Elysium Journal of Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

PERFORMANCE ANALYSIS OF K-MEANS ALGORITHMS IN WEBLOG DATA

K.Abirami [1],Dr. P.Mayilvaganan[2]

[1]Research Scholar,School of Computing Sciences, Vels University,Chennai, India
E-Mail id : abiramidharmarajan@gmail.com
[2]Professor - Dept. of MCA, School of Computing Sciences, Vels University,Chennai, India
E-Mail id: hodmca@velsuniv.ac.in

***Abstract*—Web mining is used to discover interest patterns which can be applied to many real world problems like refining web sites, better understanding the user behavior, product approval etc. Data mining software is one of a number of analytical tools for analyzing data. In this paper we are studying the various clustering algorithms for segmentation model. The basic idea of clustering is to define the similarity between the distance, the distance that represents the data between the data to measure the similarity of the size of the data are classified, until all the data gathering is completed. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. Our main aim to show the performance of K-means algorithm and will be most suitable for the users.**

***Keywords*—Web Mining, K-means algorithms, Clustering methods.**

1. INTRODUCTION

Data mining is a new kind of data processing technology and efficiently extracts useful information [1]. Data mining it is an Extraction of hidden, analytical information from large databases .It is also called as Knowledge Discovery from Databases .It performs an Identification and assessment of hidden patterns in database [1]. Web mining can be classified into three areas: 1) Web content mining: refers to discovery of useful information from web page contents i.e. text, multimedia data like images, audio, video etc. 2) Web structure mining: it refers to analyzing, discovering and modeling link structure of web pages and/or web site to generate structural. 3) Web usage mining deals with understanding user behavior while interacting with web site, by using various log files to extract knowledge from them.

One of the most important tasks of Web Usage Mining is web user clustering which forms groups of users presenting having common welfares and behavior by analyzing the data collected in the web servers. The K-means is most popular algorithm for clustering and well known for its simplicity and low time complexity [1]. However, it has some major drawbacks like quality of the resulting clusters heavily depends on the selection of initial centroids, clusters produced are of varying sizes, hence unbalanced and may also lead to empty clusters.

2. WEB USAGE MINING PROCESS

The main aim of the innovation system is to find web user clusters from web server log files [2]. These discovered clusters show the characteristics of the underlying data distribution. Clustering is useful in characterizing user groups based on patterns, categorizing web documents that have similar functionalities.

This method allows for the collected works of Web log information for Web pages. This usage data provides the paths leading to accessed Web pages [2]. This information is often gathered automatically into access logs via the Web server

Web Usage Mining is a four-step process. The first step is data collection, the second step is data pre-processing, the third step is pattern discovery and the last step is pattern analysis.

2.1 PREPROCESSING

The pre-processing stage involves cleaning of the click stream data and the data is partitioned into a set of user transactions with their respective visits to the web site. “Consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery”.

It performs a series of processing of web log data covering data cleaning, user identification, session identification, path completion and transaction identification.

2.2 DATA CLEANING

It is the process of removing irrelevant items such as jpeg, gif, sound files and references due to spider navigation to improve the quality of analysis. User Identification is the process of identifying users by using IP address and user agent fields of log entries. A user session is considered to be all of the page accesses that occur during a single visit to a Web site.

2.3 PATTERN DISCOVERY

It is the process of removing irrelevant items such as jpeg, gif, sound files and references due to spider navigation to improve the quality of analysis. User Identification is the process of identifying users by using IP address and user agent fields of log entries. A user session is considered to be all of the page accesses that occur during a single visit to a Web site.

2.4 PATTERN ANALYSIS

Pattern Analysis is the final stage of WUM (Web Usage Mining), which involves the validation and interpretation of the mined pattern.

- Validation: to eliminate the irrelevant rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process.
- Interpretation: the output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations.

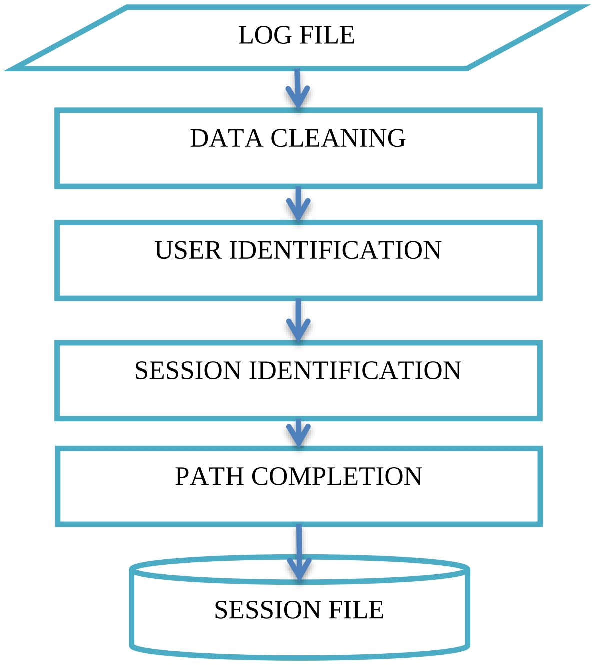


Fig.1. Weblog data preprocessing Phase

2.5 USER IDENTIFICATION

User identification phase will be processed after preprocessing. This step should be to identify unique users. If you use firewalls and proxy servers will be complex to record this information [3]. In EPA web log, each user has individual IP address. So, each IP address represents different user.

