An Approach To Webpage Prediction Method Using Variable Order Markov Model In Recommendation Systems

T. Gopalakrishnan[1], P. Sengottuvelan[2], A. Bharathi[1], R. Lokeshkumar[1]

[1] Department of information Technology, Bannari Amman Institute of Technology, India
[2] Department of Computer Science,Periyar University PG Extension Centre, India
gopalakrishnan.ct@gmail.com, sengottuvelan@rediffmail.com, abkanika07@gmail.com, rlokeshkumar@yahoo.com

Abstract

With the continuous increase of web applications and services, web usage data have been overloaded and handling that dynamic web information is very challenging task. Personalized web recommender systems are evolving to provide better and tailored experiences for online users than ever before. The personalization technique is carried out for the each individual user by considering their interests and search behavior stored in the web server access logs. Recently a variety of recommendation systems to predict user future request has been proposed, but the quality of these system results only low prediction accuracy. Hence this paper presents a new framework for effective recommendation system to reduce the searching time of the user and to reach the user's future intention (request) webpage with the improved prediction accuracy by integrating fuzzy c-means clustering and variable order markov model recommendation system. Experimental setups are carried out initially by applying preprocessing technique on the web log followed by fuzzy c-means clustering process to identify the similarity patterns. Finally web page recommendation is performed using variable order markov model to predict the user's next web page access by reducing the search time and better prediction accuracy.

Keywords: Fuzzy c-means clustering, Web Recommendation systems, Variable-order Markov model, Web usage mining, Web page prediction

1 Introduction

The web is enormous, diverse and more dynamic in nature. Extraction of interesting and search related information from web has become more important and as a result of that web mining has attracted lot of attention in recent times [1]. web mining concentrates on the pattern analysis from the world wide web. The overall process of discovering previously unknown, potentially valuable knowledge from web data is known as Web Mining. The web mining can be sub categorized into the three sub categories: web structure mining, web content mining and web usage mining. This research focuses on web usage mining.

With the advancements of electronic devices and hardware technologies, the world is experiencing systems usage in the common man life leads to the increase of data content on the world wide web. The number of websites currently exist in the WWW has been overwhelmed and offers multiple choices for the web users. Web users will struggle to find useful and relevant pages with these available multiple choices and have a tendency to make poor choices of the websites while surfing the web due to an inability to deal with the vast amounts of information. Hence, recommendation system will suggest way to reach out the most relevant webpages on specific search topics with user satisfaction too.

But, how to produce effective webpage recommendations to WWW users automatically without any excessive input from those users is a challenging task and hot research topic. Significant effort has been devoted by various researchers in developing effective webpage recommender systems; nevertheless, a number of challenges and problems, as listed below in the literature review section describe the difficulties faced in the development tenure of effective webpage recommender systems (Figure 1).

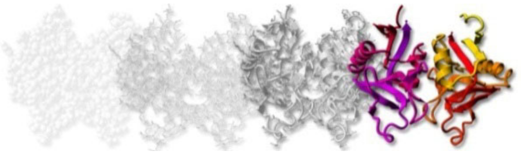


Figure 1. Webpage Prediction from different links

The best methodology to webpage recommendation is based on the rules which are built upon sequential web access patterns that consist of frequently visited Webpages. These web access patterns are usually learned by relating sequence mining techniques to the web usage data, which is often obtained from web server logs. Such extracted user's history of navigation

*Corresponding Author: T. Gopalakrishnan; E-mail: gopalakrishnan.ct@gmail.com

DOI: 10.3966/160792642018031902010

patterns within a specific time is known as session. The interesting patterns could be extracted from the web logs cached for the future page prediction. Also, those logs contain details of sequences of pages that users have been visited frequently along with their timestamp.

With this constraint, when we put effort to solve this webpage prediction problem in the existing recommendation system the following is noticed.

- Searching time for the web user is getting increased.
- Worse prediction accuracy.

To overcome the above listed problem a novel webpage recommendation system has been proposed in this paper by integrating fuzzy c-means clustering technique and variable order markov model.

Organization of the paper. The remainder of this paper is organized as follows: The Section 2 introduces existing related research about web recommendation systems for webpage prediction. Section 3 describes the detailed formulation of our proposed model for webpage prediction in recommendation system. Section 4 describes about the data collection from web server logs. Section 5 provides the techniques used for data pre-processing techniques in proposed system. Sections 6 describes about the clustering technique used for webpage prediction. Section 7 describes about the variable-order Markov model for web recommendations. Section 8 presents experimental results of the proposed system. Finally, Section 9 concludes the paper by summarizing the key contributions and findings of this study.

2 Related Work

This section provides an overview of research work related to Web mining especially for predicting user navigation behavior for web recommendation system. There are number of contributions has made in this area, which has been discussed in the literature. Xing [2] described hybrid-order tree-like Markov model which proposes the combination of two methods. First, a tree-like structure Markov model method that aggregates the access sequences by pattern matching followed by varying-order Markov models.

Sherchan [3] proposed a trust prediction model especially for web service which uses hidden markov model with multivariate Gaussian probability density functions to adjust the different QoS parameters.

In Liang and Zhao [4] to improve the recommendation accuracy, it is important to use a variety of models that compensated for each other's shortcomings. First step is to select significant sentences from web pages. Second we extract features from the significant sentences and construct relevant concepts. At last we use the similarity of web pages to cluster them into different themes. Thus the results show that the combination of the two complementary models can improve the precision rate, coverage rate and matching rate effectively and also help improve the overall solution.

Nigam and Jain [5] proposed different schemes are respectively as prefetching only, Prefetching with Caching and Prefetching from Caching which concludes prediction modeling is analyzed and performed by using dynamic nested Markov model on web logs. Nigam [6] also presented dynamic nested markov model analyzes the web mining in the time complexity and coverage of the prediction state's.

