International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 92-94
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

A NEW HYBRID CLUSTERING APPROACH FOR WEB MINING

Preeti
Department of CSE
Punjab Technical University
Kapurthala, India

Priyanka Kalia
Department of CSE
Punjab Technical University
Kapurthala, India

Malti Rani
Assistant Professor, CSE
Punjab Technical University
Kapurthala, India

***Abstract*— Clustering analysis is the foremost technique in data mining; its process will affect the clustering results directly. In this paper knowledge discovery in database process is explained which is used for the mining of the knowledge data. In the proposed work a new hybrid approach is used so as to enhance its speed or increase its efficiency. In this, we need to breakdown the dataset into subsets and then employ the reduced iteration approach over the subsample in order to get the clusters. This process will preserve lot time and enhance the functioning for the large datasets.**

Keywords— ***Data mining (DM), web mining(WM), web structure mining and web content mining, Knowledge discovery database process(KDD)Process***

I. INTRODUCTION

Nowadays, Quantity of data stored in computer files and databases is rising at a phenomenal rate. At the same time user of these data are expecting more knowledgeable information from them. In order to get the knowledge information from the data various data mining techniques are used. Data mining is exploration phase of the "knowledge discovery in database" process. It is the process of finding out the patterns in huge data sets which engages the schemes at the intersectional of artificial intelligence, database and machine learning. Generally data mining process is used to mine knowledgeable information from the large dataset and renovate the information into understandable structure for future. In DM communities, there are three kinds of mining: data, text and web mining [1]. In data mining the data is organized in database in a structured manner, while text mining deals with the shapeless text. Web mining is positioned between the both semi structured and Unstructured data. The mining may diverge from structured and unstructured. Foremost areas where KDD application used are marketing, fraud detection, telecommunication and manufacturing.

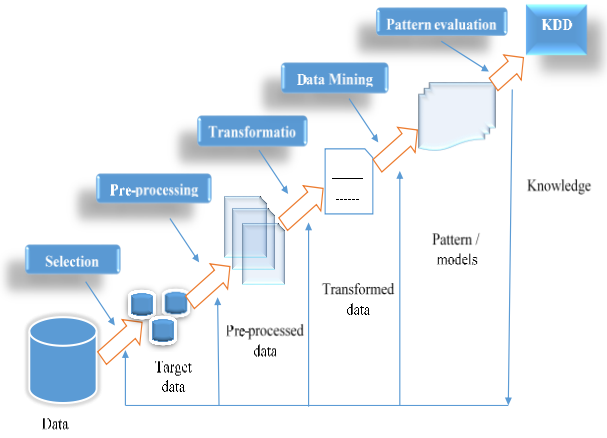

Fig. 1: Architecture of data mining (KDD Process)

WM is an approach of DM techniques which is used to determine the patterns from the world. Moreover this technique is used to find out as well as extract the information from web documents and services. This is further divided into 3 categories: web usage mining (wum), web content mining (wcm) and web structure mining (wsm). A set of items which belongs to a particular class is known as cluster. Or we can say, identical items belongs to one cluster whereas unlike items are belongs to another. Clustering composes a significant or valuable cluster by using automatic approaches of those items which have identical features. In order to clearly compose the idea, refer book management in library being an item. Lots of books are available over various topics in a library and how to keep those books in a particular manner so that the readers can have number of books over specific topic without any trouble is a challenge. And by the use of clustering approaches, one can maintain books in such a way, that the books which have some kind of relation will be kept in one cluster or over one

shelf and label it with a significant name. If the reader desire to take books over a particular topic, then he needs to go to that particular shelf instead of going through the entire library[5].
The rest of the paper is organized as follows: Section II covers the literature survey, Section III deals with the related work, Section IV consists of proposed Methodology and finally section V will conclude the paper.

II. LITERATURE SURVEY

Kavita Sharma et.al "Web Mining: Today and Tomorrow" this paper presented the study about how to extract the useful information on the web and also give the superficial knowledge and comparison about data mining. This paper describes the current, past and future of web mining. Here we introduce online resources for retrieval Information on the web. Furthermore, this paper also describe the web mining through cloud computing i.e. cloud mining. That can be seen as future of Web Mining.

Omer Adel Nassaret.al "The integrating between web usage mining and data mining techniques" In this paper, a number of web usage mining scenarios are presented which depends on the available information. The integration between the Web usage mining and data mining techniques are presented for the processes at different stages which consist of pattern discovery stages, and initiates banks cases which have analytical mining technique. In this paper a universal structure for fully integrating domain Web usage mining and data mining techniques is represented [2].

Amainderkaur et.al (2010) [10] "Effect of noise on the performance of clustering techniques". The objective of the paper is to show the effect of noise on the performance of various clustering Techniques. Clustering is being widely used in many application including medical, finance and etc. Clustering may be applied on database using various approaches, based upon distance, Density, hierarchy, and partition. This paper discusses about the performance of clustering technique in the presence of noise. Noise can appear in many real word datasets and heavily corrupt the data Structure. We have calculated the results in the presence of different percentage of noise by using k-means and PAM algorithms [3].

