"FP BASED HYBRID APPROACH FOR PREDICTING WEB USER'S FUTURE REQUEST"

[1]Rahul Kaul, [2]ShraddhaKumar
[1]M.Tech Scholar, [2]Assistant Professor
[1]Department of Computer Science & Engineering, SDBCT Indore, India
[2]Department of Computer science & Engineering, SDBCT Indore, India
[1]kaulrahul07@gmail.com, [2]shraddha.kumar@sdbct.ac.in

ABSTRACT:

The use of Web technology has increased by an excellent extent within the recent times. Millions of users spend time on internet to get data or for recreational activities. Along with satiating their own purpose, the users leave behind a detailed path of all the online pages accessed and also the frequency with which they're accessed. This information is of dominant use to several industrial domains like ecommerce websites, social networking sites, entrepreneur franchises, etc. This web log contains lot of information so it is preprocessed before modeling. The web log file is preprocessed and converted into the sequence of user web navigation sessions. The web navigation session is the sequence of web page navigated by a user during time window. The user navigation session is finally modeled through a model. Once the user navigation model is ready, the mining task can be performed for finding the interesting pattern. Modeling of web log is the essential task in web usage mining. The prediction accuracy can be achieved through a modeling the web log with an accurate model to improve the performance of the servers, caching is used where the frequently accessed pages are stored in proxy server caches. Pre-fetching of web pages is the new research area which when used with caching greatly increases the performance. In this paper, a better algorithm for predicting the web pages is proposed. Clustering of web users according to their location using clustering is done and then each cluster is mined using FP-Growth algorithm to find the association rules and predict the pages to be pre- fetched for storing in cache. HMM is used to focus on the hidden stages of web log files.

Keywords: Web Usage Mining, Semantic Web, Domain, FP Algorithm, Sequential Pattern Mining and Hidden Markov Model, Prediction, web log.

1. INTRODUCTION

1.1 Web Usage Mining:

In recent times, Web Usage Mining has emerged as a popular approach in providing Web personalization [1]. Web usage mining is concerned with finding user navigational patterns on the World Wide Web by extracting knowledge from web usage logs (we will refer to them as web logs). The assumption is that a web user can physically access only one web page at any given point in time that represents one item.

The process of [2] Web Usage Mining goes through the following three phases are.

- Preprocessing phase: The main task here is to clean up the web log by removing noisy and irrelevant data. In this phase also, users are identified and their accessed web pages are organized sequentially into sessions according to their access time, and stored in a sequence database.
- Pattern Discovery phase: The core of the mining process is in this phase. Usually, Sequential Pattern Mining (SPM) is used against the cleaned web log to mine all the frequent sequential patterns.
- Recommendation/Prediction phase: Mined Patterns

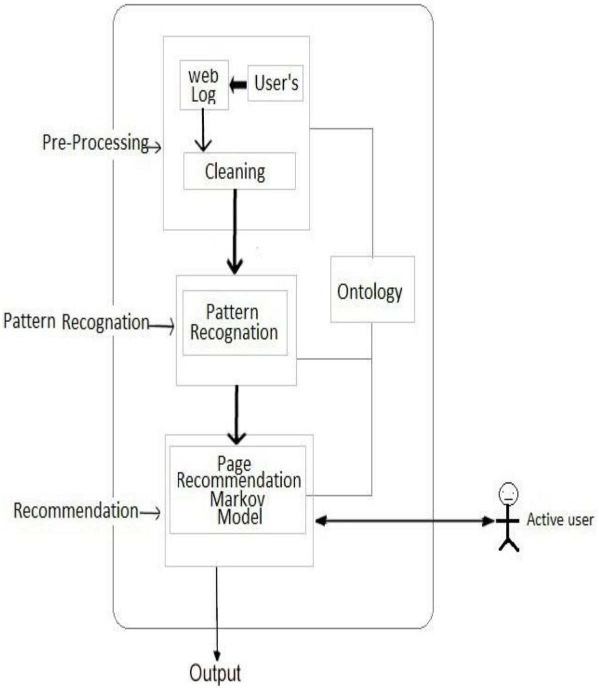
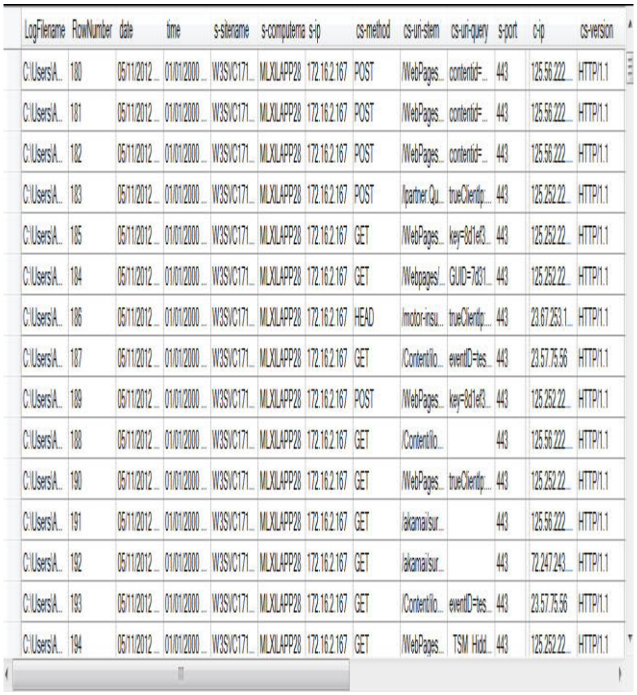


