INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

AN EXTENDING RECOMMENDATION SYSTEM FOR WEB INFORMATION RETRIEVAL

HIMANGNI RATHORE[1], HEMANT VERMA[2]

1. Computer Engineering Department, Vindhya Group of Technology and Science, Khandwa Road, Indore, India
2. Computer Engineering Department, Vindhya Group of Technology and Science, Khandwa Road, Indore, India

Accepted Date: 11/10/2015; Published Date: 01/11/2015

Abstract: - Web is a huge source of information a number of internet users visit on different web sites and extract their required data. That is direct source of information which is used by end client. On the other hand some additional data generated on the parked domain web server which is used by web site administrator and used for deciding the future business trends and future service planning. That essential information is recovered from the web server log files, knowledge extraction from these raw files are also called the web usage mining. In this presented work web usage mining is investigated and a new data model for web recommendation is reported. in order to develop the proposed recommender system the user session web accessed log data is accessed and classified on the basis of the time based fashion. This kind of analysis demonstrates the user web access browsing behaviour in different time slots. Thus according to the user behaviour analysis in different time domains a predictive model namely hidden Markov model is applied on the recovered data. That uses the probability estimation techniques for finding the new navigational web access trend. The proposed data model is implemented using the visual studio environment and the performance of the predictive algorithm is computed. The performance of the implemented system is evaluated in terms of accuracy, memory consumption, error rate and time consumption. According to the obtained results the presented technique enhancing the performance as the training data is increases.

Keywords: Web mining, Recommender systems, Collaborative filtering, HMM, K-Nearest neighbor, Web access log file.

Corresponding Author: MS. HIMANGNI RATHORE

Access Online On:

www.ijpret.com

How to Cite This Article:

Himangni Rathore, IJPRET, 2015; Volume 4 (3): 93-105

PAPER-QR CODE

INTRODUCTION

Web is a diverse source of information where different kinds of knowledge and data are available. Some of the information is direct recoverable form web pages and some information is found in hidden formats such as web access log and web link organizations. According to the nature of data availability the web mining techniques are classified as web usage mining, content mining and structure mining. Data mining algorithms are implemented to find the knowledgeable pattern in such type of data. In this proposed work the web accessed log is analysed for knowledge discovery. Web access log analysis is also termed as web usage mining. Basically the web servers contains more than one websites and for keep track the traffic information a log file is managed. This web access log files contains the entire information for each user request and their response. Such log file is known as web access log where all the users' access data or web usage information is available. In this proposed work a new recommendation system is investigated and designed. The recommendation system are the data analysis technique by which users previous or historical navigational patterns are analysed and based on their navigational behaviour future trends are predicted. For that purpose recommendation system consist of predictive algorithms and clustering techniques. The predicted future trends of user navigational patterns are help to understand which kind of data a user is looking and searching. Thus these systems are much helpful for making heuristics in e-commerce web sites, social networking web sites and others.

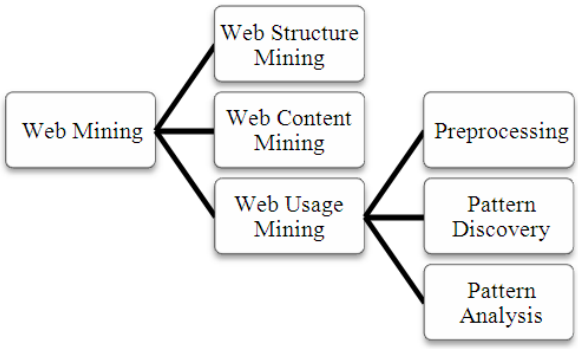

Figure 1 Web mining

OBJECTIVES

The key objective of the presented study is to develop and design an accurate recommendation engine using web access log analysis. In order to develop such system the following tasks are involved in this work.

Investigation of web access log analysis techniques: in this phase the different web log analysis and attributes are studied.