Borges and Levene [7] proposed the model by applying a Sperman footrule metric to evaluate the Variable Length Markov Chain (VLMC). Popa and Levendovszky [8] compared traditional and new markov models on the basis of prediction of user navigation behavior, which suggests that newly introduced Hybrid Order Tree Like model and Selective model provides high accuracy than the traditional models . Awad and Khalil [10] proposed Hidden Markov Model for predicting the user browsing behavior regarding e-business.

In [14], an overall design for preprocessing, clustering and dynamic link suggestion tasks had given and these concepts was used to segment user sessions to clusters or profiles that can later form the backbone for personalization. In [15], a fuzzy rough approximation approach is proposed to cluster the usage patterns. The degree of user's interest is predicted by calculating the time duration on a web page. And it is characterized by fuzzy linguistic variable. Each user session is represented as a fuzzy vector and the fuzzy rough approximation method is used to cluster the users having similar interest. The algorithm converges quickly because two clusters with same upper approximation are merged at each iteration. From these experimentations we can ensure that adaptive web sites will improve themselves by learning user access patterns.

Markov model is the most commonly used prediction model used for predicting the next page to be accessed by the Web user. Traditional markov model predicts by matching the user's current access sequence with the user's historical Web access sequences to predict the next webpage a user will most likely access [9]. By having this information, it can be deduced the future requests of the user.

Thus these entire prediction models such as first order markov model, second order markov model are probably used to predict the user future request in an efficient way. But all these recommendation system shows only less prediction accuracy and also searching time of the system for the user request was getting increased. In accordance with first and second order markov model predicts user future request by only observing at the last action in first order and last two action in the second order performed by the user. These model considers only the very recent searches which obviously doesn't not show the better prediction accuracy and also by applying a clustering technique it shows only little improvement in searching time of the system. By integrating clustering technique (fuzzy c means algorithm) and variable order markov model, a novel recommendation system has been proposed to provide better prediction accuracy and thus by improving a searching time of the system by the user request. In variable-order Markov model(VOMM), the predictions of the next page is done by analyzing at the last *k* searches of the user, leads to a state-space which holds all the possible sequences of *k* actions.

3 Proposed Work

The webpage prediction for recommender systems has gained its importance with the increased web sites and encounters from the large sets of logs cached on the server (Figure 2). This article proposes a method to improve the performance of the web server by analyzing user behavior by prefetching and integrating Fuzzy clustering with variable-order Markov model to achieve better web page access prediction accuracy.

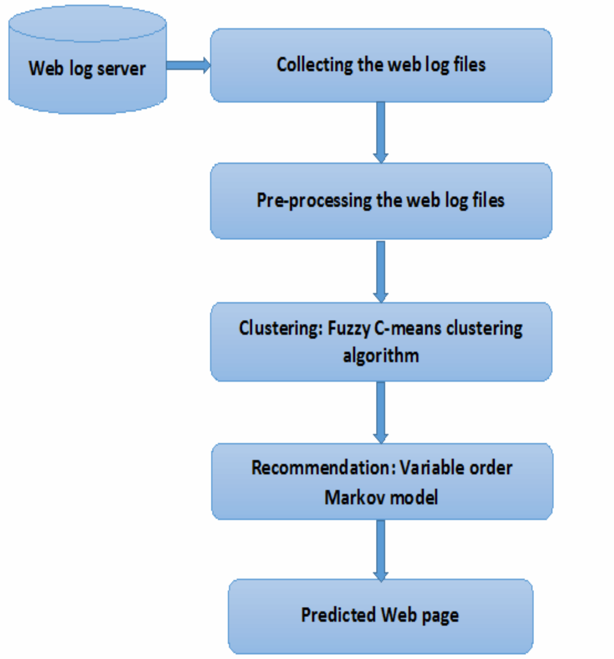

Web log server
Collecting the web log files
Pre-processing the web log files
Clustering: Fuzzy C-means clustering algorithm
Recommendation: Variable order Markov model
Predicted Web page

Figure 2. The proposed model

Intially, the web log files are pre-processed and that is further clustered using the Fuzzy C-means clustering technique. After clustering of the log files, the variable-order Markov Model is applied to used to predict the next suitable web page.

The web log file is pre-processed and then clustered using Fuzzy C-means clustering technique. Clustering of web log files is used to reduce the searching time of the prediction system. The fuzzy c-means clustering clusters the log files into related categories. The variable-order Markov model is applied to predict the next page action and it improves the accuracy of the prediction model. The proposed system has the following stages:

(1) Data Collection: *Collect the web access log information from web servers*.

(2) Pre-processing Log Files: *Pre-process the Web server log files (i.e. Data Preprocessing or Filtering)*.

(3) Clustering Log Files: *Perform clustering by applying Fuzzy C-means clustering technique*.

(4) Web Recommendations: *Apply Markov model and display the webpage with highest probability as the Predicted webpage*.