Figure1: Phases of Web Usage Mining

Web Usage Mining is the field of web mining which deals [3] with finding the interesting usage pattern from the logging information. The logging information is stored in a file known as web log file. Web log file contains lot of information like IP address, date, time, web page [4] requested etc.

1.2 Web Log: The web log is a registry of web pages accessed by different users at different times, which can be maintained at the server-side, client-side or at a proxy server, each having its own benefits and drawbacks on finding the users' relevant patterns and navigational sessions [5].

- Server Log: the server stores data regarding requests performed by the client, thus data regard generally just one source. Server Log details are given in Figure 2.



LogFilename	RowNumber	date	time	s-sitename	s-computerna	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	c-ip	cs-version
C:\Users\A...	180	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	contentid=...	443	125.56.222...	HTTP/1.1
C:\Users\A...	181	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	contentid=...	443	125.56.222...	HTTP/1.1
C:\Users\A...	182	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	contentid=...	443	125.56.222...	HTTP/1.1
C:\Users\A...	183	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/partner.Qu...	trueClientIp:...	443	125.252.22...	HTTP/1.1
C:\Users\A...	185	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/WebPages...	key=8d1ef3...	443	125.252.22...	HTTP/1.1
C:\Users\A...	184	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Webpages/...	GUID=7d31...	443	125.252.22...	HTTP/1.1
C:\Users\A...	186	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	HEAD	/motor-insu...	trueClientIp:...	443	23.67.253.1...	HTTP/1.1
C:\Users\A...	187	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Contentlo...	eventID=tes...	443	23.57.75.56	HTTP/1.1
C:\Users\A...	189	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	key=8d1ef3...	443	125.252.22...	HTTP/1.1
C:\Users\A...	188	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Contentlo...		443	125.56.222...	HTTP/1.1
C:\Users\A...	190	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/WebPages...	trueClientIp:...	443	125.252.22...	HTTP/1.1
C:\Users\A...	191	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/akamai/sur...		443	125.56.222...	HTTP/1.1
C:\Users\A...	192	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/akamai/sur...		443	72.247.243...	HTTP/1.1
C:\Users\A...	193	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Contentlo...	eventID=tes...	443	23.57.75.56	HTTP/1.1
C:\Users\A...	194	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/WebPages...	TSM Hidd...	443	125.252.22...	HTTP/1.1

Figure2: A Sample of Serer Side Web Log

- Client Log: it is the client itself which sends to a repository information regarding the user's behavior (can be implemented by using a remote agent (such as Java scripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.);
- Proxy Log: information is stored at the proxy side, thus Web data regards several Websites, but only users whose Web clients pass through the proxy.

2. LITERATURE REVIEW

Due to rapid growth in the number of internet users, the user perceived latency has become a serious issue for the web service providers. Researchers have been done which combines different techniques from multiple domains to overcome this issue.

To reduce perceivable network latency, researchers focused on pre-fetching popular documents. The integration of pre- fetching and caching techniques greatly improves the performance and also reduces the running time of the applications by 50%.

