Efficient Mining of Interest Patterns on Click Stream Data

P. Dhana Lakshmi

Abstract*: Nowadays, large amount of data is generated daily in e-commerce applications as click stream data. Because of the availability of this tremendous amount of data analyzing the user browsing behaviour and finding frequent navigation patterns of different web pages accessed by web users is an key element for retailers to optimize the website and personalized the web services of different e-commerce websites. User browsing behaviour is evaluated based on user interests on web pages or products. There are different parameters are considered while analyzing the click stream data for calculating frequent navigation patterns and context based customer behaviour in online data bases. In this paper we developed different models for optimizing and personalizing web service and sequential frequent patterns using the parameters: browsing path, frequently visited web pages, time duration of web pages and user interest. These novel models uses the parameters and applied on click stream data to optimize the web pages and improve the personalized recommendation.

Keywords: Data stream, FP-Growth Algorithm, CURE Clustering, Frequent patterns.

I. INTRODUCTION

Due to the explosive growth of information available on the internet, extracting useful information is an important research area. Mining the information on the web may be a content, structure or usage. Web usage mining is a best useful data mining technique to data stream for extracting web usage patterns[11]. It is also for website design, personalized recommendation, to create an adaptive website. Generally when the customer(s) or user interacts with any web application[13], all the user clicked information are stored in a web server as logs are called as click stream data. This click stream data is used for analyzing the customer behaviour and find the insights of customer for different online applications. There are different applications of click stream data: Traffic Analysis, Time Series Analysis, web usage and text analysis. Data streams are also useful to analyze the streaming data in knowledge systems and for monitoring the possible disease from sensors warns of patients and recognizing enemy equipment in army. Predicting user future and classifying user navigation patterns based on content of web pages retrieved for target user and their log data collected from web server. User future requests are predicted by constructing user navigation profiles. Building navigation profiles depend on both web server log files and extraction of web page content.

Web usage navigation patterns classification may be done in two perspectives:

1. User perspective
2. Website Perspective

Generally identification of navigation patterns from the user perspective is used for predicting the most likely accessible web pages for the current users and to improve the personalized web recommendations. This identification used for suggesting related web pages which are unseen by the current user. Classification of navigation patterns in the website perspective used for web masters for organize the website in such a way that user can easily navigate the website based on their needs[12]. Before optimizing the website structure by web masters they find and remove the unnecessary links between web pages. This is achieved by regularly observing the user navigations and their interest and update accordingly.

In order to achieve all these there is an efficient and most widely used technique in web mining. In web mining to extract interesting and frequently navigated user patterns, web usage mining is useful similarly to mine the knowledge from the contents of web pages web content mining is used. Combining of web usage with web content mining gives effective classification of navigation patterns and accurate web user future requests.

This paper is organized as follows section2 discusses about related work section 3 introduces different models used for finding context based behaviour and finding frequent patterns section4 discuss about experimental results section5 is about conclusion.

II. RELATED WORK

Due to rapid rate of continuous, unbounded occurrence of streaming data extracting frequently accessed usage patterns[1] in less time is a critical issue. Click streaming is an one of the emerging research area in web analytics. Web masters utilizes this data to identify the user navigation patterns and for finding potential users and for optimizing the website.[2] implemented a different models for browsing behaviour of all users for any website using click stream data to resolve this both in online and offline. Before processing data is stored as static in offline mode where as in online when a transaction occurs maintaining the data is in different data structures.

Revised Manuscript Received on December 15, 2019.
***** Correspondence Author
Dr. P. Dhana Lakshmi *, Associate Professor ,Dept. of CSSE, Sree Vidyanikethan Engineering College, A.Rangampet, Chittoor (District), Andhra Pradesh, India E-mail; mallidhana5@gmail.com

To retrieve frequent patterns over data streams there are many algorithms are developed for mining streaming data. Apriori and FP-Growth are most widely used techniques[3] for finding frequent patterns.[6] also uses browsing sequences and FP-Growth algorithm for extracting frequent patterns and preprocessing the log records. An improved FP-Growth algorithm for streaming data is developed by [4] which uses multiple granularities for finding frequent patterns under window framework. Generally streaming data changes as the time changes to find recent updated frequent patterns.[5] presented an algorithm for online streaming data in terms of tree called DS tree and monitoring tree. Generally all the data items are represented as prefix tree[7] for storing critical information about frequent items. Over the past decade most of the research is about finding frequent patterns on data streams using count, landmark and sliding window models[8,9]. These techniques are implemented using fixed window size and streamed data is processed and stored in the form of batches. Due to the popularity of personalized recommendations [10] develops a method for finding recent frequent patterns instead of approximate patterns.