4 Data Collection

In any user session, all the navigation activity on the web site is cached using a log file in the web server. The web log files collected are given as input for the analysis in the web page prediction process. The web log file mainly contains the following fields:

- ip_address - user's IP address
- user_Id - user name from the browser machine
- base_url - requested resource path
- date - request access date and time
- method - HTTP request type
- file - html file requested by the specific user
- catdesc - category description of the web page
- protocol - protocol currently used for transmission
- code - status/error code
- bytes - total number of bytes transferred.
- referrer - previously visited site by the user
- user_agent - type and version of the browser

5 Pre-Processing Log Files

Generally in the web usage mining, the preprocessing [10] is considered as a basic and essential task. Preparing cached log data for analysis by removing irrelevant data items is known as Pre-processing. The quality of the data is an important issue in the mining process. About the 80% of mining efforts are often spend to improve the quality of data [17]. We obtain mostly incomplete, noisy and inconsistent data from the server logs. The attributes that we can look for in quality data depends upon the accuracy, completeness, consistency, timeliness, believability, interpretability and accessibility. Preprocessing is needed for obtaining the above said attributes to extract the interesting patterns of knowledge. The following steps explains the steps in preprocessing the web log files [18]:

Data cleaning. Removal of noisy and irrelevant data from the web logs [13] is the basic step for data cleaning. When the user request the any HTML web pages, the embedded images are also be downloaded and gets stored in the web server. But these are not explicitly requested by the users which are avoided.

The suffix of the each URL are verified for the filtering process. The algorithm for data cleaning is given in Table 1.

```
Input Given: Web Logs file
Output Expected: FilteredLog Table
Steps:
1. Read the stored web logs record
2. If (suffix/url represent image's/ any
       multimedia file extensions) then
   If (the request is explicit request) then
       Add the records to FilterLog Table
   If (status code not equal to failure) and
       (user agent != crawler, spider, robot)
       and(method == 'GET') then
   Add the respective records to FilterLog Table
3. Repeat Step 1 and 2 till end of Log File
```

Table 1. Records before & after cleaning

	Number of Records
Before Data cleaning	1,999
After Data cleaning	948

The records which have the extension of *.gif, *.jpeg, *.css, *.cgi, etc are removed as it was categorized as non- HTML web page. Also, request from auto search engines and the poor status code are also removed. Nearly about 50%-60% irrelevant records are removed.

In the proposed system, data cleaning removes the irrelevant data from the given log files. For the given web log file, the log files before data cleaning was 1,999 and after data cleaning is 948. It nearly removes the 50%-60% of unwanted data from the log file which in-turn also reduces the processing time of the recommender system. For example, the following type of record is removed from the log file, because it is in the *.jpg file extension.

```
"itime=1413777159  date=2014-10-20
time=09:22:26  devid=FGT1KC3912802483
vd=root       type=utm
subtype=webfilter    action=passthrough
  cat=255              dstip=184.31.219.143
  dstport=80 eventtype=urlfilter
  hostname=i1.sdlcdn.com     level=notice
  logid=13317          method=domain
  msg="URL has been visited"          profile=Staff
  rcvdbyte=4329 reqtype=referral     sentbyte=393
  service=httpsessionid=154566683
  srcip=172.16.2.202
  srcport=53597 url=/img/metroUI/header/account-
  bg.jpg"
```

User identification. Users are distinguished based upon the requests from different IP addresses and as a result, the different users are identified. This can be done in various ways, by analyzing using IP addresses, authentication mode, cookies, etc. The following s algorithm is used user identification:

```
Input Given: FilterLog Table
Output Expected: Stored Log file with distinguished
                 users
Steps:
1. If the request from the browser is from new IP
   address is unique ,then it means new user;
2. If IP address is same but user name is not unique,
   agent log, OS and browser are different then it is a
   different user.
3. Construct site topology to verify access path &
   identify users
4. Repeat step 1, 2 and 3 till end of FilterLog Table
```

In the proposed system, the unique users are identified based on the '*log_id*' field. For the given log file, four unique users are identified. Each user have the different number of records (Table 2).

Table 2. Number of users count in log file

User Count	**No. of Records per User**
1	329
2	484
3	84
4	51
Total User: 4	Total Records: 948

Session identification. The stipulated time duration spend on each particular web page by the single user is a taken as a session. The session identification process [12] is a process of categorizing the individual user access logs into specific sessions. Both starting and ending time of each session is calculated by identifying the login and logout time The following are the common rules to identify user session:

(1) *If the user is identified as new user, then there is a new session.*

(2) *Else for the same session, if the refer page is null, then there is a new session.*

(3) *If the time between page requests exceeds >18 minutes, it is considered as new session.*

The formula for calculating the session time is given below:

$$session = \sum time\,(site_i \rightarrow site_j)$$

$$frequency = \sum w_v$$

In the proposed system, the session time is calculated based on the "time" field. In this each user has different sessions (Table 3).

Table 3. Each user's session time

User Count	User's Spent Time (Hours: Minutes: Seconds)
User 1	05:39:52
User 2	06:24:15
User 3	12:00:00
User 4	03:23:40

Path completion. The missing page references are filled with the construction of the site topology. Referrer page is taken from the constructed site topology to perform path completion. Figure 3 shows the site topology for path completion.

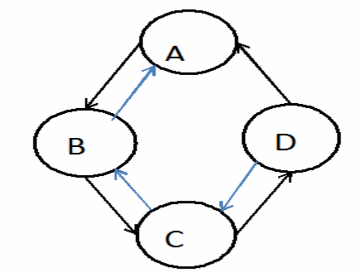


Figure 3. Site topology

The existing page reference sequence mostly will be in the pattern of A – B – C – D – C – B – A. Let us consider that the web user is accessing the page P using back button, but when it is recorded in the web server it will be stored as the sequence pattern: A – B – C – D – A. So, this type of missing paths is completed by using site topology and referrer log.

In the proposed system, path completion helps to reduce the unwanted and unreachable data from the given log files. For the given web log file, the log files before path completion is 948 and after path completion is 475. It nearly reduces the 50% of unwanted data after path completion (Table 4).

