One of the important tools in web mining is mining of web user's behavior that is considered as a way to discover the potential knowledge of web user's interaction. Nowadays, Website personalization is regarded as a popular phenomenon among web users and it plays an important role in facilitating user access and provides information of users' requirements based on their own interests. Extracting important features about web user behavior plays a significant role in web usage mining. Such features are page visit frequency in each session, visit duration, and dates of visiting a certain pages. This paper presents a method to predict user's interest and to propose a list of pages based on their interests by identifying user's behavior based on fuzzy techniques called fuzzy clustering method. Due to the user's different interests and use of one or more interest at a time, user's interest may belong to several clusters and fuzzy clustering provide a possible overlap. Using the resulted cluster helps extract fuzzy rules. This helps detecting user's movement pattern and using neural network a list of suggested pages to the users is provided.

1. Introduction

During the past few years, the World Wide Web has become the biggest and the most popular way of communication and information dissemination (Kansara & Mishara, 2013). The Web grows by nearly millions of electronic pages on a daily basis, adding to the hundreds of millions of existing pages on-line. Because of its rapid and chaotic growth, the resulting network of information does not maintain appropriate organization, which makes the structure of Web sites more complex. When searching and browsing the Web, most users are often overwhelmed by huge amount of information and they are faced with a big challenges to determine the most relevant information at convenient time (Chitraa & Davamani, 2010; Santra & Jayasudha, 2012; Rajabi et al., 2014).

*Corresponding author.
E-mail addresses farzanehkabirbeyk@gmail.com (F. Kabirbeyk)

doi: 10.5267/j.msl.2015.2.008

Personalized web is a process in which provided information or service of a web site is adapted for a particular user's requirements (Forsati et al., 2008). User behavior modeling is a primary factor in each personalized system executed implicitly by user information (Qaderian, 2008). Recommendation systems major instances of personalization systems have shown to greatly help Web users in navigating the Web, locating relevant and useful information, and receiving dynamic recommendations from Web sites on particular products or services, which fit their interests. To build Web recommendation systems, the Web usage mining methodology is one of the primary techniques implemented in the literature. Such features includes page visit frequency in each session, visit duration, and dates of visiting a certain page.

In this paper, web use mining, fuzzy clustering, and fuzzy rules are implemented to predict user's future demands by producing a list of favorite web pages through neural networks. Various user interactions on the web are tracked and web page categories are created according to user's interests. Users could be turned into permanent customers by providing what they need using fuzzy clustering method. Due to the user's different interest and use of one or more interest in a time, their use may belong to several clusters. Fuzzy clusters provide a possible overlap, and by resulting cluster, it would extract fuzzy rules. Finally, it helps user's movement pattern and using neural network, a list of suggested pages to the users is generated.

This paper is organized as follows. Section 2 includes available approaches and methods in web personalizing based on web usage mining. In section 3, some examples are reviewed. Section 4 includes the proposed method of this research. Section 5 provides details about using recommended method, data collection, and instruments. Section 6 provides information about evaluating criteria and methods, results of experiments and compares them with other methods. Finally, section 7 provides conclusion and advantages of suggested method.

2. Requirements

This section briefly explains different concepts used for the proposed method.

2.1 Web Personalization

Web personalization can be defined as any kind of operation that comply website services or information associated with a particular user or group of users. This is accomplished by applying knowledge of user's observations and personal interests, combined with the content and the website structure. The primary purpose of a web personalization system is to reach information of users' interest without expecting the explicitly request.

2.2 Web Usage Mining

Web Usage Mining deals with the knowledge extraction from server log files to derive useful patterns of user access. Web usage mining tries to capture and model users' behavioral patterns and profiles who interact with a web site. Such patterns can be applied to better understand the behaviors of various user segments, to improve the organization and structure of the site, and to create personalized experiences for users by providing dynamic recommendations of products and services (Tyagi et al., 2010; Thakare & Gawali, 2010; Suresh et al., 2011).

2.3 Clustering

Data clustering is the process of categorizing data elements into classes or clusters so that items in the same class could become as similar as possible, and items in different classes would become as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering, various

measures of similarity could be applied to place items into classes, where the similarity measure controls how the clusters are formed.