Analysis of recently developed recommendation systems: in this phase different web based recommendation systems are studied and their applications are investigated.

Implementation and design of a hybrid recommendation system: in this phase using the optimum techniques a new recommendation system is developed.

Performance study of proposed recommendation systems: in this phase the implemented recommendation system's performance is evaluated.

METHODOLOGY

A recommender engine is designed using a hybrid approach in which two different classifiers are combined for data analysis and prediction of next opened web URL. Now in these days the web based applications and the applications of web based techniques are growing in rapid manner. Therefore new development techniques and their supportive techniques are also rapidly growing. The main reason behind the popularity of web applications is the availability and connectivity. By which the product vendors and end clients are directly connected to each other and can make service request and services any time when required. In the similar way for finding the end client need and understanding of the required data by end client's recommender systems are developed. These recommendation systems analyse the historical user web access patterns and predict the navigational directions. Such kinds of system are much helpful in e-commerce development. In this presented work the web based recommendation system are investigated and based on available optimum techniques a new hybrid recommendation system is developed. The proposed recommendation system consumes the client end accessed web log and based on the different user sessions frequent accessed data is recovered. These frequent access patterns are help to find interest of a user. In addition of that for personalizing the user data KNN algorithm is applied finally using the HMM (hidden Markov Model) the predictive system is developed. This predictive system accepts the current user accessed data sequence and based on the current navigation future trends are predicted. There are number of data sources available for web data information. In this the client end accessed data is analysed.

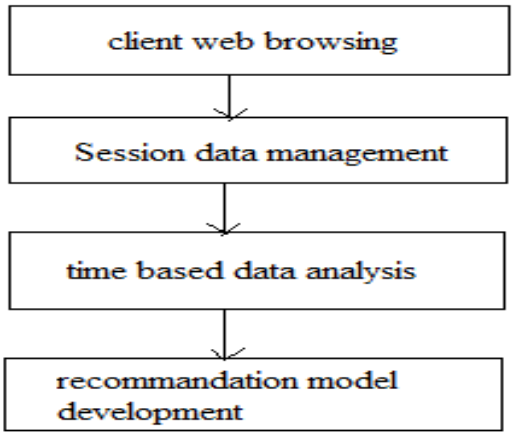

Figure 2 Layered module process

The main advantage of client end web data analysis is that the entire web access navigational information be obtainable at the client end. Thus when the user navigated through their web browser the system extracted the navigated web page information. This information is preserved in a data base table. in a single system that is possible more than one user access the web data. Thus for identification of user accessed data, the data is preserved according to the user sessions. Thus when browser is closed the session is end for a user and for new instance of browser a new session is activated. Due to literature collection that is observed the user behaviour of web navigation is fluctuating according to the time. Therefore a single user never follows similar access patterns in a day. Therefore the data is personalized according to the time based fashion. Using the observation of different session based web data access a predictive model is prepared namely hidden Markov model. This system needs to develop two different kind of data access patterns first the transitional information and secondly the observational information. These data matrix are used with the current navigational patterns of data access.

System architecture

It shows the proposed recommendation system architecture in this diagram the end client is simulated who are making request from web for data. The browser information and navigated URL is accessed using the system and managed according to the user session as described in above. The session-wise log data is then classified using KNN algorithm where the time is provided as query to the database. The classified access log data is then used with the Hidden Markov Model. The HMM model process the time based classified data and next web page is recommended with the performance outcomes.