Table 4. Data reduce after path completion

	Number of Records	Data Reduce % (approx..)
Before Path completion	948	55
After Path completion	475	75

6 Clustering Log Files

Fuzzy C-means clustering is used for the process of clustering log files in the web page prediction system. Fuzzy C-means (FCM) is a technique of overlapped clustering which allows single data object to belong to two or more clusters. In fuzzy clustering, each point has a degree of belonging to the clusters, as in fuzzy logic, rather than belonging completely to any one of the cluster. Thus, any points present in the cluster edge will belonging to the set of cluster but lesser degree to than the points in the center of cluster.

While clustering any spot x will take the pair of coefficients giving the degree of being in the k^{th} cluster $w_k(x)$. With fuzzy c-means, the centroid of the cluster will be the mean of all points and weighted by their degree of belonging to the cluster. The formula is given below:

$$C_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}$$

The degree of belonging, $w_k(x)$, is associated inversely to the distance from x to the cluster center as calculated in the previous pass. It also depends on the parameter m that controls how much weightage is given to the closest center. The fuzzy c-means (FCM) algorithm is given below.

Input Given: Pre-processed log file
Output Expected: Clustered categories
Steps:

1. Choose the number of clusters needed
2. Assign randomly to each point coefficients for being in the clusters
3. Repeat until the algorithm has converged
4. Compute the centroid for every cluster, using the technique above
5. For every point, compute its coefficients of being in the clusters, using the above formula

Fuzzy c-means clustering algorithm clusters the pre-processed log files into set of related categories. FCM is used to reduce time consumption in the prediction system.

In the proposed system, the clustering process is done by gathering the similar log files into the same cluster. The given log file is grouped into four categories such as Entertainment, Sports/Business, Education, Social Networking and Education, based on the "catdesc" field. The clustering process reduces the searching time of the recommender system (Table 5).

Table 5. Categories based on the user access

Cluster Name	Number of Records
Education	107
Entertainment	63
Sports / Business	250
Social Networking	22

7 Web Recommendations

Caching popular web objects very close to the user's search will have a chance to reduce latency by allowing users to fetch data from a nearby node rather than from a distant server node. Predictive caching technique becomes an important method where in the forthcoming webpage is likely to be requested by the user are predicted based on the history of the user log.

Major method applied for this intent is Markov

Model. Markov model is the one of the most frequently used prediction model technique to predict the next page to be accessed by the web user based on the patterns of their previous webpages access.

The number of previous actions used to predict the next action of the web user [16] is based upon the state-space provided in the markov model. If the next action is predicted only by the last action of the web user it is called first-*order Markov model (FMM)*. Relatively other model's will do the predictions by considering the last two actions performed by the particular web user is known as the *second-order Markov model*. Those state will resemble to all possible pairs of actions that can be performed in an obtained sequence known as *Variable-order Markov model (VoMM)*, where evaluation to predict the next possible sequences of K^{th} *actions* by looking at the last K^{th} actions.

Lets consider the <*Ac*, *St*, *T*> may be three important parameters where in any markov model,

→*Ac will be the* set of all possible *actions* that can be performed by the user;

→*St will be* set of all possible states for which the Markov model is built; and

→*T represents* |St| × |*Ac*| the *Transition Probability Matrix* (TPM), where each entry at t_{ij} will corresponds to the probability of performing the action *j* when the process is in state *i*.

Let us assume $p_1 = \{p_1, p_2, ..., p_m\}$ will reamin as a set of webpages accessible in a particular website. Consider *p* as a user session contains the sequence of pages visited by the specific user at a particular visit. Assuming that the user has visited L pages, then $prob = (p_i \mid W)$ will be the probability that the user visits pages p_i next. Page P_{l+1} the user will visit next is estimated by:

$$\begin{aligned} p_{l+1} &= \arg\max_{p\in 1p}\{P(P_{l+1} = p \mid W)\} \\ &= \arg\max_{p\in 1P}\{P(P_{l+1} = p \mid p_l, p_{l-1,\ldots}, p_1)\} \end{aligned}$$

This mentioned probability, $prob = (p_i \mid W)$, is estimated by analyzing the complete sequences of all users in past (or training data), denoted by W. Certainly, the longer *L* and the larger W, will produce more accurate $prob = (p_i \mid W)$. Longer L & W will be practically impossible and also leads to unnecessary complexity.

With these, higher probability is predictable by assuming the sequence of the pages visited by specific users will follows the markov process which imposes a limitation on the number of previously accessed kth pages .

Morethan that, the probability of visiting a page P_i does not depend on all the pages in the session, but only on the small set of *k* preceding pages, where *k*<< l. With these, the equation becomes:

$$P_{l+1} = \arg\max_{p\in 1P}\{P(P_{l+1} = p \mid p_l, p_{l-1}, ..., p_{l-(k-1)})\}$$

where *k i*s the number of the previous pages visited and it identifies the order of the Markov model. Based on these considerations, the resultant model of the above equation is called as all k^{th} *order Markov model*. The Markov model will be initiated by calculating the highest probability of the last webpage visited because during a Web session, the user can only linkage the page he/she is currently visiting to the next webpage. The probability of $P(p_i \mid S_j^k)$ foe estimating the k value is expected as follows from a past data set.

$$P(p_i \mid S_j^k) = \frac{Frequency\ ((S_j^k, p_i))}{Frequency\ ((S_j^k))}$$

This above said formula will be used to calculate the conditional probability as the ratio of the frequency of the sequence occurring in the training set to the occurrence of the page occurring directly after the sequence.

The traditional Markov model's predicts the next action of the web user only by considering the last action performed by the web user, where else VoMM carries the prediction by considering the last *k* actions of the web user and in-turn improves the prediction accuracy.