III. FREQUENT PATTERN AND WEB LOG STATE MODELS

Generally, most of the e-commerce website applications will add more number of products into a website in order to attract more customers and its productivity in terms of sales. For example in e-bay, Amazon.com, snapdeal.com applications all products are arranged category wise in a tree structure.

When a user visits a web site, extraction of the user browsing behaviour is useful for finding frequent patterns and also calculate the number of times a page is visited by the user in a single or multiple sessions, navigation path, total duration of web page visited by the user. A sequence of web pages visited by the user for a particular time period is known as session. If the user enters into a website, it automatically creates a session ID for each visited web page by the user. It is terminated automatically if the user not requested any web page for some time.

Consider a visiting path sequence is denoted ass $P_i = \{ wp_1, wp_2, wp_3 \text{---------} wp_n \}$

Wp_i is a web page or url of any website

P_i represents browsing path for an user i

C_i is a category

I_i is a item i

For example browsing path for uses1 is $P1=\{ C1, i_1, C_1, i_2, i_1 \}$

IV. METHODOLOGY:

The proposed method consists of following steps.

1. For the given dataset or given web log data perform weblog preprocessing.
2. Apply CURE Clustering technique to identify optimal number of web users.
3. Identify the navigation patterns and build the user pattern profiles.
4. Find the user frequent navigation patterns and predict the user future requests.
5.

The below figure1 shows the process of finding frequent patterns.

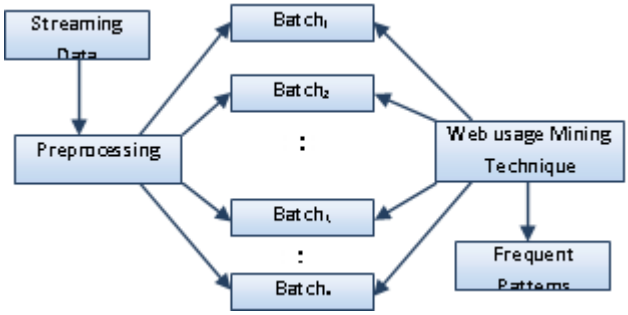


Fig. 1. Frequent Pattern Mining Model on Click Stream Data.

Generally if any user communicates with any website all the communication details are stored in the form of logs for any web server. This log file contains basic information such as IP Address, URL and number of bytes accessed.

Example: <ip_addr>--<date><method><file><protocol> <code> <bytes> <referrer><user_agent>

203.23.6.128 - - [21/sep/1999:03:09:21 -0600] "GET /xyz.html HTTP/1.1" 200 5789 "http://www.lycos.com/cgi-bin/pursuit?query=advertising+psychology-&maxhits=20 &cat=dir" "Mozilla/4.5 [en] (Win98; I)"

All the fields in web log data may not be useful for retrieving useful information so unwanted fields for finding frequent patterns is removed from web log data In preprocessing stage at the same a record with status code other than 200 is also removed. After filtering the useful fields apply CURE Algorithm which will work for large databases. Actually CURE functioning as similar to Hierarchical clustering. This clustering efficiently work on distributed stream data and eliminates more outliers. The resultant of this clustering gives optimum number of clustered users. The resultant optimum clusters considered as similar navigation behaviour of each category of web site.

Logic based Web State Model:

It is an another model to optimize the website as per user requirements and it is also used for finding frequently accessed patterns. In this model each web page is considered as a state, Navigation of one of web page to another web page is considered as transition and the time spent on a web page is considered as duration and number of times accessed an web page is assumed as frequency. Based on this transition matrix frequent navigation patterns are calculated using Logic based theory.

The following procedure is used to implement the Logic based web state model after collecting the data form web server log file.

1. Perform web log file preprocessing
2. Identify the similar web pages accessed by the web users and classify them based on similar services provided by the site.
3. Calculate the frequency between web pages and model transition matrix for an each website based on the following steps

i. For a single session calculate the transition frequency for the same user.