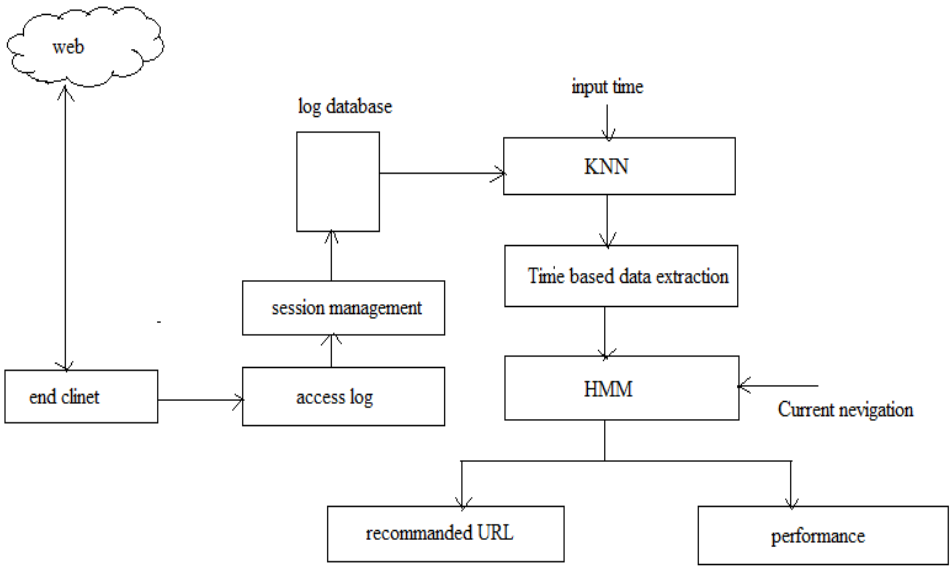

Figure 3 Architecture

Algorithm study

Hidden Markov Model is a double implanted stochastic process with two hierarchy levels. It can be used to model much more complex stochastic processes as compared to a traditional Markov model. In a specific state, an observation can be generated according to an associated probability distribution. It is only the observation and not the state that is visible to an external observer [20]. N is the number of states in the model. We denote the set of states' S = {S1; S2; ..., SN}, where Si, i= 1;2;...;N is an individual state. The state at time instant t is denoted by qt. M is the number of distinct observation symbols per state. We denote the set of symbols V = {V1; V2; ...VM}, where Vi, I = 1; 2; ...; M is an individual symbol. The state transition probability matrix A = [aij], where

$$a_{ij} = P\left(q_{t+1} = S_j \middle| q_t = S_i\right), 1 \le i \le N, 1 \le j \le N; t = 1,2 \ldots$$

Here $a_{ij} > 0$ for all i, j. Also,

$$\sum_{j=1}^{N} a_{ij} = 1\,, 1 \le i \le N$$

The remark symbol probability matrix B = {$b_j(k)$}, where

$$b_j(k) = P\left(V_k \middle| S_j\right), 1 \le j \le N, 1 \le k \le M \; and$$

$$\sum_{k=1}^{M} b_j(k) = 1\,, 1 \le j \le N$$

The initial state probability vector r= πi , where

$$\pi_i = P(q_1 = S_i), 1 \le i \le N$$

Such that

$$\sum_{i=1}^{N} \pi_i = 1$$

The remark sequence O = O_1; O_2; O_3; ...O_R, where each remark O_t is one of the symbols from V, and R is the number of remarks in the sequence. It is manifest that a complete specification of an HMM needs the approximation of two model parameters, N and M, and three possibility distributions A, B, and π. We use the notation ʌ = (A; B; π) to specify the complete set of parameters of the model, where A, B implicitly contain N and M. An observation sequence O, as mentioned above, can be generated by many possible state sequences. Consider one such particular sequence Q = q_1; q_2; ...; q_R; where q1 is the initial state. The probability that O is generated from this state sequence is given by

$P(O|Q,\lambda) = \prod_{t=1}^{R} P(O_t|q_t,\lambda)$ Where statistical independence of observations is assumed Above Equation can be expanded as

$P(O|Q,\lambda) = b_{q1}(O_1)b_{q2}(O_2) \dots \dots b_{qR}(O_R)$ The probability of the state sequence Q is given as $P(Q|\lambda) = \pi_{q1}a_{q1q2}a_{q2q3} \dots \dots a_{qR-1qR}$ Thus, the probability of generation of the observation sequence O by the HMM specified by can be written as follows:

$$P(O|\lambda) = \sum_{all\ Q} P(O|Q,\lambda)\,P(Q|\lambda)$$

Deriving the value of $P(O|\lambda)$ using the direct definition of is computationally intensive. Hence, a procedure named as Forward-Backward procedure is used to compute $P(O|\lambda)$.