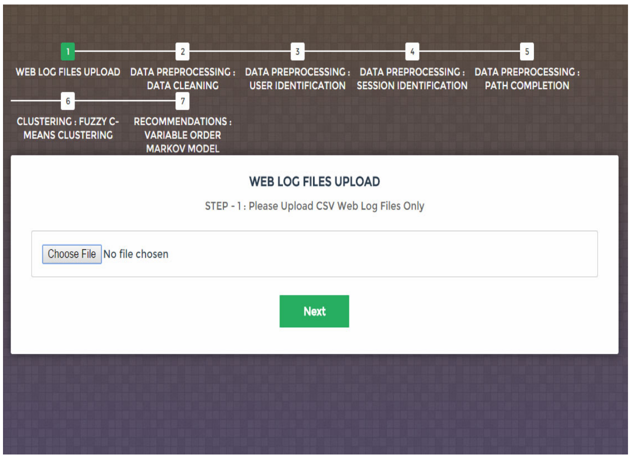
Thus the Variable-Order Markov model (VoMM) creates the opportunity of achieving higher prediction accuracy with the improved domain coverage than the other single-order Markov. This specific idea had motivated us to design a variety of techniques for intelligently coupling different order Markov models so that the resulting model will provide the low state complexity with improved prediction accuracy, and also retains the domain coverage of the Variable-Order Markov model (VoMM). The algorithm for predicting the web page using K^{th} order sequence is assumed as below:

```
Input Given : Set of required clusters
Output Expected: Predicted webpage
Steps:
(1) Find the specific cluster/domain to which current
    web page belongs.
(2) Identify the session, if it is the same session and
    session time < Time threshold then
(3) Find out the predicted value for all the web pages
    within the same cluster to which current web page
    belongs.
(4) Identify the web page with maximum value of
    predicted web page.
(5) Find the prediction value for all k^th order Markov
    model.
(6) Identify the candidate web page which as
    maximum value of transition probability as
    predicted web page among all the k^th order.
```

8 Experimental Results

The weblog are collected for the period of six months from August 2014 to January 2015 from our institute web server and client side machine. We have implemented this work using PHP in the proposed system contains Intel core i3 processor with 4GB RAM.

Initially, the weblog files are uploaded as the basic input for the webpage recommendation system (Figure 4). The format of web log file uploaded to the system is Comma Separated Values (CSV).



1 WEB LOG FILES UPLOAD
2 DATA PREPROCESSING : DATA CLEANING
3 DATA PREPROCESSING : USER IDENTIFICATION
4 DATA PREPROCESSING : SESSION IDENTIFICATION
5 DATA PREPROCESSING : PATH COMPLETION
6 CLUSTERING : FUZZY C-MEANS CLUSTERING
7 RECOMMENDATIONS : VARIABLE ORDER MARKOV MODEL

WEB LOG FILES UPLOAD

STEP - 1 : Please Upload CSV Web Log Files Only

Figure 4. Web log files upload

As a first step after uploading the log files data cleaning is performed in the pre-processing phase. Data cleaning removes the records which have the extension *.gif, *.cgi *.css, etc. other than HTML/JSP/ASP files, and the records with the failed/error status code. During this process request obtained from auto search engines such as crawlers, robot are also removed. As for as our experimentation, for given web log file, the number of records before data cleaning was 1,999 and after performing the data cleaning process it is reduced to 948 records (Figure 5).

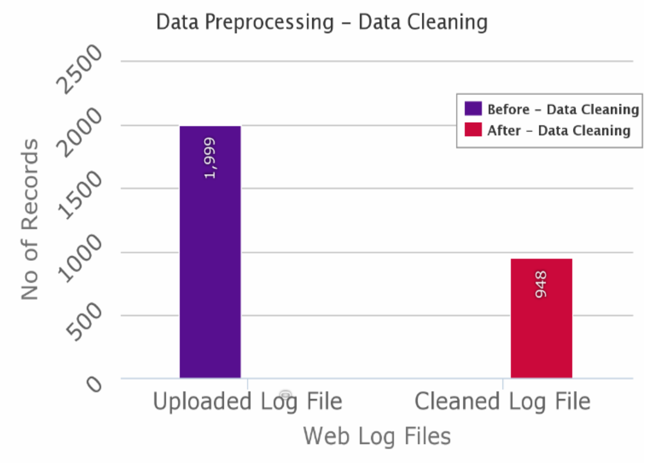
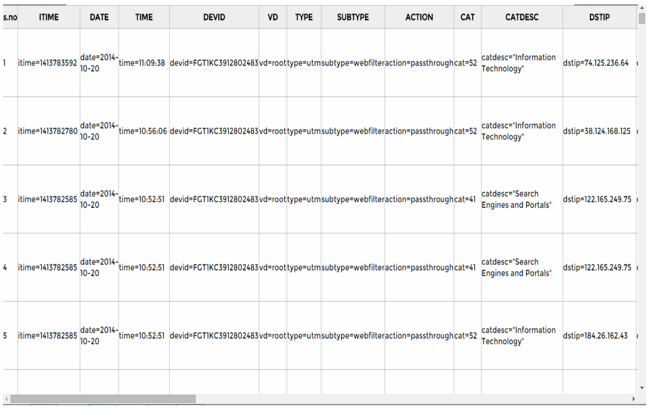


Figure 5. Cleaned web log file

Above Figure 6 describes the view of cleaned web log file. The log files are removed based on the file extensions, failed status code, and requests from auto search engines. Approximately, 30%-35% of the log data is reduced in the data cleaning process. This will reduces the processing time of the recommendation system.