2.3.1 Fuzzy Clustering

Fuzzy clustering is a process, which categorizes elements, typically usage clicks or usage sessions into different groups, where each element belongs to different groups with various degrees of membership. In fuzzy clustering, the data points may belong to more than one cluster, and associated with each of the points are membership grades, which indicate the degree to which the data points belong to the different clusters. One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm (Bezdek, 1981) and explain more about it.

3. Related work

Zhong and Li (2010) presented a method where the user profile is made up of three stages. The first step is to determine the useful data while in second step, K-means algorithm is implemented for user session clustering. They reported that the method could an effective role to improve the user model compared with the previous methods. Maheswari and Sumathi (2014) applied the theory of distribution in Dempster-Shafer's theory where the belief function similarity measure in this algorithm added to the clustering task the ability to capture the uncertainty among Web user's navigation performance. Their results show the considerable performance of the proposed algorithm. Azimpour and Azimi (2011) introduced a measure of similarity to compare the user's session. Web session clustering, in their survey, strongly depends on the similarity measure. The results indicated that the proposed method was very effective in recording user session characteristics. Valera and Chauhan (2013) proposed an efficient sequential pattern mining algorithm to determine frequent sequential web access patterns. They believe that the aim of discovering frequent sequential access patterns in web log data is to obtain information about the navigational behavior of the users. The access patterns are retrieved from a Graph, which is then applied for matching and generating web links for recommendations. Mobasher et al. (2001) presented techniques for web personalization based on association rule. Göksedef and Gündüz-Öğüdücü (2010) reviewed four different combination methods and suggested some methods to correct each of them. Also they did comparative evaluation on these four different techniques, which show how different techniques and hybrid method proposed could be effective in prediction accuracy. The proposed hybrid approach, combined the results of different techniques and generated a set of proposal for a new user. They proved that the proposed system had higher accuracy than the hybrid systems. Castellano et al. (2006, 2007, 2008) empirically proposed a fuzzy clustering method (Fuzzy C-Means) for profile detection that could be effective for website personalization. The profile distribution method suggested websites through clusters extracted from relevant data. Using fuzzy C-Means clustering algorithm enabled creation of interwoven clusters that could eliminate uncertainty in user navigation behavior. Primary empirical results are provided to indicate clusters created by data mining among the website’s input data.

4. The proposed method

According to the general scheme of a usage-based Web personalization system, three different modules can be distinguished in the proposed system:

- Log file preprocessing: usage data stored in access log files are analyzed, cleaned and filtered in order to extract user sessions and models of user behaviors representing the basic structures which encode the access patterns exhibited by the users during navigation.
- Knowledge discovery: a number of user categories characterizing the common interests of groups of users are derived by a fuzzy clustering process. Then, a recommendation model is

generated via a fuzzy strategy for establishing the associations between user categories and pages to be recommended.

- Recommendation: interesting pages are dynamically suggested to the current user by exploiting the previously discovered recommendation model. Specifically, when a user requests a new page, his current partial session is matched with the session categories previously identified and derives the degrees of relevance for each page by means of a fuzzy inference process.

These modules are organized into two main macro-modules:

- An offline module, which includes the first two modules to extract a recommendation model from Web usage data;
- An online module, which performs the effective recommendation task.

4.1 Pre-processing

Data exists in this section are as the web server logs. This pre-processing is performed to determine web access sessions. Pre-processing on the web server logs ought to be executed before applying web mining algorithms and pre-process is split into three steps:

4.1.1 Data cleaning

Data Cleaning, which eliminates redundant and useless records contained in the log file to only keep the information concerning accesses to resources of the Web site (typically Web pages). In fact, at this stage of pre-processing, the removed data is as follows:

- requests for image files
- request which response except Get and Put
- failed requests

4.1.2 User session identification

At this phase, the user sessions are detected from log files. Set of pages visited by a user during a specific website mining called user session. The sessions represent user behavior so that they are important in the process of pattern discovery.

4.1.3 Forming the data

This is the last step of preprocessing of data. The data must be in a particular format so web mining techniques can be applied on them. The data set can be stored in a relational database.