2.6 USER SESSION IDENTIFICATION

The purpose of user session identification is to determine the division of access each user has a separate session. The simplest method is to use an expiration time, i.e. the time spent in a page passes a certain threshold, and it is assumed that the user has started a new session. The default time for user session identification is thirty minutes [1]. In this paper for user session identification is considered 30 min expiration time. This default value is used in various studies [1]. Long and convoluted user access paths along with low use of a web page indicate that the web site is not laid out in an intuitive manner. With the help of this analysis, one can re-structure the web site with the navigation results.

3. CLUSTERING ALGORITHMS

Cluster analysis groups objects based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar/related to one other and different from the objects in other groups. Clustering algorithm is classified in to two categories: 1) Decomposition (top-down) 2) Agglomerative (bottom-up). If K-Means and Bisecting K-Means algorithms clusters are decomposed. But Hierarchical clusters are bottom-up approach. It is deterministic. The clusters will be creating a complete binary tree. The various clustering algorithms are compared and find which one is the best.

3.1 K-MEANS ALGORITHM

K-mean is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters.

Basic K-means Algorithm for finding K clusters [*]

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

K- Means algorithms follows.

```
K-means-cluster (in S : set of vectors : k : integer)
{  let C[1] ... C[k] be a random partition of S into k parts;
   repeat {
         for i := 1 to k {
           X[i] := centroid of C[i];
           C[i] := empty
          }
         for j := 1 to N {
           X[q] := the closest to S[j] of X[1] ... X[k]
           add S[j] to C[q]
   }
}


  until the change to C (or the change to X) is small enough
}
```

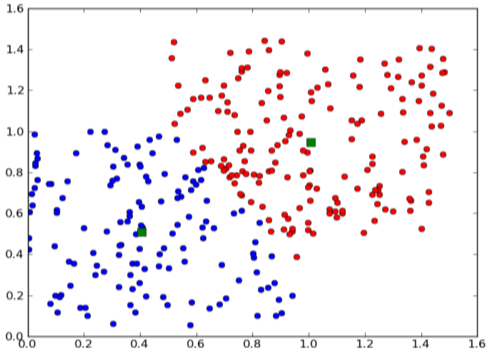


Fig.2.

Three clusters are placed to one area

In this case we splitted the data in 2 clusters, the blue points have been assigned to the first and the red ones to the second. The squares are the centers of the clusters.

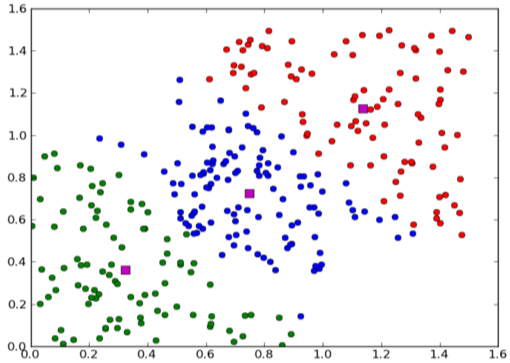


Fig.3. Three clusters are placed to one area

After using K-Means algorithm three clusters are placed in areas separated in different places in order. The pink square indicates centroids of the clusters. Web log data set, in the form of log file are collected from college web site which consists of various reports and summaries. These files are in nonstandard format. Extraction of useful information is done, which includes taking in account bandwidth and web usage reports and summaries. Here the log files are first read and then parsed. Parsing means analyzing 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.

4. LIMITATIONS

K-means clustering has some of the limitations which need to get overcome. Several people got multiple limitations while working on their research with K-means algorithm. Some of the common limitations are discussed below.

4.1 OUTLIERS

Outlier has been observed by several researchers that, when the data contains outliers there will be a variation in the result that means no stable result from different executions on the same data. Outliers are such objects they present in dataset but do not result in the clusters formed. Outliers can also increase the sum of squared error within clusters. Hence it is very important to remove outliers from the dataset. Outliers can be removed by applying preprocessing techniques on original dataset.

4.2 NUMBER OF CLUSTERS

Determining the number of clusters in advance is always been a challenging task for K-means clustering approach. It is beneficial to determine the correct number of clusters in the beginning. It has been observed that sometimes the number of clusters is assigned according to the number of classes present in the dataset. Still it is an issue that on what basis the number of clusters should be assigned [8].