s.no	ITIME	DATE	TIME	DEVID	VD	TYPE	SUBTYPE	ACTION	CAT	CATDESC	DSTIP
1	itime=1413783592	date=2014-10-20	time=11:09:38	devid=FGT1KC3912802483	vd=root	type=utm	subtype=webfilter	action=passthrough	cat=52	catdesc="Information Technology"	dstip=74.125.236.64
2	itime=1413782780	date=2014-10-20	time=10:56:06	devid=FGT1KC3912802483	vd=root	type=utm	subtype=webfilter	action=passthrough	cat=52	catdesc="Information Technology"	dstip=38.124.168.125
3	itime=1413782585	date=2014-10-20	time=10:52:51	devid=FGT1KC3912802483	vd=root	type=utm	subtype=webfilter	action=passthrough	cat=41	catdesc="Search Engines and Portals"	dstip=122.165.249.75
4	itime=1413782585	date=2014-10-20	time=10:52:51	devid=FGT1KC3912802483	vd=root	type=utm	subtype=webfilter	action=passthrough	cat=41	catdesc="Search Engines and Portals"	dstip=122.165.249.75
5	itime=1413782585	date=2014-10-20	time=10:52:51	devid=FGT1KC3912802483	vd=root	type=utm	subtype=webfilter	action=passthrough	cat=52	catdesc="Information Technology"	dstip=184.26.162.43

Figure 6. View of log file after data cleaning

After data cleaning process, identification of user is performed. Identification of individual web users who accessed a website is an important step in web usage mining and it is based on the “log_id” field in the collected weblog file. For the uploaded weblog file, 4 unique users are identified in the prediction system and the Figure 7 shows the each user have accessed different number of records.

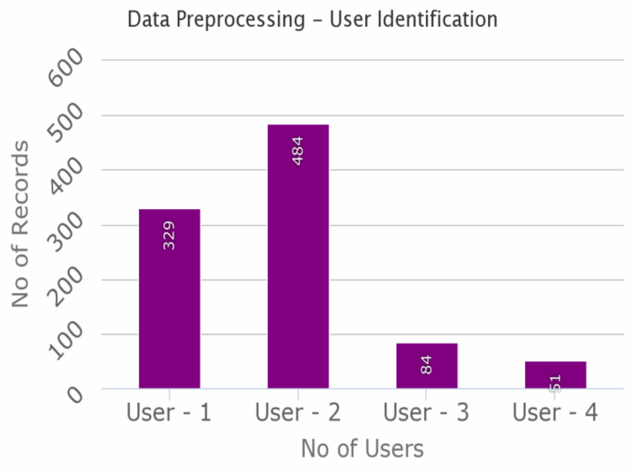


Figure 7. Identification of users in web log file

Session Identification should be performed for an every individual user uniquely identified. A user session can be mentioned as a set of pages visited by the same user within the particular duration. The total time spend by unique user is identified in the session identification which is shown in Figure 8. The session time for each user is considered by the “time” field. Each identified user has different sessions.

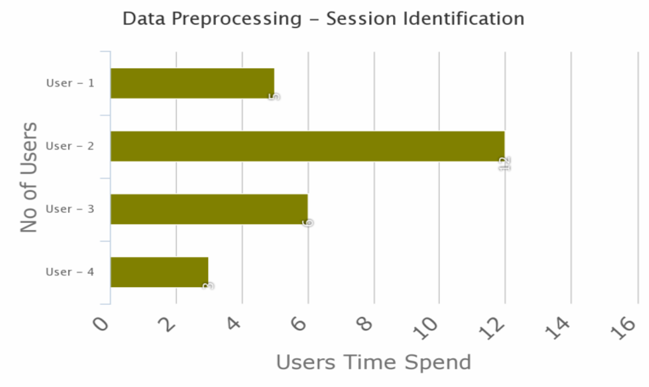


Figure 8. Session identification of each user

Path completion is verified by the system to calculate whether the user has reached the specific page or not. The unreachable pages are removed from the web log file where the path is not fully completed in hyperlinks. For the given log file, the number of records before path completion is 948 and after path completion is 475, shown in Figure 9. It nearly reduces the 45%-50% of irrelevant data from the uploaded web log file and also helps in reducing the processing time of the system.

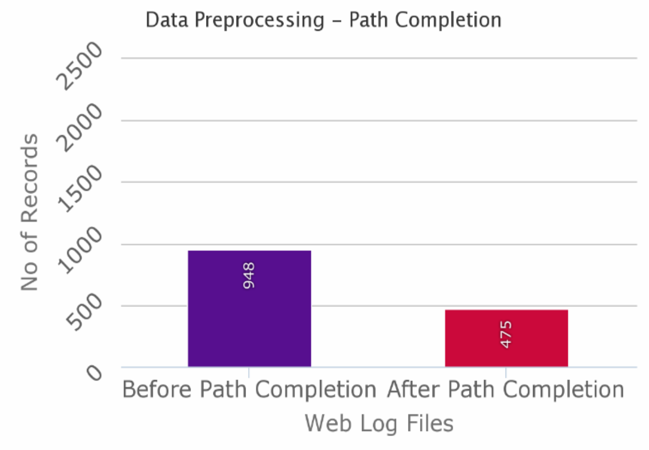


Figure 9. Performing path completion

After performing Data Pre-processing, the pre-processed log file is given as an input for fuzzy c-means clustering process. Above Figure 10 shows the clustered log files using Fuzzy C-means clustering algorithm. The given log file is grouped into four categories such as Education, Entertainment, Sports/Business, and Social Networking based on the 'catdesc' field. The clustering process is used to categorize the webpage based on the above said four categories.

After completing the clustering log files into different categories, Web recommendation is performed using the variable-order Markov Model. Above Figure 11 shows the Recommended Categories of the web pages. It is used to predict the user's next web page access. For the given log file, the recommended category is "Sports/Education".