4.2 Page Clustering

Various numbers of clustering techniques are used for clustering documents. In this work, following algorithm is used for document clustering. Assume $P = \{p_1, p_2, \dots, p_k\}$ is the set of k website's pages that will be grouped in content based clusters using following steps:

Step 1. Assign each page to a single cluster
Step 2. Merge clusters based on Jaccard coefficient similarity measure by Eq. (1).

$$sim(p_x, p_y) = \frac{|p_x \cap p_y|}{|p_x \cup p_y|} \tag{1}$$

where $|p_x \cap p_y|$ is the number of common words between two basic clusters and $|p_x \cup p_y|$ is total number of words in both clusters.

Step 3. Repeat step 2 until all documents being clustered.

The result is the set, $DC = \{DC_1, DC_2, \dots, DC_n\}$ and DC_i represents a set of URLs with similar content.

4.3 Create Session Victorization and Interest Degree Computation

Both the statistics and the user sessions are used in this step to create a model of the user interest. The most commonly used methods to evaluate user interest about pages is by counting page accesses or "hits". However, this is not sufficient. Access counts, when considered alone, can yield misleading metrics. The time collected for each successive request can give interesting clues regarding the user interest by evaluating the amount of time spent by users on each page. This approach appears reasonable, since it tends to weight content pages higher. However, a long access can completely obscure the importance of other relevant pages. Another possibility is to define interest degrees by the number of times a page was visited during the navigation. The interest degree for each page the user accessed during her/his navigation as a function of two variables:

The overall time the user spends on the page during its visit and the frequency of accesses to the page within the session. Formally, given a page $p_{ij} \in$ P accessed in the ith user session, with access time t_{ij}, the following measure is used to estimate the interest degree:

$$IG_{ij} = f_{ij} \cdot \frac{t_{ij}}{t_i} \tag{2}$$

where $f_{ij} = N_{ij} / \sum_{k=1}^{n} N_{ik}$ is the frequency of accesses to the jth page within session s_i.

The outcome of log file preprocessing is represented by a $n \times m$ matrix B = $[b_{ij}]$ where n and m are respectively the final number of retained sessions (users) and the final number of considered pages. In this matrix, named behavior matrix, each entry b_{ij} represents the interest degree of the ith user for the jth page and is as follows:

$$b_{ij} = \begin{cases} IG_{ij} & if\ page\ p_j is\ accessed\ in\ session\ s_i \\ 0 & otherwise \end{cases} \tag{3}$$

The suggested algorithm is used for determining degree of user for each page based on page observation time and pages date and pages observation frequency of each user criteria. Frequency is calculated by the following formula:

$$frequency(page) = \frac{number\ of\ visits(pages)}{\sum_{page \in visited} number\ of\ visits(pages)} \tag{4}$$

Duration, defined as the time spent on a page. If the user spends more time on a page, that page is more favored by the users and if a page is not favored by the users, the users have rejected that page quickly and will go to another page. However, length of page must also be considered. So the Duration is normalized by length of page. Duration is calculated by the following formula:

$$Duration(page) = \frac{total\ duration(page)/lenght(page)}{Max_{page \in visited\ page}\big((total\ duration(page))/lenght(page)\big)} \tag{5}$$

More recent history is higher priority. To prioritize the pages for each day assigned a constant weight between zero and one, this study considered NASA data set, for every day of this data set; dedicated a constant weight. Finally, user interest can be obtained by taking the harmonic mean of these three features.

$$Interest(page) = \frac{3 \times frequency(page) \times Duration(page) \times date(page)}{frequency(page) + Duration(page) + date(page)} \quad (6)$$

It should be noted that the amount of interest must be normalized between zero and one to be suitable for clustering.

4.4 Session categorization by fuzzy clustering

Once user sessions have been identified, a clustering process is applied in order to group similar sessions in the same category. Each session category includes users exhibiting a common browsing behavior and hence similar interests. Hence, the identified session categories represent the different user profiles, which would be successively exploited for suggesting links to pages considered interesting for a current user.

