Web Usage Mining Based On ACAKHM Algorithm

Priyanka k Prajapati[1] ,Mr.jay B Amin[2]

[1]Student,Department of Computer Engineering,L. J. I. E. T, Ahmedabad, Gujarat, India
[2]Assistant Professor,Department of Computer Engineering,L. J. I. E. T, Ahmedabad, Gujarat, India

Abstract

The web usage mining uses data mining techniques to discover interesting usage patterns from web data. Web personalization uses web usage mining techniques for the process of customization. Customization involves knowledge acquisition done by analysis of user's navigational behavior. A user when goes online would like to get the links which suits his requirements or usage in the website he visits. Clustering is a popular technique of data mining for unsupervised learning in which labels are not defined previously. K-Mean is a well known partitioning technique for forming different clusters, but it has the drawback of initial sensitivity and local optima convergence. K-Harmonic algorithm solves the initial sensitivity problem, but it stuck in local optima problem. The Ant Clustering Algorithm (ACA) can avoid trapping in local optima solution. In this paper, we will propose a new clustering algorithm using Ant Clustering Algorithm with K-Harmonic mean clustering (ACAKHM).

Keywords- *Web Mining, Web Usage Mining, Log Files, ACAKHM, Pattern Analysis*

I.Introduction

The Rapid growth of the internet has made the www a popular place for collecting information. A web mining has important task to discover useful knowledge or information from the web. Web mining can be divided in to three categories: web structure mining,web usage mining and web content mining. Web structure mining is the process of discovering hyperlink and document structure information from the web. Web usage mining is the application of data mining techniques for finding interesting and useful usage patterns from web data which makes it more demanding for web based applications. Web content mining is the process of extracting useful information from the contents of web documents[1].

Web usage mining is also called web log mining. Web Usage Mining (WUM) is the approach to extract the knowledge from analysis of web usage data about a particular website. This usage data can be obtained from server logs and can analyse the behavioral patterns and profiles those interact with the web sites[2].
The web usage mining process involves three main steps:

1) Preprocessing
2) Pattern Discovery
3) Pattern Analysis

In data mining, a method often used is clustering. Object clustering is done based on its characteristics. Companies can use clustering methods to identify patterns of data so that companies can found a certain pattern of the data. Clustering is one of the important data mining technique to discover usage pattern.

Preprocessing : Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. [2]

The different types of preprocessing in Web Usage Mining are:

Pattern Discovery: Web Usage mining can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. The following are the pattern discovery methods. [2]

1. Statistical Analysis
2. Association Rules
3. Clustering
4. Classification
5. Sequential Patterns
6. Dependency Modeling

Pattern Analysis : This is the final step in the Web Usage Mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL (Structured Query Language) processing and OLAP (Online Analytical Processing) can be used. [2]

In this paper we will propose a new clustering algorithm using Ant Clustering Algorithm with K-Harmonic mean clustering (ACAKHM). The remaining of this work in organized as follows: first describe related studies in section 2.Then section 3 describe the proposed architecture for extracting the main content from the web pages. The result of our approach describes in section 4 and finally, we describes conclusion and future work in section 5.

II EXISTING SYSTEM

In the paper, "An Optimized k-Harmonic Mean Based Clustering User Navigation Patterns" [1] by R. Gobinath, M. Hemalatha. In this, In this paper they deals with the extraction of necessary information from web access log files and applying clustering for easy analyzing of navigational patterns for web personalization. The k-harmonic algorithm is used for clustering the obtained navigational patterns from the various iterating process.

The web mining is an application process which plays an important role in analyzing the behavior of the website users. The web usage mining is a sub category of web mining has a major impact on web personalization. The main concept of the paper deals with the extraction of necessary information from web access log files and applying clustering for easy analyzing of navigational patterns for web personalization. The adapted k-harmonic algorithm is used for clustering the obtained navigational patterns from the various iterating process.

The proposed framework is based on the following process.

1) Data collection

2) Pre-processing

3) Feature extraction

4) Pattern Discovery

5) Pattern analysis

The methodology involved in this paper is shown in the following architecture.