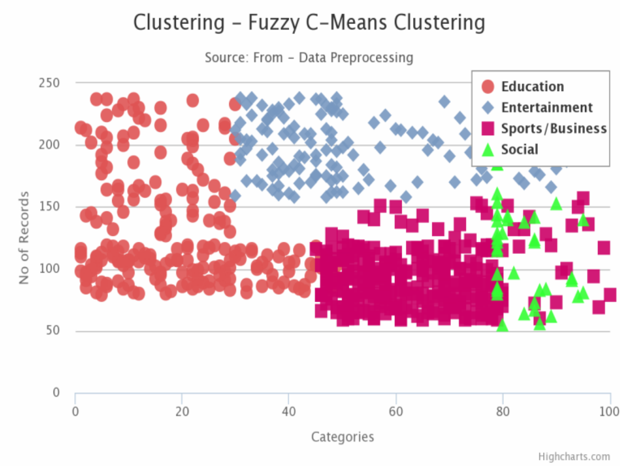


Figure 10. Clustered log files using fuzzy c-means algorithm

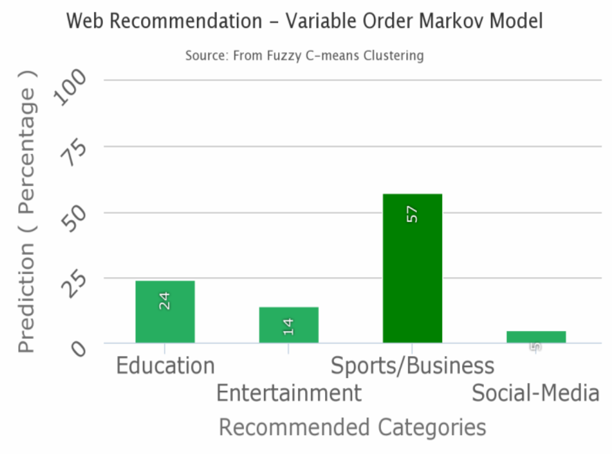


Figure 11. Page prediction for recommender system

It gives the prediction percentage of above two recommended category are sports is 57% and education is 24%. By comparing the integration of clustering technique (Fuzzy c-means algorithm) and variable order markov model it provides better prediction accuracy and less searching time of the system to the user request.

9 Experimental Validation

To evaluate the efficiency of the proposed method, performance is measured in two factors namely precision and domain coverage [20]. Precision recommendation here specifies about the quantity of correct recommendations *i.e.* proportion of the relevant recommendations to the total number of recommendations. Precision is given by the following formula,

$$precision = \frac{T(p) \cap R(p)}{R(p)}$$

Where $R(p)$ is recommendation set
$T(p)$ is session

The domain coverage of the recommender system is nothing but the proportion of relevant recommendations to all pages that should be recommended. Precision of the systems here is measured for varying number of recommended pages. Here, based on proposed system we created the practical evaluation session using PHP.

Following graph displays the improved performance evaluation of the existing algorithm *i.e.* based on traditional markov model without efficient preprocessing and our proposed method based on the Variable Order Markov Model (VoMM) for web recommendation system.

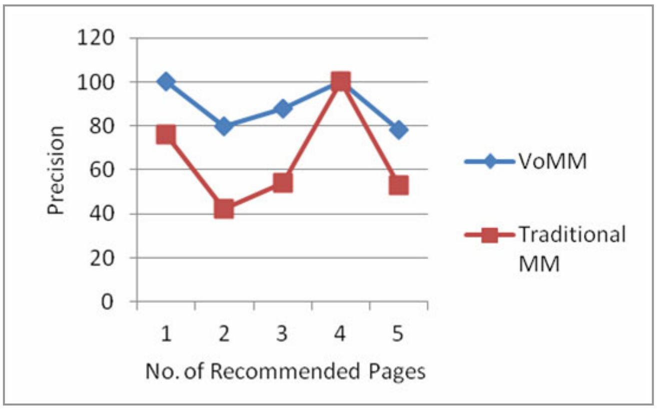


Figure 12. Precision comparative analysis

In both the precision and coverage analysis chart the value shows the variable order markov model will always have upper hand performance than the traditional markov models.

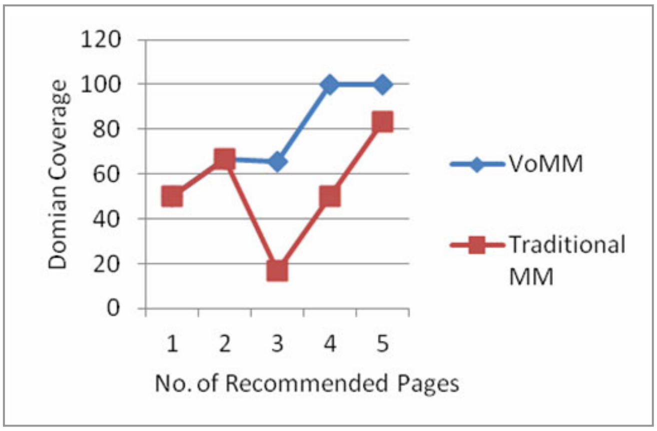


Figure 13. Domain coverage comparative analysis

10 Conclusion

This paper propose a new framework to offer better webpage prediction for personalized web recommendation system by integrating variable-order markov model and fuzzy c-means clustering technique with efficient pre-processing. As the experimental results indicates the proposed webpage prediction system displays a list of personalized recommended categories by processing the web server log files. It includes improvised techniques for efficient pre-processing and clustering of log files (past visited pages) which leads to searching of new patterns in clusters rather than searching whole web logs, thereby reducing time consumption. This in-turn increases performance of the overall webpage recommendation system by predicting the next request page quickly. Thus the webpage prediction is carried out for the each individual user by considering their interests and search behavior stored in the web server access logs are analyzed and evaluated to construct the efficient personalized recommendation technique. Based on the existing limitations, in this paper new mining approach based on combination of efficient preprocessing with FCM and variable order markov model is presented which is showing the improved performance as compared to the existing traditional markov methods. For the work we also suggest to apply two-level clustering in future to have even more accuracy.