4.3 EMPTY CLUSTERS

If no points are allocated to a cluster during the assignment step, then the empty clusters occurs. It was an earlier problem with the traditional K- means clustering algorithm.

4.4 NON GLOBULAR SHAPES AND SIZES

With the K-means clustering algorithm if the clusters are of different size, different densities and non-globular shapes, then the results are not optimal. There is always an issue with the convex shapes of clusters formed.

5. APPLICATIONS

There are diverse applications of clustering techniques in the fields of finance, health care, telecommunication, scientific, World Wide Web, etc. Some of the applications are discussed below.

5.1 CLUSTERING ALGORITHM IN IDENTIFYING DISEASED DATA

Clustering algorithm can be used in identifying the diseased data record within a dataset. Different people tried on this application by assigning labels to known samples of datasets as diseased and non-diseased. Then randomly the data samples are mixed together and different clustering algorithms were applied. The result of clustering has been analyzed to know the correctly clustered samples. Accuracy of clustering is calculated easily as the labels of samples were known initially.

5.2 CLUSTERING ALGORITHM IN SEARCH ENGINES

Clustering algorithm plays an important role in the functioning of search engines. Hence it will act as a backbone to search engines. Search engines try to group similar kind of objects into one cluster and dissimilar objects into other. The performance of the search engines depend on the working of the clustering techniques. The chances of getting the required results on the front page are more if the clustering technique is better.

5.3 CLUSTERING ALGORITHM IN ACADEMICS

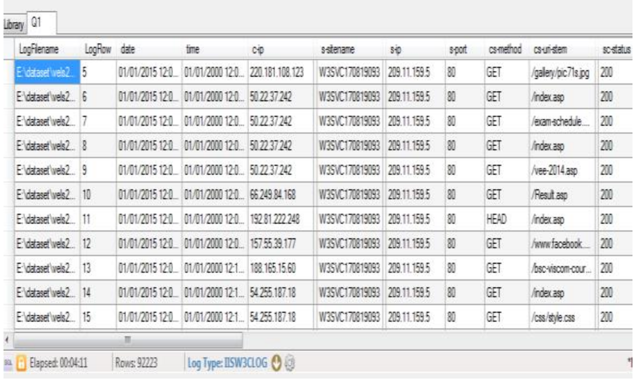
Students' academic progress monitoring has been a vital issue for academic society of higher learning. With clustering technique this issue can be managed easily. Based on the scores obtained by the students they are grouped into different clusters, where each cluster shows the different level of performance. By calculating the number of students' in each cluster we can determine the average performance of a class all together [8].

5.4 CLUSTERING ALGORITHM IN WIRELESS SENSOR NETWORK BASED APPLICATION

Clustering Algorithm can be used efficiently in Wireless Sensor Network's based application. It can be used in landmine detection. Clustering algorithm plays a role of finding the cluster heads which collects all the data in its respective cluster.

6. EXPERIMENTAL RESULTS

Experiment was carried out using a log retrieved. The web log files (Log files) in the form of LOG are collected from Vels University Chennai.

