An Efficient Algorithm To Generate Dynamic User Pattern Cluster Using K-means Clustering Techniques

Ms. Rashmi Patel
H.G.C.E., Vahelal

Mrs. Risha Tiwari
Asst. Prof., Computer Department,
H.G.C.E., Vahelal

Abstract - The expanded on-line applications are prompting to exponential development of the web content. The vast majority of the business associations are intrigued to know the web client conduct to improve their business. In this unique circumstance, clients route in static and element web applications assumes a vital part in comprehension client's interests. The static mining procedures may not be appropriate as it is for element web log documents and basic leadership. Conventional web log preprocessing approaches and weblog use designs have confinements to break down the substance association with the perusing history This thing, concentrates on different static web log preprocessing and mining strategies and their material confinements for element web mining using this techniques we can create pattern cluster so we can easily retrieve data from data source.

Index Terms—: Data Mining, Web log mining, Web Mining, Clustering, Pattern Clustering.

I. INTRODUCTION

Web mining is a process to analyze the online Web data, navigate between various Web sites and perform transaction of data across the Web. According to the types of data can be mined, web mining is classified into three types. Web Content Mining discovers information or knowledge from millions of sources across the Web. Web structure mining is the technique of finding structure information from the web. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web based applications.

II. TYPES OF WEB LOG FILE FORMAT

A. W3C(World Wide Web Consortium) Extended Log file Format Extended log is a customizable ASCII format which has different types of fields.
B. Microsoft IIS(Internet Information Services) Log file Format can record more information than the NCSA format.
C. NCSA(National Centre for Supercomputing Application) Ordinary Log file Format which is available for Web sites but not for FTP sites.

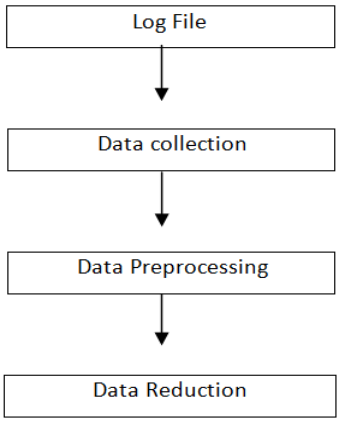

Figure.1 Existing Architecture

III. EXISTING ALGORITHM

1. Read **N** no of records from clean data source **DS**
   For i=1 to i<=N
2. For each records **R** find frequent access data item **F** from data source **DS**
3. Read frequency user access item **F**
4. If **R = F** frequent records then
5. Save for clustering frequent user access records in frequency access data source **FDS**
6. Make cluster from frequency user access records
7. Else not select records
8. End If
9. Next record

IV. DRAWBACKS OF EXISTING

1. It does not provide clustering
2. Does not Cache of visited item
3. It recommended all visited item
4. Doesn't create pattern clustering
5. It gives less performance
6. It consumes time
7. Low efficiency and less accuracy

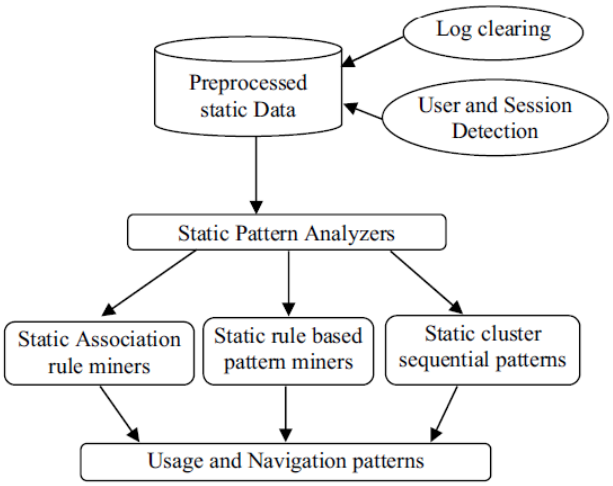

Figure 2 - Static Pattern Analyzers

V. PROBLEM STATEMENT

The mining is most useful technology and method to gain information from dataset. A general static model for logs cleaning, user & session identification and pattern discovery models are studied on the different log files. This allows guaranteeing an importance of static pattern discovery models on static web log files for decision making. Also, static models only utilize temporal logs for pattern evaluation and navigation, which affects the precision and recall. Experimental results show that the static models on static log data have high performance in terms of F-measure and error rate on the limited data size and limited fields. In future work, we will develop a mathematical model to predict the user's behavior on dynamic web contents based on server log and customer database.

VI. PROPOSED ALGORITHM

1. Read no of all records from data source
2. If the same records then put flag on record from data source
3. For each records from data source find frequent data
4. Read frequent pattern data using specified address from data source
5. If requested records from data source with specified pattern then
6. Collect and Save in pattern data source
7. Make cluster in pattern data source
8. Else not select that record
9. End if
10. Next record

VII. CONCLUSION AND FUTURE WORK

After conclude the pre-processing. The cleaned data is stored in temporary conversely, the raw data before analyze is about 13,77,738 records. Data filtering perform by removing unwanted patterns from each record in the database. Since the pre-processing techniques performed is to mine the interesting patterns, the data end with *.jpg, *.gif, *.bmp be removed. The final data after all process completed is about 38,890 records. The final data will be fed into Generalized Association Rules for rule generation and calculating the interesting rules by producing the support and confidence value.

And in future work we have to create dynamic pattern cluster for dataset. Top ten exit pages and one more future work is like this log files are not in format. So we should format all this log files so we can easily make customizable combined file for analysis.

REFERENCES

[1]. P.Dhanalakshmi, Dr. K.Ramani, Dr.B.EswaraReddy,"The Research of Preprocessing and Pattern Discovery Techniques on Web Log files" -978-1-4673-8286-1/16 $31.00, 2016 IEEE DOI 10.1109/IACC.2016.35

[2]. ShashiMehrotra, ShrutiKohli, "Comparative Analysis of K-Means with other Clustering Algorithms to Improve Search Result" - 978-1-4673-7910-6/15/$31.00, 2015 IEEE

[3]. Dilip Singh Sisodia, ShrishVerma, "Web Usage Pattern Analysis Through Web Logs: A Review" - 978-1-4673-1921-8/12/$31.00, 2012 IEEE

[4]. TingZhong Wang, D. Jin and S. Lin (Eds.), "The Development of Web Log Mining Based on ImproveK-Means Clustering Analysis" - Advances in CSIE, Vol. 2, AISC 169, pp. 613–618. Springer-Verlag Berlin Heidelberg 2012

[5] S.C. Satapathy et al. (Eds.), "Design and Implementation of an Effective Web Server Log Preprocessing System" - Proceedings of the InConINDIA 2012, AISC 132, pp. 897–905. Springer-Verlag Berlin Heidelberg 2012

[6]. J.MonishaPrivthyJeba,M.S.Bhuvaneswari, K.Muneeswaran, "Extracting Usage Patterns from Web Server" - 978-1-4673-6615-1/16/$31.00 © 2016 IEEE

[7]. Supinder Singh, SukhpreetKaur, "Web Log File Data Clustering Using K-Means and Decision Tree" - 2013, IJARCSSE All Rights Reserved

[8]. Ripal Patel, Mr. KrunalPanchal, Mr. DushyantsinhRathod, "Efficient Log Mining from Web Server Using Clustering Technique" - IJETIR1512016 - 2015

[9]. Mr. Dushyant B. Rathod, "Customizable Web Log Mining from Web Server Log" - IJEDR1302021

[10]. http://www.w3.org/TR/WD-logfile.html

[11]. http://www.extratrend.com