Methodologies on User Behavior Analysis and Future Request Prediction in Web Usage Mining using Data mining Techniques

M.SelviMohana[1], B.Rosiline Jeetha[2]

[1]M.Phil Research Scholar, [2]Associate Professor,

[1, 2] Department of Computer Science, RVS College of Arts & Science, Sulur, Coimbatore

Abstract

Web Usage Mining is a kind of web mining which provides knowledge about user navigation behavior and gets the interesting patterns from web. Web usage mining refers to the mechanical invention and scrutiny of patterns in click stream and linked data treated as a consequence of user interactions with web resources on one or more web sites. Identify the need and interest of the user and its useful for upgrade web Sources. Web site developers they can update their web site according to their attention. In this paper discuss about the different types of Methodologies which has been carried out in previous research work for Discovering User Behavior and Predicting the Future Request.

Index Terms: Web Mining, Web Usage Mining, Behavior Analysis, Future Request Prediction

1. Introduction

Web mining which is a type of data mining is used to extract web data from web pages. As data mining basically deals with the structured form of data, web mining deals with the unstructured and semi structured form of data. applications Web mining is an application of data mining which uses data mining techniques to extract useful information from web documents Web mining consist of three techniques i.e. web content mining, web structure mining and web usage mining for web data extraction

Web Content Mining: Content mining deals with extraction of data from the content of WebPages based upon pattern matching

Web Structure Mining: Describe relational structure of the WebPages and used to extract information from hyperlink structures.

Web Usage Mining: Usage Mining is the application of data mining technique to discover information from the web log data in order to understand and better serve the needs of Web based. Web usage mining is a process of mining useful information from server logs. This paper describes the methods which already used in past research work for analyzing the user behavior and predicting the future request.

2. Review of Literature

1.**[**Dilpreet Kaur] proposed an Efficient User Future Request Prediction Using KFCM. Predicts the user browsing behavior of user using Fuzzy. Finding the web pages with highest grade membership in each cluster. To overcome heavy traffic delay in response using future request. Prediction based on Session Oriented/page Oriented.

2.**[** Neeraj Raheja]Focused on Efficient Web Data Extraction Using Clustering Approach for web usage mining using cluster formulation. The results of cluster based web log searching are compared with the results of complete web log based searching i.e. caching of documents in the web log.

3.[M.Rathamani]has done the work on Cloud Mining: Web usage mining and user behavior analysis using fuzzy C-means clustering .Analyze the relation between the structure of website and log file using hybrid Clustering with Content mining.

4.[Shaily G. Langhnoja] focused on Discover Visitor Group with Common Behavior Using DBSCAN Clustering Algorithm.Finding visitor group with common behavior using local density of database using one parameter. Number of starts from estimated density distribution of corresponding nodes.

5.[Hemant N.Randhi]has done survey on Browsing Behavior of a User and

Subsequently to Predict Desired Pages: A Survey. This paper is a survey of recent work in the field of web usage mining for the benefit of research on the web log files of Web-based information services.

6.[Poonam Kaushal]proposed an Analysis of User Behavior by Hybrid Technique for predicting user next page request in a combination of Markov model and Nearest Neighbor model. Cluster data grouped into classes. It improves the performance of web page access time.

7.[Akshay Kansara] presents an Improved Approach to Predict user Future Sessions using Classification and Clustering for classification identifying potential user and clustering group potential users with similar interest. K means is an partition clustering based on distance measured by Euclidean distance.

8.[Alexandros Nanopoulos] Data mining approach on Effective Prediction of Web-user Access. The factors are order of dependency between web pages, prior present in user access due to random access of user, ordering of accessing within sequence.

9.[R. Khanchana]focused on Usage Mining Approach Based on New Technique In Web Path Recommendation Systems web page ranking is significant problem web pre fetching is to reduce latency access. Two Bias features considered for fetching the data .It takes length of time spent on visiting a page and frequency of visited pages.

3.Methodologies of Developed Work

These papers illustrate the methodologies which have been used in previous research work on user behavior analysis in web usages and predicting future user movements.

i)Future Request Prediction KFCM(Kernelized Fuzzy C Means)

[Dilpreet Kaur][1].