K. R. Suneetha, Dr. R. Krishnamoorthi [12], APRIL 2009

In this research area, they work with the web access logs to enhance the web application contribution over the web usage mining domain which is described as; Web usage mining is application of data mining methods to realize usage patterns from web access data, to better serve the requirements of web applications. The access log files include significant information about a web server access. This paper is focused on the deep investigation of NASA website's Web Log Data to find information of web site, errors, and probable visitors of the site etc. that help administrators and Web designer to enhance the system by finding systems errors, degraded and broken links using web usage mining. The demonstrated results can be used for further development of the web sites to increase its effectiveness.

Cristóbal Romero, Sebastián Ventura, Amelia Zafra, Paul de Bra [13], MAY 2009

Currently, the application of Web mining approaches in e-learning and Web-based educational systems is increasing rapidly. In this paper, Cristóbal Romero et al [13] propose an innovative architecture for a personalization using Web mining. A Web mining tool is developed with recommender engine is integrated. That helps to the instructor to find out whole Web mining procedure. The main aim is to recommend to a student the most appropriate Web pages with in the AHA System. Different experiments are performed with real data in order to test both the architecture. Finally, author describes the meaning of several recommendations.

Akshay Kansara, Swati Patel [7], MAY 2013

With the large amount of information available on web makes it challenging for providers to find relevant information to users in efficient and personalized manner. One manner to handle this issue is to use a recommendation methodology that can offers visitors to find offerings. The usage mining is the main process of mining knowledge from access pattern from web server's logs. In this paper Akshay Kansara et al [7] presents the hybrid approach of the classification and clustering to predict user events.

Bill Karakostas, Babis Theodoulidis [5], JULY 2013

Monitoring user's behavior for large numbers of web site in real time place a performance challenges, due to the decentralized location and amount of generated data. In this paper Bill Karakostas et al [5] proposes Map Reduce-style architecture. The processing of event series of Web users is accomplished by a number of cascading mappers and reducers. Using static analysis a prototype implementation is performed, additionally author demonstrate how architecture is able to obtain time series analysis over real time web data sets, based on the actual events.

Lei Shi, Alexandra I. Cristea, Malik Shahzad Awan, Craig Stewart, Maurice Hendrix [6], AUGUST 2013

Implicit user modeling has played a significant role in personalized web-based e-learning systems and that is highly important in other learning environments. The main objective is to learn from a learner's recent experiences and properties, to find the services for their personal requirements. An experimental study for understanding learning behavior patterns on basis of stronger implicit user modeling mechanisms. Additionally, focuses to get a better observation of learning behavior. The proposed web usage mining and visualization stimulated some interesting learning behavior patterns. Lei Shi et al [6] analyzed these from two perspectives: action frequency and action sequences, based on an efficient classification of behavior patterns. That helped to rank the different action categories according to user's perspective. The results of experiments are promising and present possible instructions for refining user modeling.

R. Suguna, D. Sharmila [3], APRIL 2014

To analyze the user behavior over the web domain R. Suguna et al [3] present a study according to his team, Web mining is the area of data mining. It consists of major three sub areas: (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining. Web usage mining is performed over web log files that are generated at proxy servers, web servers, and browsers as a source of data to identify user's access activities. The user's website access information is recorded in common log format. These logs are huge in size and never found in well format. Therefore, pre-processing is necessary to create a suitable web logs for extracting knowledge. Pre-processed web logs are applied to Pattern analysis techniques to obtain the information. Additionally, a review on the pattern discovery algorithms and clustering algorithms are also provided.

Mayank Kalbhor [1], 2015

Web prediction is a classification problem in which we attempt to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages or browsing history. It is very beneficial to predict user's behavior while he/she is serving the internet for many reasons like increasing browsing speed or minimization of server load etc. In this paper we have proposed fuzzy based hidden Markov model for predicting user's next process. Our proposal is modified version of all-Kth Markov model. This paper also reports the comparisons of various methods for future request prediction with their appropriate application. This complete work provide us the overview of user request prediction and also provide its possible solutions which will help in further research for new development in this area.

3. PROBLEM DOMAIN

3.1 Problem Definition:

Web mining is categorized in the three major domains, first web content mining where the web page contents are analyzed to find the important data over the web pages. In next the web structure mining, using technique organization of the web pages, structural information of the web pages and links between different pages are investigated. Finally, in the web usage mining, the web usage data, in other words the web accessing data is utilized for finding the access patterns, user interest, and web recommender data.

The proposed work is motivated to design and develop a predictive data model by which the user data access pattern can be investigated. The following issues are focused for improving the traditional data analysis approaches.