The **K-nearest-neighbour** (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set [21]. We can compute the distance between two scenarios using some distance function d(x, y), where x, y are scenarios composed of features, such that

$$X = \{x_1 x_2, x_3, \dots\}$$

$Y = \{y_1, y_2, y_3, \ldots\}$ Two distance functions are discussed here:

Absolute distance measuring: $d_A(x, y) = \sum_{i=1}^{N} |x_i - y_i|$

Euclidean distance measuring: $d_E(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$

Because the distance between two scenarios is dependant of the breaks, it is suggested that resulting distances be scaled such that the arithmetic mean across the dataset is 0 and the standard deviation is 1. This can be accomplished by replacing the scalars with according to the following function: $x^{"} = \frac{x - x'}{\sigma(x)}$ Where the un-scaled value is the arithmetic mean of feature across the data set, is its standard deviation, and is the resulting scaled value.

The arithmetic mean is defined as:$x' = \frac{1}{N} \sum_{i=1}^{N} x_i$

We can then compute the standard deviation as follows:$\sigma(x) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - x')^2}$

Distance functions As stated previously, we are only considering absolute and Euclidean distance functions. However, we may choose to provide the original unscaled values, or uses transform them using the scaling function. **K-nearest-neighbour** Now that we have established a measure in which to determine the distance between two scenarios, we can simply pass through the data set, one situation at a time, and compare it to the query scenario. We can represent our data set as a matrix D= N x P, containing p scenarios S_1…… S_p, where each scenario Si contains N features S_i = {S_1…….. S_n} A vector O with length P of output values O = {O_1…….. O_p} accompanies this matrix, listing the output value O_i for each scenario S_i. It should be prominent that the vector can also be seen as a column matrix; if multiple output values are desired, the width of the matrix may be expanded.

KNN can be run in these steps: Store the output values of the M nearest neighbours to query scenario Q in vector r = {r_1……r_m} by repeating the following loop M times: Go to the next scenario S_i in the data set, where I is the current iteration within the domain {1……P} If Q is not set or q < d (q, S_i): q ←d (q, S_i), t ←O_i Loop until we reach the end of the data set. Store q into vector c and t into vector r. Calculate the arithmetic mean output across r as follows:

$$r = \frac{1}{M} \sum_{i=1}^{M} r_i$$

Return r as the output value for the query scenario q.

Result

Accuracy In the predictive data models the amount of correctly identified patterns are known as the accuracy of the predictive system. That can be calculated using the following formula.

$$accuracy = \frac{total\ correctly\ classified\ samples}{total\ input\ samples} X100$$

