Clustering for web log mining using modified k-means Algorithm

Miss Ruchika Patil
PG student Dept of CSE
YCCE
Nagpur, India
ruchika1patil@gmail.com

Prof. Amreen Khan
Asst. Prof Dept of CT
YCCE
Nagpur, India
amreenkhan786@gmail.com

Abstract— *World Wide Web is a massive repository of web pages and links. Due to the huge amount of data available online, it has become most valuable resource to extract knowledge. Web usage mining is the area of web mining which concerns with the extraction of interesting knowledge from web log information produced by web servers. Techniques of web usage mining can be applied for web log analysis. Analyzing the web logs can help understand the user behavior and the web structure and hence can enhance the website design. Web Session Clustering is one of the crucial techniques which aims to group usage sessions on the basis of some similarity measures. We cluster web users with K-Means and modified K-means algorithm based on web user log data and perform comparative analysis of the algorithms.*

Keywords- Web mining, Preprocessing, Clustering, K-means, Bisecting K-means

*****

I. INTRODUCTION

In this internet era, web sites on the internet are useful source of information in everyday life. The growth of World Wide Web over the last two decades has resulted in a large amount of data that is available for user access. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and service [1]. The substantial increase in the number of Websites presents a challenging task for Website administrators to organize the contents to serve the needs of users. Website administrators may want to know how they can attract visitors, which pages are being accessed most or least frequently, which part of Website is most or least popular and need enhancement, etc. Web mining can be used to discover and extract useful information from the World Wide Web documents and services in order to better understand and serve the needs of Web-based applications [2].

Web mining can be categorized into three areas of interest based on which part of the Web to mine: 1) Web content mining: refers to discovery of useful information or knowledge from web page contents i.e. text, multimedia data like images, audio, video etc. 2)Web structure mining : aims at analyzing, discovering and modeling link structure of web pages and/or web site to generate structural summary. 3) Web usage mining deals with understanding user behavior while interacting with web site, by using various log files to extract knowledge from them [5].

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data and is targeted towards applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis[2][16].

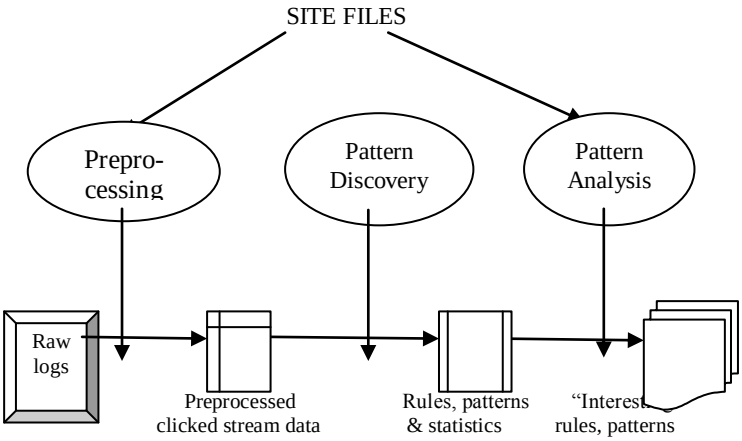

Figure 1: High level Web Usage Mining Process[2]

Web usage mining techniques can be applied for web log analysis. Analyzing the web access logs can help understand the user behavior and the web structure, hence improving the website design. Web log data set, in the form of PDF's are collected from college web site which consists of various reports and summaries. Preprocessing plays a vital role in efficient mining; hence the fields which are not relevant for mining are eliminated. Extraction of useful information is done, which includes taking in account bandwidth and web usage reports and summaries. Here the PDF files are first read and then parsed. Parsing means reading a text and converting it into useful form. It consists of displaying of IP address from the Bandwidth reports and the total bytes communicated by it. Application of the clustering technique by K-Means is done in order to get similar IP addresses and Packet combinations together, thus the clusters contains the number of packets of similar nature. A modified version of traditional K-Means will be proposed. Section II describes the literature survey done, Section III gives the glimpse of proposed work & algorithms

used, Section IV states the experimental illustration and Section V gives conclusion of the work.