In this work, the well-known Fuzzy C-Means (FCM) clustering algorithm (2004) is implemented in order to group user sessions in overlapping categories which represent user profiles. The input to the algorithm is the matrix whose the rows is session and the column is the content of web pages. As a result, FCM provides:

$$J_{FCM}(V, U, X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ic}^{m} d^2(X_i, v_i)\,, 1 \le m \le \infty \quad (7)$$

A fuzzy partition matrix $U = \{\mu_{ij}\}$ with $u_{ij} = u_i(x_j)$ where u_{ij} represents the membership degree of the visitor behavior vector x_j to the *i*th cluster. The partition matrix as follows:

$$0 < \sum_{j=1}^{n} \mu_{ij} < n\,, \forall i \in \{1, \dots, C\} \quad (8)$$

$$\sum_{i=1}^{c} \mu_{ij} = 1\,, \forall j \in \{1, \dots, n\} \quad (9)$$

1. $U = [u_{ic}]_{i=1,\dots,n}^{c=1,\dots,C}$ matrix, $U^{(0)}$

2. At βth step: calculate the center vectors $V^{(\beta)} = (v_c)_{c=1,\dots,C}$ as

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}}\,, 1 \le i \le c \quad (10)$$

3. Update $U^{(\beta)}$, $U^{(\beta+1)}$ according to:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d(x_j v_i)}{d(x_j v_k)} \right)^{\frac{2}{m-1}}} \quad (11)$$

4. If $\left\| U^{(\beta+1)} - U^{(\beta)} \right\| < \epsilon$ with $0 < \epsilon < 1$, STOP; otherwise return to step 2.

4.4 Fuzzy rules extraction

After acquiring the center clusters and degree of membership of each cluster, this method will acquire membership functions by Gaussian distribution. Providing membership functions is done in order to improve processing in real world by using Gaussian distribution. Then, it can has fuzzy rules by using fuzzy deduction system based on Mamdani method implemented because it is comprehensible for human entrances.

4.5 Recommending system for neural network

In previous phases, after acquiring users' Profiles, clustering profiles and determining degree of membership for users for each cluster, the fuzzy rules are extracted by using users' profiles and degree of membership of each cluster. In this phase, the system is taught by using neural network. After teaching neural network, the recommended system can predict users' interests and the related pages in this regard.

5. Implementation of proposed method

We use Matlab and C# software, in order to implement suggested method. The log file data of this research is collected based on Nasa log file web server. In order to implement this method, NASA log file used like other research works. As mentioned in section 4.1 preprocessing must be executed during the initial stage. After removing the non-useful data for identifying session C # programming language is used. User sessions are detected with a threshold of 30 minutes. After this step, User session vectors are extracted. Sessions are divided in two categories of training and test. After preprocessing phase and session identification, we have classified pages based on content and structure of web sites into six groups, which include computing, economical, historical, amusing, and structural subjects and the clustering operation based on users' session. Having determined cluster centers to produce membership function and to extract rules, this method used three linguistic variables which include small, medium, and high in order to express the degree of users interest to the pages and degree of membership of each user to each cluster. Based on three linguistic variables and six groups of different pages, the number of extracted fuzzy rules will result in 3^6 rules. The low validity rules will be omitted form the rules collection. Finally, the number of extracted rules is 11 and two of these extracted rules are shown in Table 1. In this table, s1 refers to users interest about computing pages, s2 refers to economical, s3 refers to historical, s4 refers to amusing, s5 refers to structural and s6 refers to the other subjects. c1-c9 expresses the title of cluster or related cluster of each pattern as well.

Table 1
The rules used for clustering the data

Rule1	if interest degree s1 is low and interest degree s2 is low and interest degree s3 is low and interest degree s4 is low and interest degree s5 is low and interest degree s6 is high then (μc1 is low) (μc2 is low) (μc3 is low) (μc4 is low) (μc5 is high) (μc6 is low) (μc7 is low) (μc8 is low) (μc9 is low)
Rule2	if interest degree s1 is high and interest degree s2 is low and interest degree s3 is low and interest degree s4 is low and interest degree s5 is low and interest degree s6 is medium then (μc1 is low) (μc2 is low) (μc3 is medium) (μc4 is low) (μc5 is low) (μc6 is low) (μc7 is medium) (μc8 is low) (μc9 is low)
Rule3	if interest degree s1 is low and interest degree s2 is high and interest degree s3 is low and interest degree s4 is low and interest degree s5 is low and interest degree s6 is medium then (μc1 is medium) (μc2 is low) (μc3 is low) (μc4 is low) (μc5 is low) (μc6 is medium) (μc7 is low) (μc8 is low) (μc9 is medium)

Neural network has been used in this study for recommender engine, three-layer network with an input layer, hidden, and output used in MATLAB software environment and hidden layer in the network has 50 neurons. The back-propagation network algorithm was used to learn from data and Levenberg Marquardt technique has been used as a learning technique.