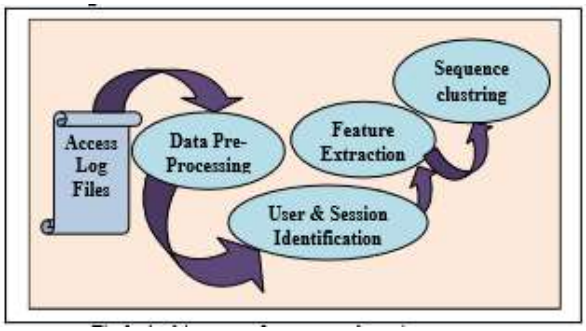

Figure 1: Architecture of sequence clustering process [2]

2.1 Comparison of Different Approaches

Table -1: Comparison of Different Approaches

Research Paper	**Methodology/ Algorithm**	**Pros.**	**Cons.**
The integrating between web usage mining and data mining techniques	Data mining techniques	Improve Performance, and Frequent pattern	.
An optimized k-harmonic mean based clustering user navigation patterns	K-Harmonic mean algorithm	K-Harmonic algorithm solves the initial sensitivity problem	Stuck in local optima problem
Cluster optimization for improved web usage mining using ant nestmate approach	ART1-neural network, AntNestmate approach	Data Redundancies occur	Time complexity is more
An efficient hybrid data clustering method based on candidate group search and genetic algorithm	candidate group search and genetic algorithm (CGSGA)	Reach to global optimal solution	Computational time is more
An efficient prediction based	Modified ant	Improved next	Grid scheduling

on web user simulation approach using modified ant optimization model and hierarchical clustering	optimization	node election	problems

III. PROPOSED APPROACH

Problem Statement: As in the current framework, we have seen that the data extraction done has not achieved the exactness with the expanding request. Furthermore the productivity in the current framework is less. In the current framework, they have not extricated the words from the sentences and passages, so it makes the framework complex.

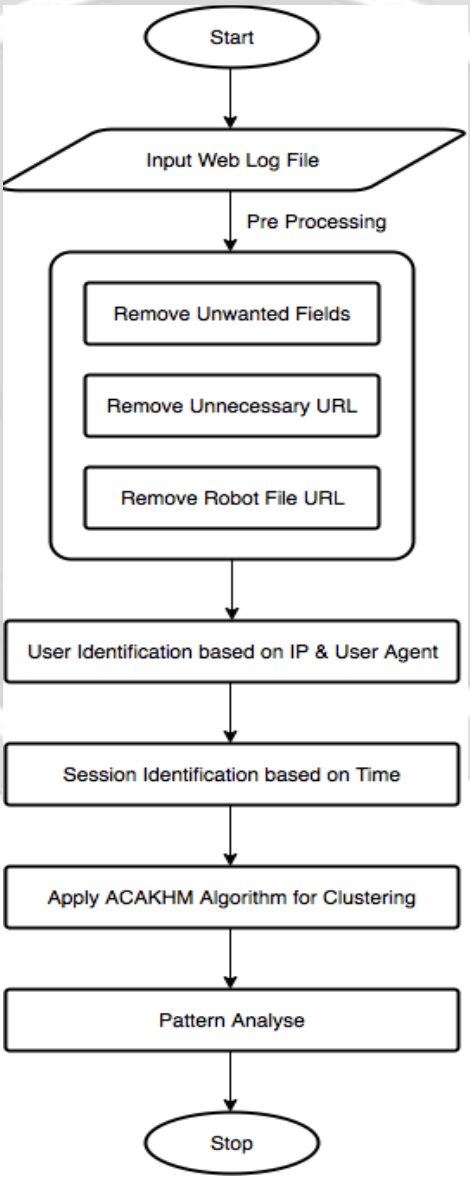

Figure 2: Proposed Work Flow

Proposed Solution steps describe as below:

1. First step is start the process and the consider web log files as an input.
2. The second step is preprocessing. In this separates all fields and remove all unwanted fields.
   a. Remove Unwanted field
   b. Remove Unnecessary URL
   c. Remove Robot file URL
3. After the preprocessing we get cleaned log files.
4. User identification based on IP and user agent
5. Session identification based on Time
6. Then Apply ACAKHM algorithm for clustering
7. Ruled based pattern analysis is apply for pattern finding
8. Finally extract the meaningful pattern from web log files.