KFCM is an algorithm which is derived from FCM modifications can be done in objective function using Kernel induced distance matrix instead of Euclidean distance and called as kernelized fuzzy c-means (KFCM) algorithm. Cluster can be identified using membership grade. User session ids and Webpage can be taken. Find the web pages with highest grade membership in each and every cluster. Allot credits to each webpage according to grade membership, page with highest credit point has higher membership and page with low credit point has low membership. The webpage which has high membership grade probably for opening in future by user.

Association Rule, Frequent Sequence, Frequent Generalized Sequence

[Ujwala Patil][1]

Association Rule is a way for discovering relations between variables in large database. Frequent Sequences: This technique is to discover time ordered sequences of URLs that have been followed by past users. Frequent Generalized Sequences: A generalized sequence is a sequence allowing wildcards in order to reflect the user's navigation in a flexible way. In order to extract frequent generalized subsequences they have used the generalized algorithm. Association rule mining is used to discover useful common browsing patterns and then to predict the further browsing sequences. Navigation pattern can be taken to compare the similarity with predicted model therefore the candidate upcoming request can be predicted and to improve the browsing performance.

Classification and K Means Clustering

[Akshay Kansara][1]

Classification is a technique referred as a collection of records. Training and Testing data two phases building and validating the data main process in classification. Clustering is used to group the similar things and not similar is called outlier. Classification identifies promising users from weblog data. Clustering group the promising users with comparable interest. Web page clustering is done by grouping pages having parallel content. Page clustering can be done the Web site is structured Orderly. The K-Means algorithm is a partition clustering and based on distance, unconfirmed. This algorithm make a cluster randomly so it's have different points. Clusters are compared using the different distances within clusters and the least sum of distances is considered as a result, *k*-means algorithm deals with the total number of clusters (*k*), the total number of runs and the distance measured.

Discover and Extracting User Behavior DBSCAN clustering algorithm

[Shaily G. Langhnoja][1]

The DBSCAN algorithm can identify clusters in large spatial data sets by looking at the local density of database elements, using only one input parameter. It is density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN requires two parameters: (eps) and Minimum number of points required to form a cluster (minPts).

This point's -neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. If a point is found to be a dense part of a cluster, its -neighborhood is also part of that cluster. Hence, all points that are found within the -neighborhood are added, as is their own-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found.

Support Vector Machine

[Poonam Kaushal[1]]

SVM (Support Vector Machine)

SVM regression technique is used for identifying user's prefetching used for analyzing the data and recognizes patterns, it is supervised learning technique. Prediction can be done by using hit rate. SVR prediction concern with request page on server and then get next page, with this request the web server sends the requested web page that web page is considered as predicted page to the cache. If the prediction is right and send a request therefore the page is in the cache after that calculate hit rate for the next new session of request then the cache will again send the request to the server.

Decision Tree, Apriori algorithm [Pooja Sharma][1]

Apriori algorithm is based on breadth first search to count the candidate data set. Generates very large candidate data set that affects the execution speed to process data sets. **Eclat algorithm** each row belongs to data and each transaction belongs to column. It mechanism is depth first. Two ways bit matrix performance one bit for each data and transaction and other bit for each row a record of columns in which bit is set. Generate candidate set by using two bits. The number of generated data set is much larger than data set generated in Apriori algorithm..

Fuzzy C Means and Expected Maximization algorithm [K.Poongothai][1]

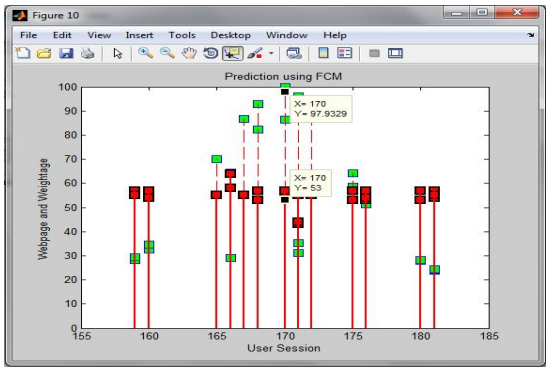
Fuzzy clustering algorithm is optimally separate similar user interests. Compared with hierarchical patterns (to discover patterns) and several function approximation techniques. Fuzzy C means clustering algorithm (to discover web data clusters) and compare with Expected Maximization cluster system to analyze the Web site visitor trends. Expectation maximization (EM) is used for clustering in the context of mixture models. This method calculates absent parameters of probabilistic models. The parameters are reused until a preferred value is achieved. The finite mixtures model assumes all attributes to be independent random variables