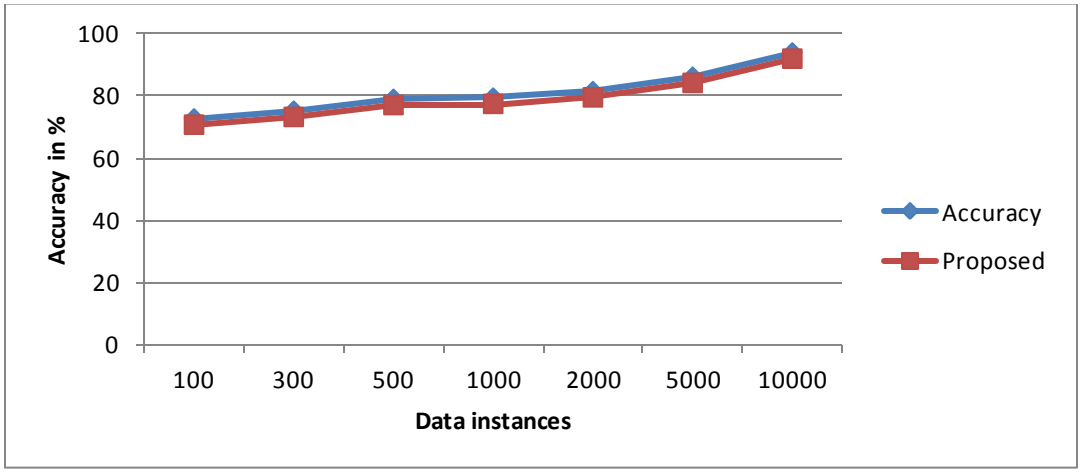

Figure 4 Accuracy

Error rate of the proposed recommender engine is given using figure 5.2 where the X axis contains the training samples as input and the Y axis shows the error rate of system. According to the evaluated results the error rate of the system is decreases as the amount of training samples are increase.

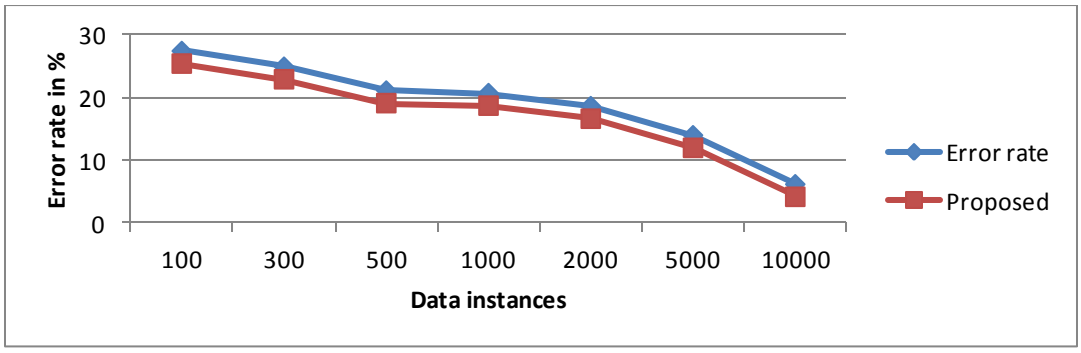

Figure 5 Error rate

Thus the proposed model is adoptable and enhancing their learning capability of system as the samples in data base is increases.

Memory used the amount of main memory required to execute the algorithm for developing the data model and providing predictive outcomes is termed as memory consumption of the system. The memory consumption of the proposed data model is given using figure

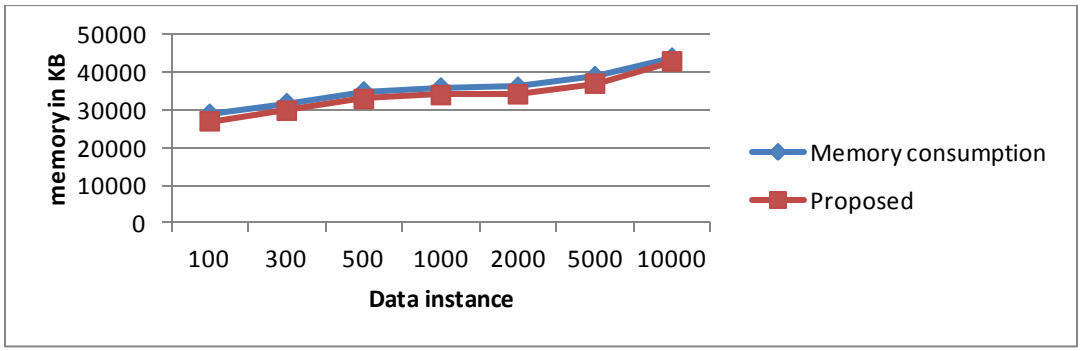

Figure 6 Memory consumption

in this diagram the amount of memory consumption is given in Y axis and the X axis shows the amount of training samples on which the data model is developed. According to the obtained results as the amount of training data is increase the memory consumption of the system is increases.