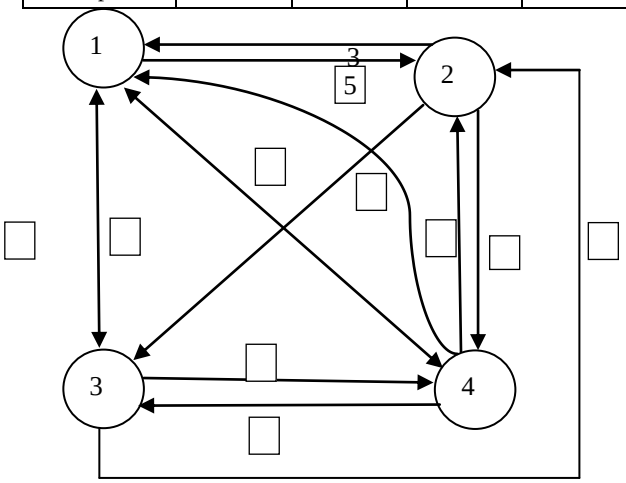
ii. Suppose if the user accessed a two consecutive web pages duration is more than 30 min it will be considered as user is available in two sessions.

4. Construct the web page state model using step3 for all web pages accessed by a web user for a single website or multiple websites.
5. Find the Logic Based pattern discovery for all user accessed web pages in different sessions.

For example, if the transition frequency between web pages is shown in TableI and Fig. 2.

Table I: Web Page Transition frequency

Web pages	Wp1	Wp2	Wp3	Wp4
Wp1	0	5	4	2
Wp2	3	0	2	5
Wp3	4	3	0	2
Wp4	3	4	5	0



State Transition Diagram:

To characterize the user behaviour of any navigation the following mathematical model is used. Consider a model M with number of web pages Pij is a probability, Navigation from web page i to webpage j is denoted by Wi and Wj and S is a set of sessions then predictable model has the following transition matrix.

$$P(Q_{n+1}=W_i/Q_0=W_0, Q_1=W_1,\text{----------} Q_n=W_n)= P(Q_{n+1}= W_j/ Q_n=W_i)$$

$$=P_{i,j}$$

$$P_{i,j} \geq 0 \text{ for all } i,j \in \{1,2,\text{-----}n\}$$

$$\sum_{j=1}^{n} P_{i,j} = 1 \text{ for all } i,j \in \{1,2,\text{-----}n\}$$

The predictable model is constructed based on finding frequently accessed web pages and rarely used web pages by using navigation between web pages.

Assume webpage 4 i.e W4 Navigation paths for this web page is calculated as examine all the pages reaches to webpage 4. The pages are 1, 2, 3. Similarly calculate the navigations of these web pages to web page 4 to access any service. The navigation paths are w2-w4, w2-w1-w3-w4, w2-w1-w3-w4, w1-w2-w3-w4.

From Table1 we calculated the transition frequency between web pages. To find the Frequent patterns apply Logic based model (L). There are two considerations while evaluating the frequent patterns one is positive and another one is negative. If there is any transition between web pages and satisfies minimum session threshold and minimum support then that is considered for frequent. If no such web pages are accessed that information is also required by the manager to remove those rare web pages for optimizing the website and personalized recommendation to users. This is similar to Equivalence relation in relation algebra.

Example:

$$L = \begin{matrix} T & T & T \\ T & F & F \\ T & T & F \\ F & F & T \end{matrix} \quad F$$

Here T represents web page is accessed and F represents web page is not visited

V. EXPERIMENTAL RESULTS

For our experimentation we extracted access entries of 2 months dataset from farmersarmy.com and click stream data collected from one of largest resource in france. This site consists of electronic devices, clothes, toys and cosmetics. All these records are organized into different sessions. This data set contain s 256 records for September 2019. Each record contains URL, through this record, for each user HTML file is retrieved because of click stream data is very large. Because of it is very rich data source and if it has different formats of data analyzation is very difficult. So therefore all these data is preprocessed before it is divided into segments. This preprocessing includes data cleaning, user identification, session identification. Sequence of web pages accessed for a particular time period is called as session. Users are identified in two ways one is based on IP Address and second one is based on session duration.

In france online Shopping market website contains 5 percent of customer users visit the website for every week. The details of collected click stream data before and after preprocessing is shown in Table2.

Table II: Click stream data

Name	Original Data	Preprocessed data
No. of Categories	3800	2320
No. of Records	1000500	800480
No. of sessions	785200	1540

The Proposed approach effectiveness is validated with CURE clustering method.

Interest patterns are evaluated for click stream data from the website.