FP-Growth Algorithm & Apriori

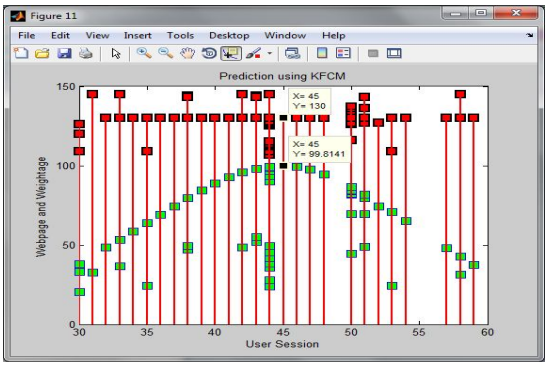
[Mr. Rahul Mishra][1]

The FP-growth algorithm is frequent items set mining uses the FP-tree structure to attain a divide-and conquer to break down the mining problem into a set of smaller problems. Performance of FP-growth compared with apriori in large databases. Frequent patterns from web log data using FP for finding the most frequently access pattern generated. Apriori for association rule mining. Finding frequent sets using candidate set generation. Apriori sets many candidates so the cost is high compared to FP and its getting slow. FP requires less memory, low cost, very fast than Apriori.

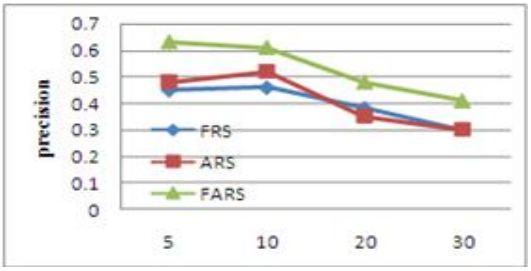
4.Results

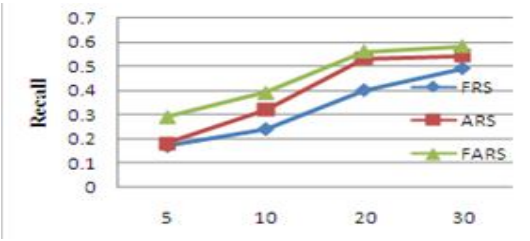


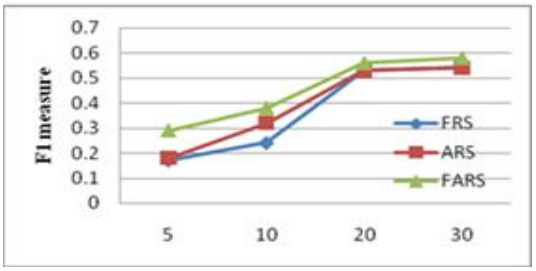
Prediction of user future web page using FCM



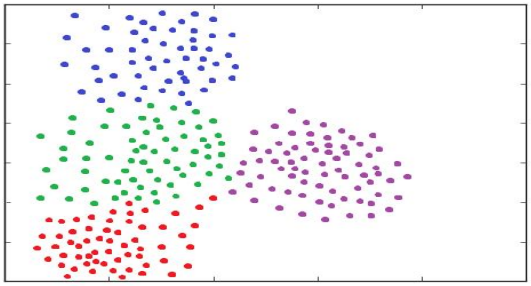
Prediction of User future web page using KFCM