.

II. LITERATURE REVIEW

Cluster discovery deals with formation of groups of users exhibiting similar browsing patterns and obtaining groups of pages that are accessed together. A group of users can be Clustered that have similar navigation patterns on a web site. In e-commerce using cluster analysis technique, a group of customers can be clustered with similar browsing navigation and common characteristics of customers can be analyzed. Various data mining methods have been have been used to generate models of usage patterns. In [2, 3], Cooley et al. covered Web usage mining process & various steps involved in it. It serves as the primary knowledge to understand fundamentals of Web usage mining.

Natheer Khasawneh and Chien-Chung Chan [6] have proposed new techniques for preprocessing Web log data including identifying unique users and sessions by making use of Website ontology. User identification algorithm with time complexity O(n) was introduced. Combination of both an IP address and a finite users inactive time to identify different users in the web log is used. For identifying website structure and break points for browsing behavior, website ontology is useful .For session identification, an ontology-based method is presented that utilizes the website structure and functionalities to identify different sessions

Weina Wang, Yunjie Zhang, Yi Li and Xiaona Zhang (2006)[15] In this paper author presented that the Fuzzy *C* Means(FCM) is one of the algorithms for clustering based on optimizing an objective function. Due to above problem, we present the global Fuzzy *C*-Means clustering algorithm(GFCM) which is an incremental approach to clustering. It does not depend on initial conditions and the better clustering results are obtained through a deterministic global search procedure. Experiments show that the global Fuzzy *C*-Means clustering algorithm can give us more satisfactory results by escaping from the sensibility to initial value and improving the accuracy of clustering.

Jin Hua Xu[4] proposed a technique to cluster web usage data. It used simple algorithm K-Mean which is basically built for clustering. So, it is easy and simple to implement for web logs clustering. This method is creates the matrix of of different users to accessed different web pages at specific session. Then the K-Mean algorithm is applied to cluster the data set. The results show the proposed algorithm is feasible, and have scalability.

Peilin Shi[7] In this paper, author proposes a rough k-means clustering algorithm based on properties of rough variable to group the gained fuzzy web access patterns. Users can effectively mine web logs records to discover interesting user access patterns. The interests of web users can be extracted by their visited web pages and time duration on these web pages during their browsing. The fuzzy linguistic variable acts as characterization of time duration on web page because linguistic variable makes users easily understand the expression of time duration and can disregard subtle difference between two time durations. The web access patterns of each users from web logs is converted as corresponding fuzzy web access pattern, which is a fuzzy vector made of fuzzy linguistic variables or 0. Every element in fuzzy web access patterns represents visited web page and time duration on this web page.

Uma Maheswari, Dr. P.Sumathi[1].The author presents a novel approach to clustering Website users into different groups and generating common user profiles. The concept of mass distribution in Dempster-Shafer's theory is used and the belief function similarity measure in the algorithm adds to the clustering task the ability to capture the uncertainty among Web user's navigation behavior. The algorithm is relatively simple to use and gives comparable results to other approaches reported in the literature of web mining

K.Poongothai[13] proposed the usage mining clusters with Expected Maximization (EM). The evolutionary clustering algorithm is proposed to segregate similar user interests. It created the framework with fuzzy C means clustering algorithm and compared with Expected Maximization cluster system. The experimental results of EM represent that by decreasing the number of clusters, the log converges toward lower values and probability of the largest cluster will be decreased while the number of the clusters increases in each web usage pattern. The results indicate the EM approach can improve accuracy of clustering to 11 more. It also shows that the precision of EM is higher than C fuzzy cluster model.