1. Traditional web usage mining techniques don't seem to be abundant economical for analyzing information.
2. As the size of data increases, the resources consumption of algorithm is also affected.

3. Accuracy of prediction is not much accurate due to outliers, data spicks.

3.2 Solution Domain

To overcome the above listed issues in predicting the web user's next page, it is required to design a new data model by which the problem is satisfied. To evaluate the large scale data a divide and conqueror method is suitable. Therefore, improved FP Growth algorithm is used for frequent pattern analysis. In this algorithm when the large data sets come into the existence, it divides the large datasets in small parts and then applies FP tree method on it. By using this algorithm the accuracy to predict web user's next page is gradually increases as compared to previous work. On the other hand the available data includes the number of unique web pages and their access patterns thus, required to optimize the algorithm by pre-processing of the data. Therefore, HMM algorithm is applied to keep in track the selection process of improved FP algorithm. That algorithm helps to find the duplicate data patterns from the available set of data. For analyzing the nearest data patterns over number of session algorithm, concept of Euclidian distance is used and the similar data is eliminated from the data set. This process reduces the amount of data for evaluation. The combination of two different models is leads to design a hybrid approach for data analysis.

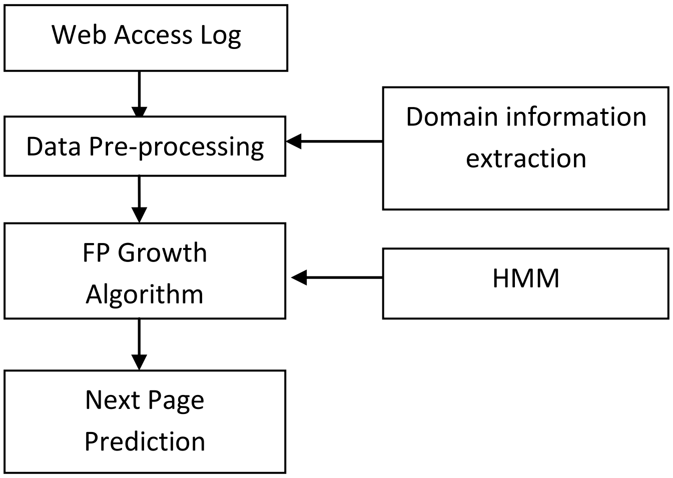


Figure 3.1 Proposed Architecture

4. PROPOSED WORK

We propose to formulate an FP Growth Algorithm to implement the solution to the problem formulation.

Algorithm: FP_Growth (WebLog)

Step 1: Generation of web log data.-The data is generated when the users access/ create any information over the internet. The weblogs are created by the web servers.

Step 2: Extraction of web log data.-The web log data is of prime importance in the entire process. Web log data extraction is done using software.

Step 3: ETL process.-It does the extraction, transformation and loading of the data extracted from the weblogs. This is also called as cleaning of data. This removes all the abnormalities from the data and makes it ready for use by the algorithm.

Step 4: Application of algorithm. The algorithm used here is the F.P Growth algorithm. It is applied to the data obtained from the ETL process. It mines the data and finds the frequent patterns in the data. It is a two- step process. It concludes by forming a F.P Tree.

Step 5: Pattern discovery.-The frequent patterns mined by the algorithm are discovered and highlighted.

Step 6: Pattern analysis. The discovered patterns are analysed and are used for distinguishing different categories of data.

Step 7: Customizations.

The F.P Tree algorithm works as follows-The proposed approach that we plan to implement follows the following steps:

Step1: In the first steps data is being collected from the Web log file and then Preprocessing is applied. In the Preprocessing the Data is being loaded and it is being converted in to the Data set having fields Client-IP, Session_ID, Country, Access Date Time, Method, URL, URL_ID, Protocol, Status, Bytes transferred. The session is calculated in 30 minutes interval of time, after 30 minutes the system will recognize the same user as next user.

Step 2: In this step there is Pattern Discovery which is performed by the Frequent Pattern (FP) which involves FP Tree which in turn FP growth .FP tree method is used in Data Mining .It consists of two passes over the Data Set .In the first Pass it scans data and find the minimum support for the each item. The item set whose support is less than minimum is discarded .The Data item that is included is the Web Site or the URL that is being visited by the User. Next steps in the First Pass in the FP tree are to generate a decreasing order on the basis of frequency of occurrence of the Item Set Which is the URL visited by the User. In the Second Pass of the FP Tree Transaction is being read .In this work the Transaction is the number of user visited the particular Web Site. The Read Transaction is iterated until all the Transaction is being completed. After Reading all the Transaction discards all the transaction which has lees support or support than the minimum threshold value.