References

[1] N. K. Tyagi, A. K. Solanki, M. Wadhwa, Analysis of Server Log by Web Usage Mining for Website Improvement, *International Journal of Computer Science Issues*, Vol. 7, No. 8, pp. 17-21, July, 2010.

[2] D. Xing, J. Shen, A New Markov Model for Web access Prediction, *Computing in Science and Engineering*, Vol. 4, No. 6, pp. 34-39, November, 2002.

[3] W. Sherchan, A Trust Prediction Model for Service Web, *Proceedings of the 2011 International Joint Conference of IEEE TrustCom*, Los Alamitos, CA, 2011, pp. 258-265.

[4] W. Liang, S.-H. Zhao, A Hybrid Recommender System Combining Web Page Clustering with Web Usage Mining, *2009 International Conference on Computational Intelligence and Software Engineering*, Wuhan, China, 2009.

[5] B. Nigam, S. Jain, Analysis of Markov Model on Different Web Pre-fetching and Caching Schemes, *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Tamilnadu, India, 2010, pp. 78-83.

[6] B. Nigam, Generating a New Model for Predicting the Next Accessed Web Page in Web Usage Mining, *Third International Conference on Emerging Trends in Engineering and Technology*, Nagpur, India, 2010, pp. 485-490.

[7] J. Borges, M. Levene, Evaluating Variable Length Markov Chain Models for Analysis of User Web Navigation Sessions, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 4, pp. 441-452, April, 2007.

[8] R. Popa, T. Levendovszky, Markov Models for Web Access Prediction, *8th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics*, Budapest, Hungary, 2008, pp. 539-550.

[9] P. Kaushal, Prediction of User's Next Web Page Request by Hybrid Technique, *International Journal of Engineering Technology and Advanced Engineering*, Vol. 2, No. 3, pp.

339-342, August, 2012, .

[10] M. A. Awad, I. Khalil, Prediction of User's Web-Browsing Behavior: Application of Markov Model, *IEEE Transactions on Systems, Man, and Cybernetics*- Part b: Cybernetics, Vol. 42, No. 4, pp. 1131-1142, August, 2012.

[11] V. Chitraa, A. S. Devamani, A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing, *International Journal of Computer Applications*, Vol. 34, No. 9, pp. 24-28, November, 2011.

[12] S. A. Raiyani, S. Jain, Efficient Preprocessing Technique Using Web Log Mining, *International Journal of Advancements in Research & Technology*, Vol. 1, No. 6, pp. 1-5, November, 2012.

[13] V. Losarwar, M. Joshi, Data Preprocessing in Web Usage Mining, *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012)*, Singapore, 2012, pp. 257-266.

[14] T. Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal, From User Access Patterns to Dynamic Hypertext Linking, *Computer Networks and ISDN Systems*, Vol. 28, No. 7-11, pp. 1007-1017, May, 1996.

[15] C. Chen, A Fuzzy Rough Approximation Approach for Clustering User Access Patterns, *IEEE World Congress on Software Engineering*, Xiamen, China, 2009.

[16] M. Deshpande, G. Karypis, Selective Markov Models for Predicting Web Page Accesses, *ACM Transactions on Internet Technology (TOIT)*, Vol. 4, No. 2, pp. 163-184, May, 2004.

[17] D. A. Grossman, O. Frieder, *Information Retrieval: Algorithms and Heuristics*, Springer, 2004.

[18] R. Suguna, User Interest Level Based Preprocessing Algorithms Using Web Usage Mining, *International Journal on Computer Science and Engineering*, Vol. 5, No. 9, pp. 815-822, September, 2013.

[19] G. Bejerano, Algorithms for Variable Length Markov Chain Modeling, *Bioinformatics*, Vol. 20, No. 5, pp. 788-789, January, 2004.

[20] N. K. Papadakis, D. Skoutas, STAVIES: A System for Information Extraction from Unknown Web Data Sources through Automatic Web Wrapper Generation Using Clustering Techniques, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 12, pp. 1638-1652, December, 2005.

Biographies

T. Gopalakrishnan received B.Tech Information Technology from Anna University in 2005 and M.E. Computer and Communication from Anna University in 2008. Currently he is working as Assistant Professor (Senior Grade) in the Department of IT, BIT, Sathyamangalam. He is pursuing part time research at Anna University.

P. Sengottuvelan received M.E. Computer Science & Engineering from Anna University in 2004. He also received Ph.D. in Computer Science & Engineering Vinayaka Missions University, Salem in 2010. Currently he is working as Associate Professor in the Department of IT, BIT, Sathyamangalam. His current research focuses on Concurrent Engineering, web mining.

A. Bharathi completed her Bachelor's in Engineering degree in Computer Science and Engineering under Bharathiar University, Coimbatore in 1997. She then completed her Maters in Engineering in Computer Science and Engineering under Anna University, Chennai in 2007. She also received her Ph.D. in Information and Communication engineering specializing in Data Mining, 2012. Her current research focuses on Information Retrieval, Data Mining and Analytics. She is guiding around 10 Ph.D. scholars.

R. Lokeshkumar received B.E Computer Science from Anna University in 2006 and M.Tech Degree in Information Technology from, Anna University in 2009. Currently he is working as Assistant Professor (Senior Grade) in the Department of IT, BIT, Sathyamangalam. He is pursuing part time research in web mining at Anna University.