Manjot kaur et.al "web document clustering approaches using k-means algorithm" This paper describes the data mining process which is used to get the valuable information from a data set and renovate the same into a significant structure. In this approach K-means clustering algorithm is a well-organized learning algorithm to which is used for the famous clustering problems. Fewer identical based clustering approaches is proposed in this paper, which is used for discovering the better initial centroids and to grant a proficient manner of assigning the data points to an appropriate clusters along with lower time complexity [4].

III. RELATED WORK

Various data mining techniques have been developed and used in various data mining projects these days. In the proposed work, k –means performance will be enhanced by using hybrid approach for better result. So as to show the result of noise on the performance of various clustering techniques, Clustering may be applied on database using various approaches, based upon distance, density hierarchy and partition. Clustering being widely used in many applications including medical, finance etc. our purpose is to study how a particular clustering technique is quick to respond to the noise in the term of time

IV. PROPOSED METHODOLOGY

In these methodology two approaches has been merged by initially partitioning the dataset into small ones and then using the reduced iteration approach to each subset in order to get the clusters. Through this process lot of time will be saved and also it helps to increase the functioning of the large datasets.

1) Break the entire data set into multiple subsamples.

2) After splitting employ the algorithm that will lesser the number of iterations.

In this we have used a terminology in which we have assembled dataset which is related to web mining. And by filtration preprocessing of that particular dataset occurs which consist of removal of both noise and missing values. Then K-means clustering algorithm is employed and the results will be observed. When the tactic of K-means clustering algorithm together with reduced iterations and K-means clustering algorithm in addition to both the reduced iterations and sub samples are applied then results will be observed again. The effectiveness of the algorithms is measured in terms of time taken in seconds.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 6, ISSN No. 2455-2143, Pages 92-94
Published Online April - May 2016 in IJEAST (http://www.ijeast.com)

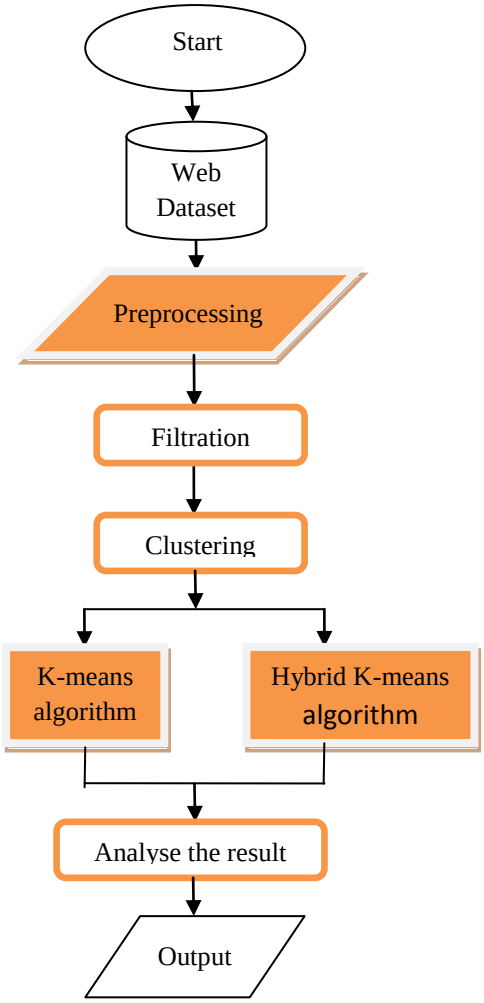

Fig 2: Proposed Methodology

V. CONCLUSION

In this paper, a new hybridizes method based on k-means clustering method proposed to cluster data. In the proposed method, to find optimal cluster centers and then initialized the k-means algorithm with this centers to refine the center.

VI. REFERENCES

[1] Q. YANG and X. Wu, **10 challenging problems in data mining research**, int. J inform. Technol. Decision making 5(4) (2006) 597-604

[2] Omer Adel Nassar and Dr. Nedhal A. Al saiyd,**"The integrating between web usage mining and data mining techniques"** 2013 5th International conference on computer science and information technology (CSIT) 2013, pp.243-247.

[3] Amainderkaur and pankajkumar**"Effect of noise on the performance of clustering techniques"** 2010 international conference on networking and information technology,IEEE, pp.504-506

[4] Manjotkaur, Navjotkaur **"Web document clustering approaches using k-means algorithm"**, IJARCSSE, volume 3, number 5,, 2013, pp.861-864

[5] Data mining in banks and financial institutions http://www.rightpoint.com/community/blogs/viewpoint/archive/2011/11/08/data-mining-in-banks-and-financial-institutions.aspx