Time complexity is the amount of time consumed for developing the data model is known as the time complexity of the system. For simulating the performance in terms of time consumption the X axis shows the amount of data produced for training and the Y axis shows the time consumed in terms of seconds. According to the evaluated performance the time consumption for model development is increases as the amount of data for training is increases.

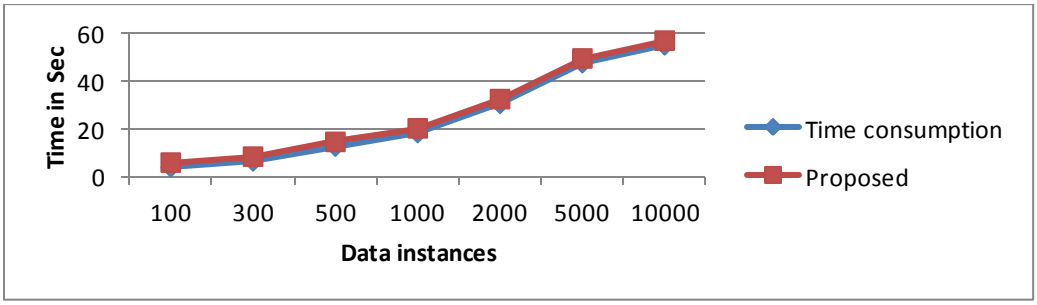

Figure 7 Time consumption

CONCLUSION

Web mining is a technique where the web data is analysed for finding or extracting the knowledgeable information. This information can be used with different kinds of application development and planning. In this presented work the web usage data is analysed for discovering the user navigational patterns and these patterns utilized for predicting the next web page according to the current trend of web navigation. Therefore the proposed study is focused on web usage data analysis, user data personalization, user navigational pattern estimation and future trend prediction. By incorporating all these requirements the web recommender engine is developed. Thus in order to design such data analysis methodology the web access log data is consumed and using the pre-processing techniques the desired attributes for evaluation is extracted. After that for personalization of web usage data the KNN classification algorithm is employed which collect the similar user data for analysis this selected user data is used with the hidden Markov model. Thus first the transition matrix and observational matrix is prepared and using the HMM computations the next web page is predicted for current sequence of input navigational pattern. The performance of predictive system is estimated in terms of accuracy, error rate, and memory usage and time complexity. The evaluated performance is summarized using table 2.

Parameters	***Remark***
Accuracy	*High accurate predictive results, and performance is enhanced when the training data is increase*
Error rate	*Low error rate because when training input increases the patterns for learning is increases and error rate is decreases*
Memory usage	*That is depends on the input training set, as the amount of data for training is increases the memory consumption of the system increases.*
Time consumption	*That is also depends on the size of training data, as the amount of data increases the amount of time for data analysis is also increases.*

ACKNOWLEDGMENT

This research paper is made possible through the help and support from everyone, including: parents, teachers, family, friends, and in essence, all sentient beings. I would like to express my

deep gratitude to Mr. Hemant Verma, my research supervisors, for their patient guidance, enthusiastic encourage-ment and useful critiques of this research work. He kindly read my paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper. for his inspiring and encouraging attitude. Initially working out to this would have been a very tedious job if proper support was not presented from his side. I would also like to thank Mark F. Hornick, Katrien Verbert, Martin Wolpers, Hao Ma, Irwin King, and many more who provides me ancillary guideline to achieve desired goal. The concerning paper is only used for educational research and development purpose that is not tested for industrial protocols.

Finally, I offer my deep gratitude to my parents who have appreciated, encouraged and assisted in our endeavor.