S. Alam [14] presented the review the existing web usage clustering techniques and proposed a swarm intelligence based PSO clustering algorithm for the clustering of sessions of web users. The proposed algorithm works independently without hybridization with any other clustering algorithm. The results showed that the proposed approach performs better than the K-means clustering algorithm for clustering web usage sessions. The results showed the performance of the algorithm is better than K-means clustering. But as analyzed there is difference between the working styles of both the algorithm, because K-means works as a partitioning method and PSO works in hierarchal way.

Keerthiram Murugesan and Jun Zhang [12] In this paper, a hybrid version clustering algorithm is presented that sumps up divisive and agglomerative hierarchical clustering algorithm. The method uses bisect K-means for divisive clustering algorithm and unweighted Pair Group Method with Arithmetic Mean (UPGMA) for agglomerative clustering algorithm. First, the document collection is clustered using bisect K-means clustering algorithm with the initial cluster value K, which is larger than the total number of clusters, K. Then the centroids of K clusters obtained from the previous step are calculated. Algorithm produces better clusters with time complexity of O(N), which is better than the O(N2) time complexity of UPGMA algorithm.

III. GLIMPSE OF PROPOSED WORK & ALGORITHMS USED

Our problem is concerned to improve the Web log content from web log files by efficient k-means (and modified) clustering methods. The proposed work represents K-means clustering algorithm and its accuracy in clustering the data of web log file. The objectives are:

1. To preprocess the Log data to get a reliable data set.
2. To cluster web users: establish groups of users exhibiting similar browsing patterns and to cluster web pages to provide useful knowledge to personalized Web services.

K-means algorithm:

1. Choose k number of clusters to be determined
2. Choose k objects randomly as the initial cluster center
3. Repeat
   3.1. Assign each object to their closest cluster
   3.2. Compute new clusters, i.e. Calculate mean points.
4. Until
   4.1. No changes on cluster centers (i.e. Centroids do not change location any more) OR
   4.2. No object changes its cluster (We may define stopping criteria as well)

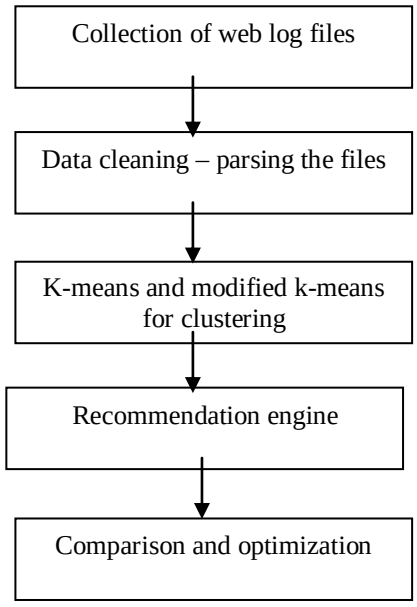

Figure 2: Methodology for proposed work

- **Bisecting K-means**:

Bisecting k-means is a variant of k-means. Instead of partitioning the data set into k clusters in each iteration, bisecting k-means algorithm splits one cluster into two subclusters at each bisecting step(by using k-means) untill k clusters are obtained.

1. Select cluster C(j) to split based on a rule.
2. Find 2 sub clusters of C(j) by using k-means algorithm (bisecting step)
   (a) Select 2 data points of C(j) as initial cluster centroid.
   (b) For each data point of C(j), compute clustering criteria function with 2 centroids and assign the data point to its best choice (calculation step)
   (c) Recalculate 2 centroids based on the data points assigned to them (update step).
   (d) Repeat steps 2(b) and 2(c) until convergence.
3. Repeat step 2 I times and select the split that produces the cluster satisfying the global function.

Repeat steps 1, 2 and 3 until k clusters are obtained

IV. EXPERIMENTAL ILLUSTRATIONS

Experiment was carried out using a log retrieved. The web log files (Log files of Sonicwall) in the form of PDF are collected from reputed Engineering college, YCCE Nagpur. These log files consists of summaries and reports of various information which consist of following : Daily bandwidth usage, Web Usage Summary, Web filter ,Summary, Attack summary, Bandwidth summary ,Top users of bandwidth, Web usage top sites : Hits, Mbytes , category etc.