IV. EXPERIMENTED RESULTS

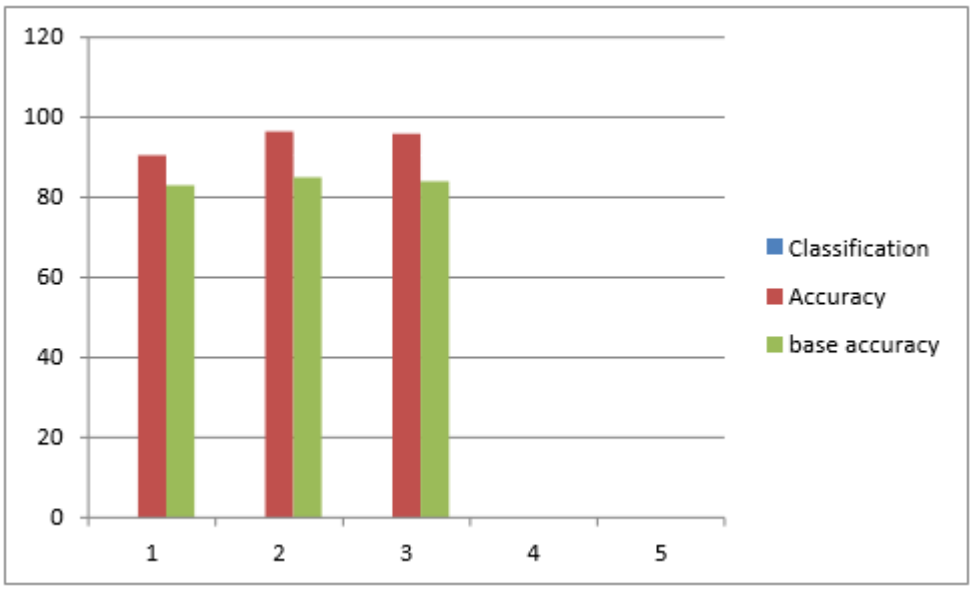

Figure 3: Accuracy comparison of base paper and proposed system

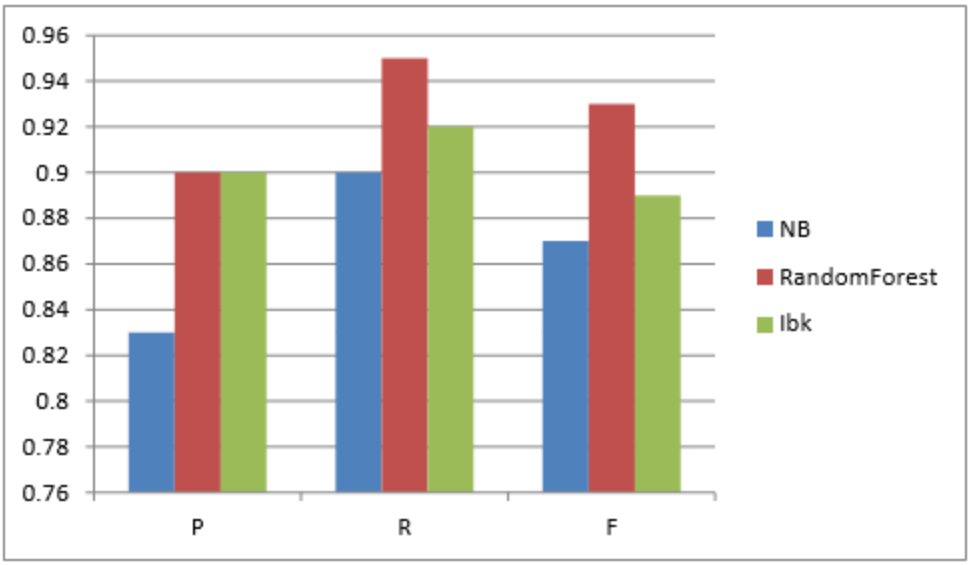

Figure 4 :Precision, Recall, F1 measure for Proposed system

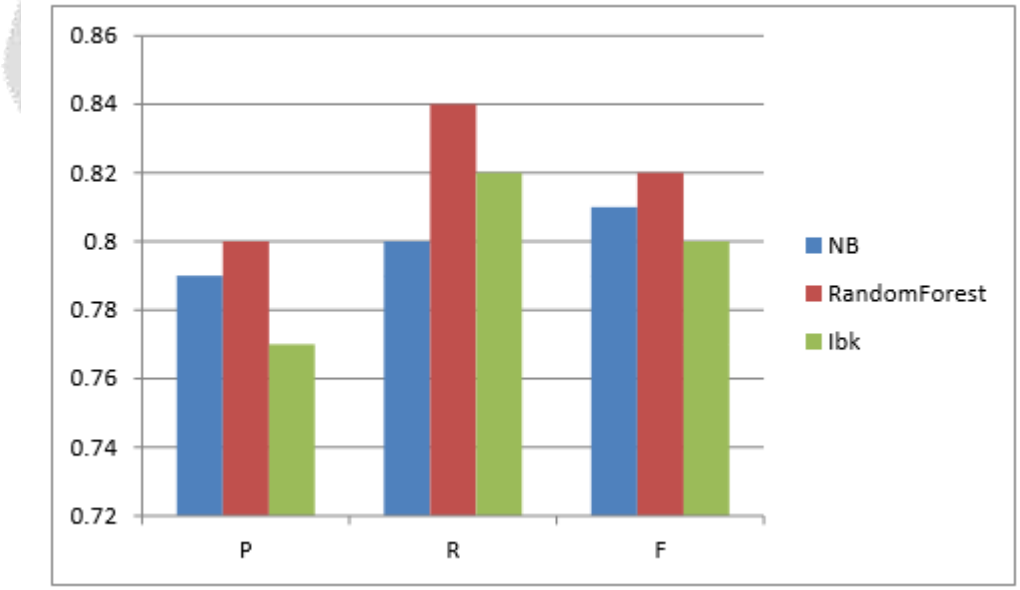

Figure 5: Precision, Recall, F1 measure for Base system

Conclusion

This paper deals with a cluster optimization technique. The web log is accessed and performs data cleaning. The cleaned web log is used for pattern analysis. This paper uses the clustering technique for discovering interesting usage patterns. Clustering is done based on user identification and session identification. In this paper we present a new algorithm using the Ant clustering algorithm with K-haromonic means clustering (ACAKHM).It overcomes initialization sensitivity of KM and KHM, and reaches a global optimal effectively. The result of ACAKHM algorithm is better than the KHM algorithm. The accurateness of the results of proposed method is improved over the existing methods.

References

1] R. Gobinath, M. Hemalatha "An Optimized k-Harmonic Mean Based Clustering User Navigation Patterns " International Conference on Computational Intelligence and Computing Research, 2013. 978-1-4799-1597-2/13/ 2013 IEEE

[2] Omer Adel Nassar , Dr. Nedhal A. Al Saiyd "The Integrating Between Web Usage Mining and Data Mining Techniques." International Conference on Computer Science and Information Technology (CSIT) *2014 International Conference on*, pp. 978-1-4673-5825-2013 IEEE

[3] Anna Alphy , S. Prabakaran " Cluster Optimization for Improved Web Usage Mining using Ant Nestmate Approach" International Conference on Recent Trends in Information Technology *Fourth International Conference on*, pp. 978-1-4577-0590-, 2011