Step3: In this step Pattern analysis is done and in this Candidate rule is generated and on the basis of candidate rule confidence is generated. On the basis of pattern analysis Prediction is done of the User's Future request.

The integration of semantic information directly in the transition probability matrix of lower order Markov models, was presented as a solution to this tradeoff problem [17]. This integration also solves the problem of contradicting prediction. , we propose to use semantic information as a criteria for pruning states in higher order (where k > 2) Selective Hidden Markov models [4], and compare the accuracy and model size

of this idea with semantic-rich hidden markov models and with traditional Hidden Markov models.

Hidden Markov Model as a proposed solution to prove semantically meaningful and accurate predictions without using complicated all Kth order. The semantic distance matrix Weight Matrix and Transition Matrix is directly used in hidden markov model

5. RESULT

We evaluate our proposed system on different parameters, which describe below:

- Success Rate
- Failure Rate
- Training Time
- Memory

1. Success Rate

The success rate of the data model is evaluated using the N-cross validation method. According to the validation methodology, the numbers of correctly identified patterns are known as the model success rate. The below given figure 5.1 provide the evaluated comparative accuracy of the proposed hybrid algorithm with the traditional markov model. For calculating the success rate, the following formula is used.

$$Success\ rate = \frac{correctly\ identified\ patterns}{patterns\ avilable\ for\ evaluation} X\ 100$$

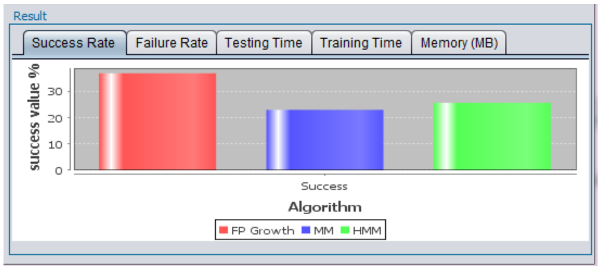


Figure 5.1 Success rate graph for FP Growth, Markov and Hidden Markov Model

We calculate success rate value for all algorithms FP Growth, Markov and Hidden Markov Model. And results shown with help of diagram. We find that FP Growth approach shows more success rate compare to Markov Model & Hidden Markov Model approaches.

2. Failure Rate:

The failure rate of the system is inversely proportional to the accuracy obtained, in terms of the percentage is calculated using the below given formula.

$$FAILURE\ RATE = 100 - success\ rate$$

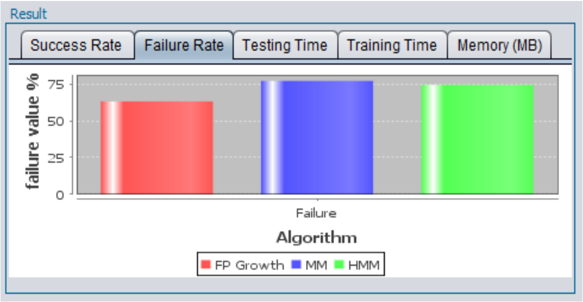


Figure 5.2 Failure rate graph for FP Growth, Markov and Hidden Markov Model

We calculate failure rate value for all algorithms FP Growth, Markov and Hidden Markov Model. And results shown with help of diagram. We find that FP Growth approach shows less failure rate compare to Markov Model & Hidden Markov Model approaches.

3. Training Time

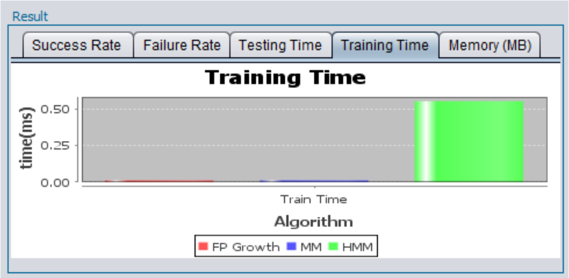


Figure 5.3 Training Time graph for FP Growth, Markov and Hidden Markov Model

We calculate Training Time value for all algorithms FP Growth, Markov and Hidden Markov Model. And results shown with help of diagram. We find that FP Growth approach shows more success rate compare to Markov Model & Hidden Markov Model approaches.