Table III: Comparison of clustering Algorithms

Algorithm	K	Time(s)	Similarity
CURE	40	5	0.15
K-Means	38	28	4.5
K-Means	30	20	2.8
K-medoids	21	15	1.9

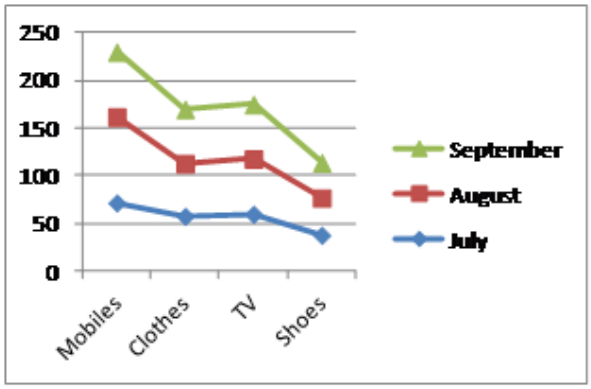


Fig. 2 Interest Patterns for different Products

Farmers army is a website which gives information regarding the new techniques and methods in farming which are easily implemented in fields. It is a platform where they encourage farmers by writing articles about them and one can get their organic products by sending email.

It is useful for the farmers, scientists, students and marketers. Farmers can know the information regarding the methods and the people who implemented these methods. The students and scientists can know the information regarding techniques in which they can improve in better way. The marketers can buy their products in large amount.

There are four posts in this web site.

*TRENDING METHODS IN FARMING#POLYHOUSE

*FARMERS SUPPORT SCHEME

*Mahatma Gandhi Village Independence Celebrations-GOVANAM TRUST

*ORGANIC MULCHING SHEETS

The author posted three posts where there are 134 views in three weeks. There are 94 views during the week of September 23. There are 14 views during the week of September 30. There are 26views during the week of October 7.

It started recently in September 2019.The best views in this website is 28 views in one day. The more clicks are on home page and people page.

The Figure3 and Figure4 shows day wise and monthly wise statistics of farmers army website

Fig. 3. Day wise statistics

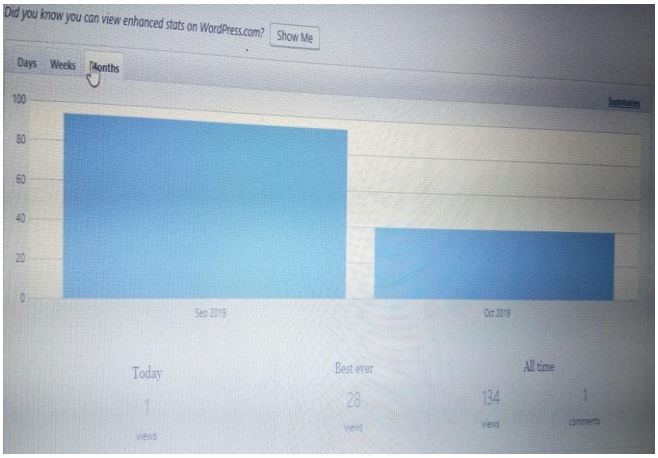


Fig. 4. Monthly statistics

Fig. 5. Total and best views of Farmers army.com

The below figure(s) shows the traffic driven to each and every post s and the information regarding the posts published, date of publication and number of views on the date of publication.

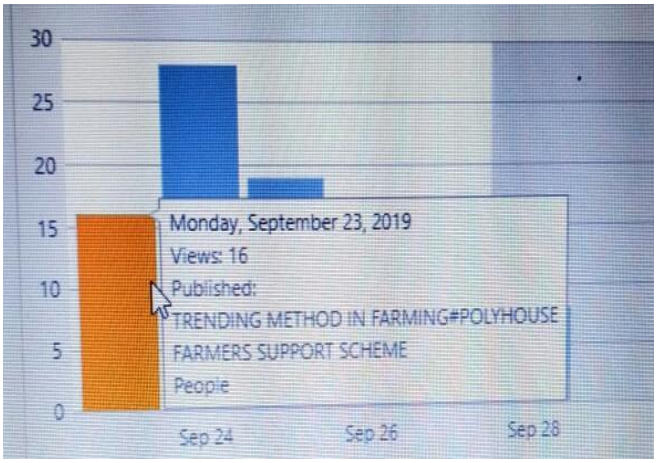


Fig. 6. STASTICS OF VIEWS ON SEP 23

Figure6 Shows the posts called TRENDING METHOD IN FARMING#POLYHOUSE AND FARMERS SUPPORT SCHEME posted on September 23 and 16 people have visited this website.