[4] Suvarna P. Patil, Anuradha D. Thakare, C. A. Dhote. "An efficient hybrid data clustering method based on Candidate Group Search and Genetic Algorithm.". 978-1-4673-7231-2/15 IEEE, 2015.

[5] Trapti Agrawal, Shailendra Srivastava, Abhishek Mathur "An Efficient Prediction Based on Web User Simulation Approach Using Modified Ant Optimization Model and Hierarchical Clustering." International Conference on Machine Intelligence Research and Advancement,*2013 International Conference on*, pp. 978-0-7695-5013-. IEEE, 2013

[6] H. Jiang, S. Yi, J. Li, F. Yang, X. Hu, "Ant clustering algorithm with k-harmonic means clustering," doi:10.1016/j.eswa.2010.06.061 ELSEVIER 2010.

[7] Pablo Loyola, Pablo E. Rom´an and Juan D. Vel´asquez "Clustering-Based Learning Approach for Ant Colony Optimization Model to Simulate Web User Behavior" 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technolog *International Conference on*, pp. 978-0-7695-4513 - IEEE, 2011.

[8] Eltahir, Mirghani, and Anour F. Dafa-Alla. "Extracting knowledge from web server logs using web usage mining." In *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on*, pp. 413-417. 978-1-s4673-6232-0/13. IEEE, 2013.

[9] Citra Lestari N, Shaufiah, Angelina Prima K."K-harmonic means algorithm for clustering telecommunication customer data" Researchgate publication 2010

[10] Bhargav, Appasani, and Munish Bhargav. "Pattern discovery and users classification through web usage mining." In *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on*, pp. 632-636. 978-1-4799-4190-2/14. IEEE, 2014.