Preprocessing plays a vital role in efficient mining, hence the fields which are not relevant for mining are eliminated. Extraction of useful information is done, which includes taking in account bandwidth and web usage reports and summaries. Here the PDF files are first read and then parsed. Parsing means reading a text and converting it into useful form. It consists of displaying of IP address from the Bandwidth reports and the total bytes communicated by it.

I. Collection web log files:

- The web log files (Log files of Sonicwall) in the form of PDF are collected from college.

II. Preprocessing:

- Data cleaning.
- Field extraction: The log entry comprises of several fields which need to be isolated for further processing.
- Extraction is done by Parsing the pdf .which is the process of analyzing a string of symbols (here regular expressions)

 ISSN: 2321-8169
223 – 227

III. Application of clustering algorithms.

Step 1: Reading

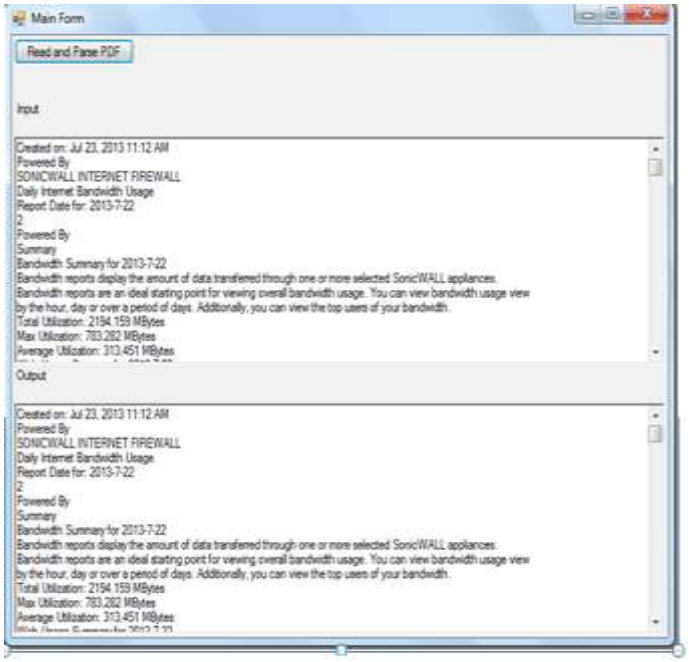

Figure 3: Reading the Pdf

The input data set are the PDF files which are the web log files collected from the college. These log files consists of various information related with bandwidth usage. In the first module, all the text from the PDF files is read first.