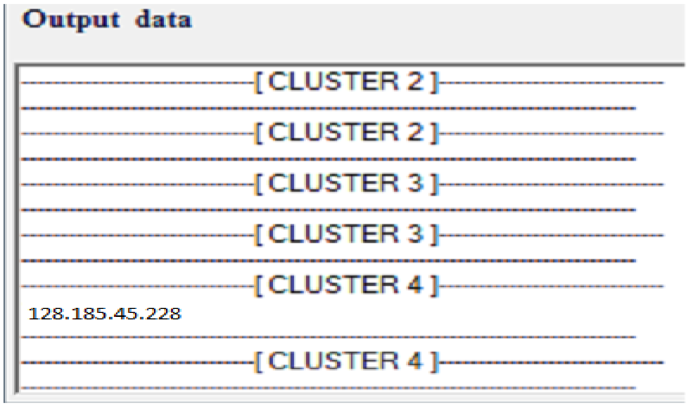


LogFilename	LogRow	date	time	c-ip	s-sitename	s-ip	s-port	cs-method	cs-uri-stem	sc-status
E:\dataset\vels2...	5	01/01/2015 12:0...	01/01/2000 12:0...	220.181.108.123	W3SVC170819093	209.11.159.5	80	GET	/gallery/pic71s.jpg	200
E:\dataset\vels2...	6	01/01/2015 12:0...	01/01/2000 12:0...	50.22.37.242	W3SVC170819093	209.11.159.5	80	GET	/index.asp	200
E:\dataset\vels2...	7	01/01/2015 12:0...	01/01/2000 12:0...	50.22.37.242	W3SVC170819093	209.11.159.5	80	GET	/exam-schedule....	200
E:\dataset\vels2...	8	01/01/2015 12:0...	01/01/2000 12:0...	50.22.37.242	W3SVC170819093	209.11.159.5	80	GET	/index.asp	200
E:\dataset\vels2...	9	01/01/2015 12:0...	01/01/2000 12:0...	50.22.37.242	W3SVC170819093	209.11.159.5	80	GET	/vee-2014.asp	200
E:\dataset\vels2...	10	01/01/2015 12:0...	01/01/2000 12:0...	66.249.84.168	W3SVC170819093	209.11.159.5	80	GET	/Result.asp	200
E:\dataset\vels2...	11	01/01/2015 12:0...	01/01/2000 12:0...	192.81.222.248	W3SVC170819093	209.11.159.5	80	HEAD	/index.asp	200
E:\dataset\vels2...	12	01/01/2015 12:0...	01/01/2000 12:0...	157.55.39.177	W3SVC170819093	209.11.159.5	80	GET	/www.facebook....	200
E:\dataset\vels2...	13	01/01/2015 12:0...	01/01/2000 12:1...	188.165.15.60	W3SVC170819093	209.11.159.5	80	GET	/bsc-viscom-cour...	200
E:\dataset\vels2...	14	01/01/2015 12:0...	01/01/2000 12:1...	54.255.187.18	W3SVC170819093	209.11.159.5	80	GET	/index.asp	200
E:\dataset\vels2...	15	01/01/2015 12:0...	01/01/2000 12:1...	54.255.187.18	W3SVC170819093	209.11.159.5	80	GET	/css/style.css	200

Elapsed: 00:04:11 | Rows: 92223 | Log Type: IISW3CLOG

Fig.4. Reading web log file and Parsing the file

The input web log file is the log file, which is first read by the system. The useful text is extracted from a large data and then parsed. Parsing involves displaying of various IP address and total bytes communicated by them.



```
Output data

-----------------------[CLUSTER 2]-----------------------
----------------------------------------------------------
-----------------------[CLUSTER 2]-----------------------
----------------------------------------------------------
-----------------------[CLUSTER 3]-----------------------
----------------------------------------------------------
-----------------------[CLUSTER 3]-----------------------
----------------------------------------------------------
-----------------------[CLUSTER 4]-----------------------
128.185.45.228
----------------------------------------------------------
-----------------------[CLUSTER 4]-----------------------
----------------------------------------------------------
```

Fig 5.Formation of clusters using K-Means

Fig 5 shows various empty clusters generated using K-means. Here, we can see a single cluster (cluster1) consisting of large portion of the dataset.

7. CONCLUSION

In this paper, we have made a survey on work carried out by different researchers using K-means clustering approach. This article present for clustering algorithm on weblog data. The preprocessing technique is applied in the log parser tool. Here we have compared algorithms theoretically and experimentally on parameters such as time taken to build model, clustered instances,etc. Finally, the Performance of K-means algorithms were used the web log data is efficient. In future Work, the accuracy will be calculated using k-means and compare another clustering algorithms. For which algorithms were best is suitable for web user.

REFERENCES

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000

[2] Hesam T. Dashti, Tiago Simas, Rita A. Ribeiro, Amir Assadi and AndreMoitinho "MK-means - Modified K-means clustering algorithm" ,*IEEE* ,978-1-4244-8126-2/10/$26.00 ©2010

[3] FAHIM A.M., SALEM A.M., TORKEY F.A., RAMADAN M.A." Anefficient enhanced k-means clustering algorithm" J Zhejiang Univ SCIENCE A 2006 7(10):1626-1633

[4] Pritesh Vora and Bhavesh Oza "A Survey on K-mean Clustering and Particle Swarm Optimization", International Journal of Science and Modern Engineering (*IJISME*) ISSN: 2319-6386, Volume-1, Issue-3, February 2013

[5] Bangoria Bhoomi M. "Enhanced K-Means Clustering Algorithm to Reduce Time Complexity for Numeric Values International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 876-879

[6] M.SANTHANAKUMAR, C.CHRISTOPHER COLUMBUS," Web Usage Based Analysis of Web Pages Using RapidMiner" E-ISSN: 2224-2872, Volume 14, 2015.

[7] Kaushik H. Raviya ** Kunjan Dhinoja, "An Empirical Comparison of K-Means and DBSCAN Clustering Algorithm" PARIPEX - INDIAN JOURNAL OF RESEARCH, Volume : 2 | Issue : 4 | April 2013 ISSN - 2250-1991.

[8] Shraddha Shukla and Naganna S, " A review of K-means Data clustering Approach, International Journal of Information & Computation Technology. Volume 4, Number 17 (2014), pp. 1847-1860.

[9] Ruchika R. Patil, Amreen Khan, " Bisecting K-Means for Clustering Web Log data" International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 19, April 2015.