4. Memory

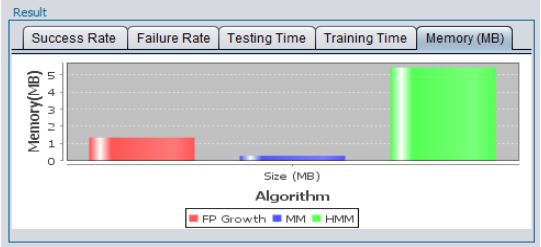


Figure 5.4 Memory uses graph for FP Growth, Markov and Hidden Markov Model

We calculate memory required during runtime for all algorithms FP Growth, Markov and Hidden Markov Model. And results shown with help of diagram. We find that FP Growth approach requires less memory as compared to Markov Model & Hidden Markov Model approaches.

6. CONCLUSION

Web usage mining model is kind of mining to server logs. Web usage mining used for the improvement of improving the requirement of the system performance, the customers relation and realizing enhancing the usability of the website design. The main goal of the proposed system is to identify usage pattern from web log files. FP Growth Algorithm is used for this purpose. Apriori is a classic algorithm for association rule mining. The main drawback of Apriori algorithm is that the candidate set generation is costly, especially if a large number of patterns and/or long patterns exist. The main drawback of Markov model is that it can't focus on the hidden stages of web log files. The FP-growth algorithm is one of the fastest approaches for frequent item set mining. The FP-growth algorithm uses the FP-tree data structure to achieve a condensed representation of the database transaction and employees a divide-and conquer approach to decompose the mining problem. Our experimental result shows that the FP-growth & HMM method is efficient and scalable for mining both long and short frequent patterns. In future the algorithm can be extended to web content mining, web structure mining.

REFERENCES

[1] Mayank Kalbhor [1] "Fuzzy Based Hybrid Approach for User Request Prediction Using Markov Model" [IEEE International Conference on Computer, Communication and Control (IC4-2015)]

[2] Meera Narvekara, Shaikh Sakina Banu "Predicting User's Web Navigation Behaviour Using Hybrid Approach" International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

[3] R. Suguna, D. Sharmila, “Clustering Web Log Files – A Review”, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 4, April – 2014, ISSN: 2278-0181

[4] Mirghani. A. Eltahir, Anour F.A. Dafa-Alla, “Extracting Knowledge from Web Server Logs Using Web Usage Mining”, 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE),2013 IEEE

[5]Bill Karakostas, Babis Theodoulidis, “A Map Reduce Architecture for Web Site User Behaviour Monitoring in Real Time”, 2nd International Conference on Data Management Technologies and Applications (DATA), 29 - 31 July 2013, Reykjavik, Iceland

[6]Lei Shi, Alexandra I. Cristea, Malik Shahzad Awan, Craig Stewart, Maurice Hendrix, “Towards Understanding Learning Behaviour Patterns in Social Adaptive Personalized E-Learning Systems”, Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois, August 15-17, 2013

[7]Akshay Kansara, Swati Patel, “Improved Approach to Predict user Future Sessions using Classification and Clustering”, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064,Volume 2 Issue 5, May 2013, www.ijsr.net

[8]V.Chitraa, Dr.Antony Selvadoss Thanamani, ”Web Log Data Cleaning For Enhancing Mining Process”, International Journal of Communication and Computer Technologies, Volume 01 – No.11, Issue: 03 December 2012 ISSN NUMBER : 2278-9723

[9]Kavita Das,O. P. Vyas, “Issues of Learning the Browsing Language”, International Journal of Computer Applications (0975 – 8887)Volume 14– No.5, January 2011

[10]Hamid Rastegari and Siti Mariyam Shamsuddin, “Web Search Personalization Based on Browsing History by Artificial Immune System”, Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 3, November 2010 ISSN 2074-8523; Copyright © ICSRS Publication, 2010

[11]S.Vijayalakshmi, V.Mohan, S.Suresh Raja, “Mining Of Users’ Access Behaviour for Frequent Sequential Pattern From Web Logs”, International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2010, DOI : 10.5121/ijdms.2010.2304

[12]K. R. Suneetha, Dr. R. Krishnamoorthi, “Identifying User Behaviour by Analyzing Web Server Access Log File”, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009, 327

[13]Cristóbal Romero, Sebastián Ventura, Amelia Zafra, Paul de Bra, “Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems”,0360-1315/$ - see front matter 2009 Elsevier Ltd. All rights reserved.doi:10.1016/j.compedu.2009.05.003