Step 2: Obtaining IP addresses

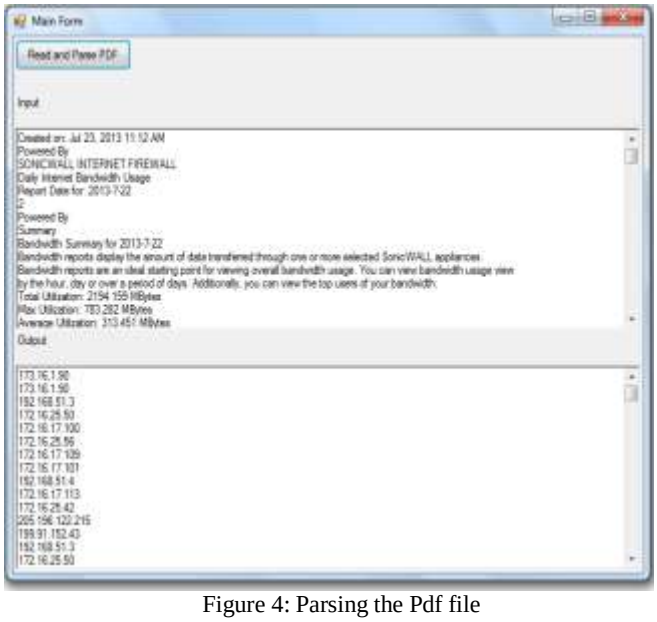

Figure 4: Parsing the Pdf file

The preprocessing procedure involves the data cleaning tasks which performs the task of removing irrelevant records. All the fields which are not relevant (eg. the images) or all the information which are not useful for further mining process are removed. In the web log files, only the bandwidth usage reports were taken in account which consists of various users, their connections, costs (INR), Mbytes and Mbytes %. In the second step, parsing of the PDF file is done. Parsing is the process of analyzing a string of symbols. Here, the text is read and converted into useful form web logs are extracted and displayed.

Step 3: Obtaining IP address along with total bytes communicated.

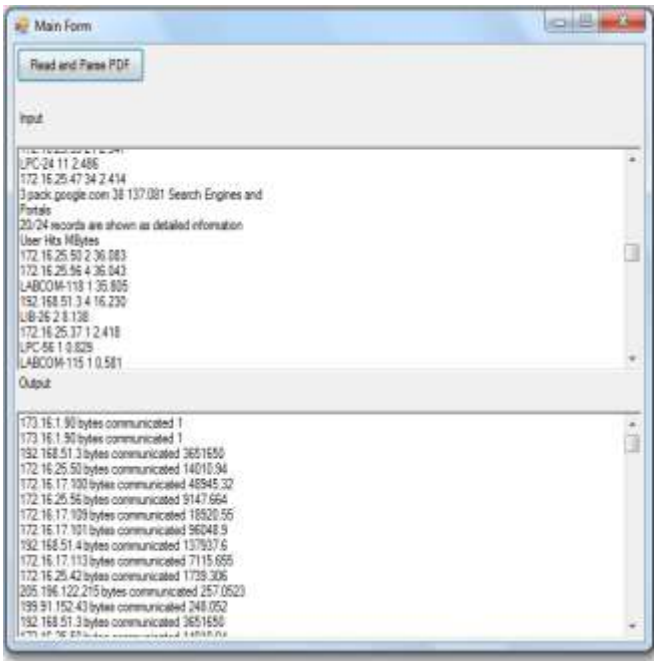

Figure 5: Total bytes communicated by IP

In step 3, the output shows the total bytes communicated by the particular IP address.