REFERENCES

1. Mark F. Hornick, and Pablo Tamayo, "Extending Recommender Systems for Disjoint User/Item Sets: The Conference Recommendation Problem", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 8, August 2012

2. Bussa V. R. R. Nagarjuna, Akula Ratnababu, Miriyala Markandeyulu, A. S. K. Ratnam, "Web Mining: Methodologies, Algorithms and Applications", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-3, July 2012

3. Sheetal Chouhan, Manish Shrivastava and Kavita Deshmukh, "A Noble Approach of Web Log Mining", VSRD-IJCSIT, Vol. 2 (7), 2012, 590-596

4. Bhaiyalal Birla, Sachin Patel, "An Implementation on Web Log Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014

5. M. Vengateshwaran,E. V. R. M Kalaimani, "Web Mining Research Direction and Open Source Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014

6. Personalization: Collaborative Filtering vs Prediction Based on Benefit Theory, November 05, 2007,http://myshoppal.typepad.com/blog/2007/11/personalization.html

7. Magdalini P. Eirinak, "New Approaches to Web Personalization", Ph.D. Thesis, Athens University of Economics and Business Dept. of Informatics May 2006

8. "Recommendation Systems", http://infolab.stanford.edu/~ullman/mmds/ch9.pdf

9. "Data Mining - Classification & Prediction Introduction", http://www.idc-online.com/technical_references/pdfs/data_communications/Data_Mining_Classification_Prediction.pdf

10. Ida Mele, "Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content", WSDM'13, February 4–8, 2013, Rome, Italy, Copyright 2013 ACM

11. Suresh Shirgave and Prakash Kulkarni, "Semantically Enriched Web Usage Mining For Predicting User Future Movements" ,International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.4, October 2013

12. Kamika Chaudhary, Santosh Kumar Gupta, "Web Usage Mining Tools & Techniques: A Survey", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013

13. Maryam Jafari, Farzad Soleymani Sabzchi and Amir Jalili Irani, "Applying Web Usage Mining Techniques to Design Effective Web Recommendation Systems: A Case Study", ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 2, No.8 , March 2014

14. Ms. Shital C. Patil, Prof. R. R. Keole, "The Role of Web Content Mining and Web Usage Mining in Improving Search Result Delivery", IJCSMC, Vol. 3, Issue. 3, March 2014, pg.7 – 14

15. P. Senthil Pandian, Dr. S. Srinivasan, "Perfections And Psychiatry User Profile In Web Sites Using Web Usage Mining & Clustering Sesions", International Journal of Research in Computer and Communication Technology, Vol 2, Isue 2, Feb-2013

16. Mr. Akshay Upadhyay, Mr.Balram Purswani, "Web Usage Mining has Pattern Discovery", International Journal of Scientific and Research Publications, Volume 3, Issue 2, February 2013

17. Yılmaz Atay and Halife Kodaz, "Application of Web Mining Using Clonal Selection Algorithm", Lecture Notes on Software Engineering, Vol. 1, No. 3, August 2013

18. Farzad Soleymani Sabzchi, Shahram Jamali, Maryam Jafari, "Mining users` navigation patterns for building web pages recommendation system", Journal of Advances in Computer Research(Vol. 4, No. 2, May 2013), Pages: 15-24

19. D.A. Adeniyi, Z. Wai, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics (2014)No. of Pages 23

20. Shweta Jaiswal, Atish Mishra, Praveen Bhanodia, "Grid Host Load Prediction Using Grid Sim Simulation and Hidden Markov Model", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 7, July 2014

21. Rui Zhang, H. V. Jagadish, Bing Tian Dai, Kotagiri Ramamohanarao, "Optimized Algorithms for Predictive Range and KNN Queries on Moving Object", Volume 35, Issue 8, December 2010, Pages 911–932, 2010 Elsevier

22. Mining Web Graphs for Recommendations Hao Ma, Irwin King, *Senior Member, IEEE,* and Michael R. Lyu, *Fellow, IEEE*

23. Context-Aware Recommender Systems for Learning: A Survey and Future Challenges Katrien Verbert, Member, IEEE, Nikos Manouselis, Member, IEEE, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, Student Member, IEEE, and Erik Duval, Member, IEEE