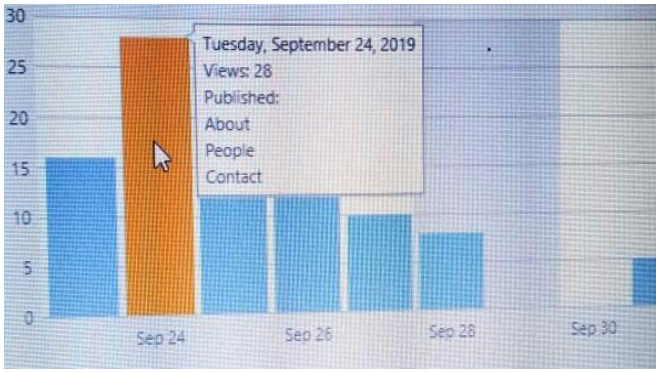


Fig. 7. STASTICS OF VIEWS ON SEP 23

The Fig.27 shows the post called ORGANIC MULCHING SHEETS published on September 24 and 24 people have visited this website.

VI. CONCLUSION

Nowadays due to increased number of web applications in the internet, analyzing the web usage data to attract more customers and to predict the future interest is an popular research in web usage mining. Real time data analysis is more valuable than historical data for creating an adaptive websites, Recommend the products to the user in optimized way. In sensor network analysis, stock data analysis applications due to increase of data streams mining frequent patterns is also an challenging issue. In this paper an effective algorithm is CURE is applied on click stream data to extract user interest in terms of frequent patterns in a short period of time. Therefore we developed different models for optimizing and personalizing web service and sequential frequent patterns using the parameters: browsing path, frequently visited web pages, time duration of web pages and user interest. These novel models uses the parameters and applied on click stream data to optimize the web pages and improve the personalized recommendation.

ACKNOWLEDGMENT

This work is supported by University Grants Commission (UGC) under Minor Research Project titled “Development of Mathematical Model for the Prediction of Customer Behavior in Online Databases”.

REFERENCES

1. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J. (2002). Models and issues in data stream systems. In *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems* (pp. 1–16). ACM Press.
2. Giannella, C., Han, J., Pei, J., Yan, X., Yu, P.S. (2003). Mining frequent patterns in data streams at multiple time granularities. *Next generations on data mining* (pp. 191–212).
3. Han, J., Pei, J., Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of the 2000ACMSIGMOD international conference of management of data* (pp. 1–12).ACM Press.
4. C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu, Mining frequent patterns in data streams at multiple time granularities, in *Data Mining: Next Generation Challenges and Future Directions*, (AAAI/MIT Press, 2004), pp. 191–212.
5. J. H. Chang and W. S. Lee, Finding recent frequent itemsets adaptively over online data streams, in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, 2003), pp. 487–492.
6. Huiping Peng “Discovery of Interesting Association Rules Based on Web Usage Mining” 2010 International Conference.
7. Han J., Pei J., Yin Y. and Mao R., “Mining frequent patterns without candidate generation: A frequent-pattern tree approach” Data Mining and Knowledge Discovery, 2004.
8. Lee, .J and C.S. WaNG 2007. An efficient algorithm for mining frequent inter-transaction patterns Inf. Sci 177, pp. 3453-3476.
9. Li, J, D.Maier, K.Tufte, V.Papadimos and P.A Tucker, 2005. No Pane no gain: Efficient evaluation of sliding window aggregation over data streams. ACM. SIGMOD. Rec, 34:39-44.
10. Silvestri, C and S.Orlando, 2007. Approximate mining of frequent patterns on streams. Intell Data Anal, 11: pp: 49-73.
11. Gnabasambandan P, Poonkuzhali S, “Click stream Analysis on web usage mining”, International Journal of Pure and Applied Mathematics, Vol 119 No.16 2018,PP. 891-899.
12. Qiang Su, Lu Chen, “A method for discovering clusters of e-commerce interest patterns using click- stream data”, 2014, PP. 1-11.
13. Quanshu Zhou, Hairong Ye, Zuohua Ding, “Performance Analysis of Web Applications based on user Navigation”, 2012 International Conference on Applied Physics and Industrial Engineering,2012, pp.1319-1328.