Step 4: Application of clustering algorithm

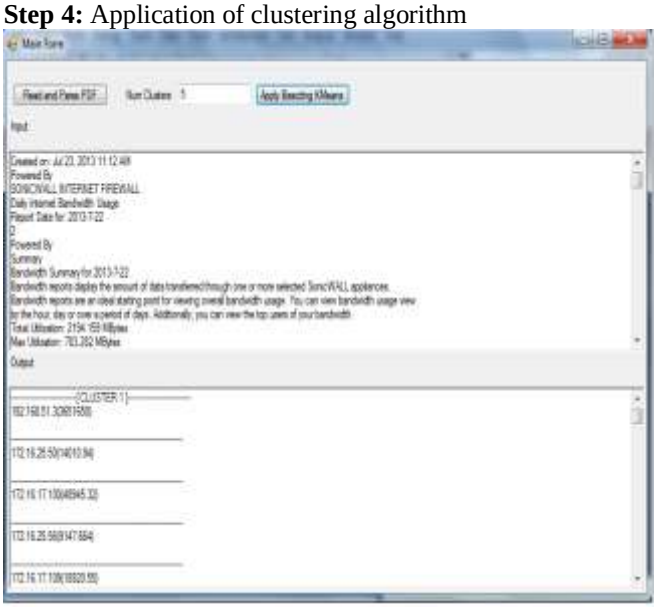

Figure 6a: User clusters

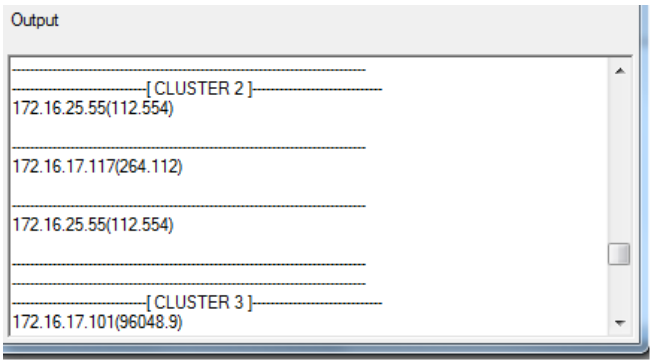

Figure 7b: User clusters

IV. CONCLUSIONS

This paper presented a brief description about Web usage mining, various activities involved in the process and clustering techniques. Our experiments emphasized on clustering algorithms- K-means and its modified version. Further work will be focusing the comparison and optimization of proposed algorithm.

REFERENCES

[1] Uma Maheswari, Dr. P.Sumathi, "A New Clustering and Preprocessing for Web Log Mining", IEEE World Congress on Computing and Communication Technologies, 2014

[2] J. Srivastava, R.Cooley, M. Deshpande, and P. N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Explorations Newsletter, 2000.

[3] T.Kanungo, N. Netanyahu, Angela Y. Wu., "An effective k-means clustering algorithm: Analysis and implementation", IEEE Transactions on pattern analysis and machine intelligence, 2002

[4] JinHuaXu and HongLiu, "Web User Clustering Analysis based on K-Means Algorithm", International Conference on Information, Networking and Automation, IEEE 2010.

[5] Varnagar, C.R., Madhak N.N., Kodinariya, T.M., Rathod, J.N., "Web usage mining: A review on process, methods and techniques", IEEE-Information Communication and Embedded Systems, 2013.

[6] Natheer Khasawneh and Hien-Chung Chan, "Active User-Based and Ontology-Based Weblog data preprocessing for Web Usage Mining", IEEE International Conference, 2006.

[7] Peilin Shi, "An Efficient Approach for Clustering Web Access Patterns from Web Logs", International Journal of Advanced Science and Technology, 2009.

[8] Aisan Maghsoodi, Merlijn Sevenster, Johannes Scholtes, Georgi Nalbanto Philips Research, "Sentence-based Classification of Free-text Cancer Radiology Reports", 25th International Symposium on Computer-Based Medical Systems , 2012, pp.978-4678-2051.

[9] D.Vasumathi, A.Govardhan, K.Suresh, "Effective Web Personalization Using Clustering", International Conference on Intelligent Agent & Multi-Agent Systems, IEEE 2009.

[10] T. Vijaya Kumar and H.S.Guruprasad, "Clustering Web Usage Data using Concept Hierarchy and Self Organizing Map", International Journal of Computer Applications , 2012.

[11] Odukoya, O.H, Aderounmu, G.A. And Adagunodo E.R. "An Improved Data Clustering Algorithm for Mining Web Documents", International Conference on Computational Intelligence and Software Engineering (CiSE), IEEE 2010.

[12] Keerthiram Murugesan and Jun Zhang, " Hybrid Bisect K-means algorithm", International Conference on Business Computing and Global Informatization, IEEE 2011.

[13] K. Poongothai, M.Parimala and Dr. S. Sathiyabama," Efficient Web Usage Mining with Clustering", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[14] Shafiq Alam, Gillian Dobbie, Patricia Riddle, "Particle Swarm Optimization Based Clustering Of Web Usage Data", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[15] Weina Wang, Yunjie Zhang, Yi Li and Xiaona Zhang "The Global Fuzzy C-Means Clustering Algorithm", Sixth World Congress on Intelligent Control and Automation, 2006

[16] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web", IEEE International Conference on Tools with Artificial Intelligence, 1997