To suggest pages to users, the current session given as input to the neural network and using fuzzy rules extraction for meeting determines the cluster. When determining the appropriate number of clusters, then, pages that have not been visited, are only pages with the highest level of relevance, which are suggested to the user and included in the suggested list.

6. Evaluating proposed method

Precision refers to the ability of suggested system to produce accurate suggestions. In other words, suggestion precision means the ratio of true suggestion to the total number of suggestions. Recall refers to ability of suggested method to produce suggestions which are obvious for users. In fact; recall refers to ratio of true recognized suggestions in relation to remained pages in the same session.

6.1 Suggested model

In order to teach neural net, this method has used 70 percent of session collections for learning and 30 percent of them for testing. In order to evaluate suggested method, it compared with the provided method (Castellano et al., 2011). The newer method is used for evaluation method because it is more accurate than KNN, FAR, and NP method. The comparison between, precision and recall of recommended method with newer method is shown in Fig. 1 and Fig. 2.

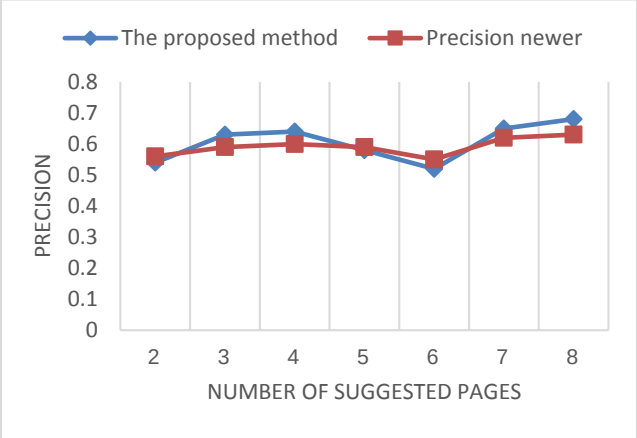

Fig. 1. Comparison of Precision Algorithms

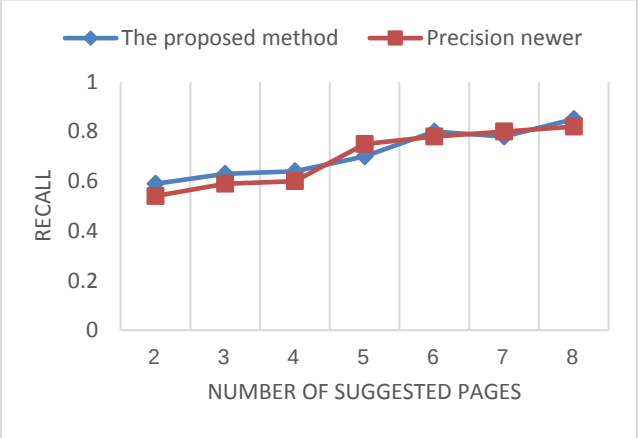

Fig. 2. Comparison of Recall Algorithms

As we can see in Fig. 1, the precision of suggested method is more than newer method about size of suggested pages of 3, 4, 7, and 8. In addition, as we can see from the results of Fig. 2, the proposed method yields more recall than the newer method for size of suggested pages of 2, 3, 4, 6 and 8.

7. Conclusion

As mentioned earlier, web is associated with development of a decentralized and distracted process. This process results in a large mass of connected results, which do not have any logical organization. Since presentation of knowledge in the real world is associated with some uncertainties, and given the

fact that users differ from each other with respect to interests, it is difficult to determine the user's behavioral pattern. In this research, we have used the fuzzy clustering method and a modeling method provided by the fuzzy network to propose a structure and helped deal with various user inclinations, which pose uncertainty as a big problem. The proposed model considered uncertainties in user's behavior and produced a list of priorities based on those uncertainties. The proposed method aims to create users' profiles and detect their common behavioral patterns, implicitly. Finding users' movement pattern has been accomplished by using fuzzy clustering technique and web usage mining. Also, the fuzzy rules of proper clusters have been extracted based on the degree of users' interest to the pages and degree of membership of each user to each cluster. The proposed method employs the extracted rules to find the proper cluster for users and suggests a list of pages to new users by using neural network. Result of experiment shows that, suggested algorithm has very high degree of precision and recall in recommending pages for users.