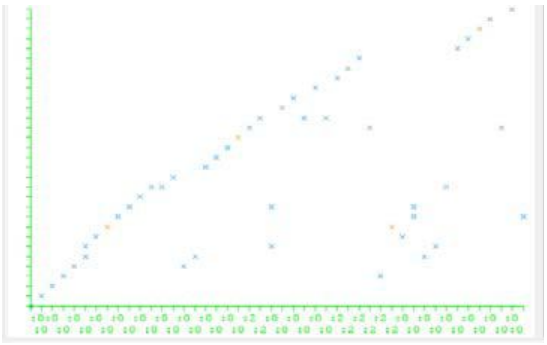




Density connected cluster using DBSCAN



Predict user Future Session using k Means



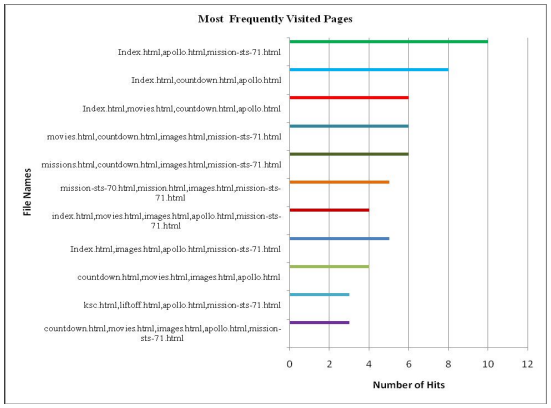
Customer Behavior Analysis using Apriori

FREQUENT PATH REPORT

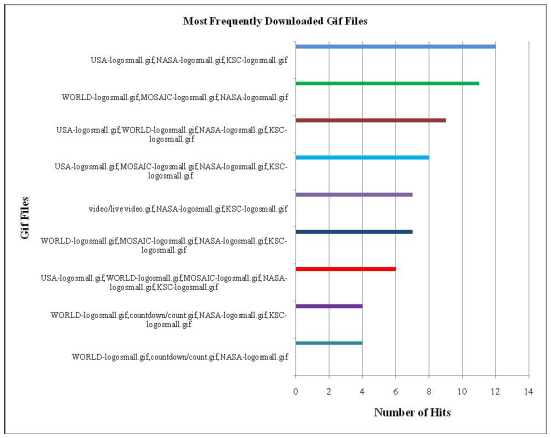
/A.JSP:/B.JSP:/C.JSP://D.JSP:/F.JSP/G.JSP

FREQUENT BINARY PATH

/A.JSP→C.JSP
/B.JSP→C.JSP
/B.JSP→D.JSP



Most frequently visited pages using FP



Most frequently downloaded pages using FP

5. Summary of Literature review

Author	Method	Application	Year
Shaily G. Langhnoja	DBSCAN	Density-connected cluster	2013
Dilpreet Kaur	KFCM	Robust than FCM, highest probability	2013
M.Rathamani	Fuzzy c means	Cost reducing, Security	2012
Author	**Method**	**Application**	**Year**
Neeraj Raheja	Web log searching	Log partition.	2014
Akshay Kansara	k means	Accuracy Coverage	2013
Preeti Sharma	Apriori	Frequent navigation pages retrieved	2011
Monika Verma	Dynamic CandidateGeneration	Execution time less	2012
Mr. Rahul Mishra	FP-Growth Algorithm & Apriori	Frequent patterns	2012

6. Conclusion

In this paper review the methodologies in the area of web usage mining focusing on user behavior analysis and future request prediction. Methods for navigation pattern discovery and future prediction analysis are discussed by reviewing different research papers. These methods can be applied on different websites to evaluate the performance and effectiveness.

7.References

[1].S.K. Pani, et al., "A Survey on Pattern Extraction from Web Logs" IJICA, Volume 1, Issue 1, 2011.

[2].Ujwala Patil and Sachin Pardeshi, "A Survey on User Future Request Prediction: Web Usage Mining" ISSN 2250-2459,

Volume 2, Issue 3, 2012

[3] V.Chitraa,"A Survey on Preprocessing Methods for Web Usage Data" (IJCSIS) Vol. 7, No. 3, 2010

[4] R.Dhanya," An Efficient Web Learning Web Text Feature Extraction with Exponential Particle Swarm Optimization" International Journal of Advances in Computer Science and Technology, Volume 3, No.2, February 2014

[5] D. Uma Maheswari, "A Study of Web Usage Mining Applications and its Future Trends, International Journal of Engineering Research & Technology, Vol. 2 Issue 9, September – 2013