References

Azimpour-Kivi, M., & Azmi, R. (2011, June). A webpage similarity measure for web sessions clustering using sequence alignment. In *Artificial Intelligence and Signal Processing (AISP), 2011 International Symposium on* (pp. 20-24). IEEE.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers.

Castellano, G., Mesto, F., Minunno, M., & Torsello, M. A. (2007). Web user profiling using fuzzy clustering. In *Applications of Fuzzy Sets Theory* (pp. 94-101). Springer Berlin Heidelberg.

Castellano, G., Fanelli, A. M., & Torsello, M. A. (2006, September). Mining usage profiles from access data using fuzzy clustering. In *Proceedings of the 6th WSEAS International Conference on SIMULATION, MODELLING AND OPTIMIZATION (SMO'06)* (pp. 157-160).

Castellano, G., Fanelli, A. M., Plantamura, P., & Torsello, M. A. (2008, July). A Neuro-Fuzzy Strategy for Web Personalization. In *AAAI* (pp. 1784-1785).

Castellano, G., Fanelli, A. M., & Torsello, M. A. (2011). NEWER: A system for NEuro-fuzzy WEb Recommendation. *Applied Soft Computing*, *11*(1), 793-806.

Chitraa, V. & Davamani, A. S. (2010). A survey on preprocessing methods for web usage data. *International Journal of Computer Science and Information Security*, 7(3), 78-83.

Forsati, R. & Meybodi, M. R. (2008). *An algorithm based on structure of connected pages and information of users for suggesting web pages*. The second Iran data mining conference, industrial Amir Kabir university.

Göksedef, M., & Gündüz-Öğüdücü, Ş. (2010). Combination of Web page recommender systems. *Expert Systems with Applications*, *37*(4), 2911-2922.

Kansara, N. & Mishara, S. (2013). An improved fuzzy clustering technique for user's browsing behaviors. *International Journal of Emerging Trends & Technology in Computer Science* (IJETTCS), 2(2), 2278-6856.

Maheswari, B. U., & Sumathi, P. (2014, February). A New Clustering and Preprocessing for web log mining. In *Computing and Communication Technologies (WCCCT), 2014 World Congress on* (pp. 25-29). IEEE.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001, November). Effective personalization based on association rule discovery from web usage data. In*Proceedings of the 3rd international workshop on Web information and data management* (pp. 9-15). ACM.

Qaderian, M. (2008). Improving user model in website automatically by using semantics of specific degree of concepts. *M.A. Thesis, Amir Kabir University, information technology and computer engineering faculty*.

Rajabi, S., Harounabadi, A., & Aghazarian, V. (2014). A recommender system for the Web: Using user profile and machine learning methods. *International Journal of Computer Applications*, 96(11), 8875-8887.

Santra, A. K., & Jayasudha, S. (2012). Classification of web log data to identify interested users using Naïve Bayesian classification. *International Journal of Computer Science Issues*, *9*(1), 381-387.

Suresh, K., MadanaMohana, R., RamaMohan Reddy, A., & Subramanyam, A. (2011, May). Improved fcm algorithm for clustering on web usage mining. In*Computer and Management (CAMAN), 2011 International Conference on* (pp. 1-4). IEEE.

Thakare, S. B., & Gawali, S. Z. (2010). A effective and complete preprocessing for Web Usage Mining. *International Journal on Computer Science and Engineering*, *2*(03), 848-851.

Tyagi, N. K., Solanki, A. K., & Wadhwa, M. (2010). Analysis of server log by web usage mining for website improvement. *International Journal of Computer Science Issues*, *7*(4), 17-21.

Valera, M., & Chauhan, U. (2013, July). An efficient web recommender system based on approach of mining frequent sequential pattern from customized web log preprocessing. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Verma, V., Verma, A. K., & Bhatia, S. S. (2011). Comprehensive Analysis of Web Log Files for Mining. *IJCSI International Journal of Computer Science Issues*, *8*(6). 199-202.

Zhong, J., & Li, X. (2010). Unified collaborative filtering model based on combination of latent features. *Expert Systems with Applications*, *37*(8), 5666-5672.