Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 1,Issue 12, December -2014

# SURVEY ON AMELIORATE DATA EXTRACTION IN WEB MINING BY CLUSTERING THE WEB LOG DATA

Jasmine M Chaniara[1], Prof. Firoz Sherasiya[2]

[1]*M.E. [Computer Engineering], Darshan Institute of Engineering & Technology, Rajkot,jasmine.chaniara@gmail.com*
[2]*M.Tech. [Computer Engineering], Darshan Institute of Engineering & Technology, Rajkot, firoz.sherasiya@darshan.ac.in*

**Abstract** — *Web mining process can largely express as discovery and analysis of suitable data information from the World Wide Web. It can be express as the exploration and analysis of different pattern, during the web mining of web log files and linked data from a particular weblog file, in a manner that will be more efficiently for the user while using web. Web usage mining is a part of web mining, which defines data mining techniques to find the important relevant information as per the usage of user of web. The initial phase of the web usage mining is the processing of the data .Session reconstruction is the most important work of web usage mining since it directly emphasizes on the quality of the patterns which are extracted frequently. Similarly another factor affecting the web data extraction is the depend upon how the clusters are form of the web data in the web usage mining. Forming the cluster of the web data efficiently enhance the searching speed, decrease the searching time and give the most relevant data set to the user. There are several algorithms are apply for the clustering the data to partition it according to the web usage factors like most viewed pages , ranking of the page, dataset, etc. Clustering is one of the main web data analysis methods and k-means algorithm is one of the popular algorithms. [1]There exist already many new technique were already proposed to improve the efficiency and performance of the k-means algorithm, but it need extra efforts and parameters to improve the efficiency. But with the initial centroid[1] the efficiency of k mean will be improved without any other inputs. In this paper we proposed by taking a web log of a user searching into consideration that if we divided the whole web log data into cluster using k mean with initial centroid algorithm instead of only K mean algorithm and on those clusters we will applied different web extracting techniques to retrieve that search item then the combinations result will be more efficient.*

***Keywords***-*Web mining, Kmeans, Clustering, Kmeans with initial centroid, web log, page ranking*

## I. INTRODUCTION

The world is now converting everything into digital data. Everything is now online from the starting of alarm of the day till the late night. Any need of information is available on the www (World Wide Web) or more frequently we can say Internet. Internet is a largest source of digital data. We can found very answer from the World Wide Web, so there must huge data storage and servers are there to serve us. It may not be possible that anyone is untouched from the digital world. Everyone has their own data and requirement of the information into the internet. Due to such a huge usage of internet day by day there are many problems are emerging into the front of us. Explosive growth of data from terabyte to petabytes creates massive data sets. Due to such huge amount of data the day by day the unnecessary data will go on increasing and users has to surf more and more to get their required information from the internet. Such a situation on Web data create a problem to particular users to while searching on the internet. So to rectify this problem the Web data mining is done. That is to remove the unnecessary data from for the user and give them the required information or data they are looking for. So from such a huge amount of data the necessary, useful and some worthy information is extracted is called as web mining. Web mining is very helpful in e-commerce, online transactions, and stocks, Similarly on Science: Remote sensing, bioinformatics, scientific Society. We are drowning in data, but starving for knowledge. So simply we can state that data mining refers to extracting or "mining" knowledge from large amounts of data, extracting the meaningful and important relevant data from the huge web data. It is an application of data mining techniques to discover the patterns from the web which is called as the web mining [1]. Web mining can be further divided into the three categories which are:

- Web content mining.
- Web usage mining
- Web structure mining.

Another relative and important concept of increasing the efficiency of data retrieval is clustering. It's main work are exploratory data extracting , and similar technique for statistical analysis of data, similarly it is used in many fields,

 

including machine learning ,reorganization of pattern , analysis of image, information gaining, and bioinformatics. Clustering analysis itself is not a specific algorithm or task, but general task to be solved. Plenty of different algorithms are used into clustering based on the their category [2]. The clustering algorithms are divided broadly into the five sub categories. They are :-

- Partition Clustering Algorithm
- Hierarchical clustering
- Density Based Algorithm
- Grid based Method
- Model Based Method

Cluster of data sets are formed within a cluster on the basis of having high similarity between one another, but are rather dissimilar to object in other clusters. It is an unsupervised learning technique.

## II. BACK GROUND THEORY

### 2.1 Web Mining

Web mining is an application or tool of data mining techniques to discover the patterns from the web logs and websites. It crawl through the various web sources to collect the required information, which enable one to find the relative patterns and depend upon that analysis is done on that web data.

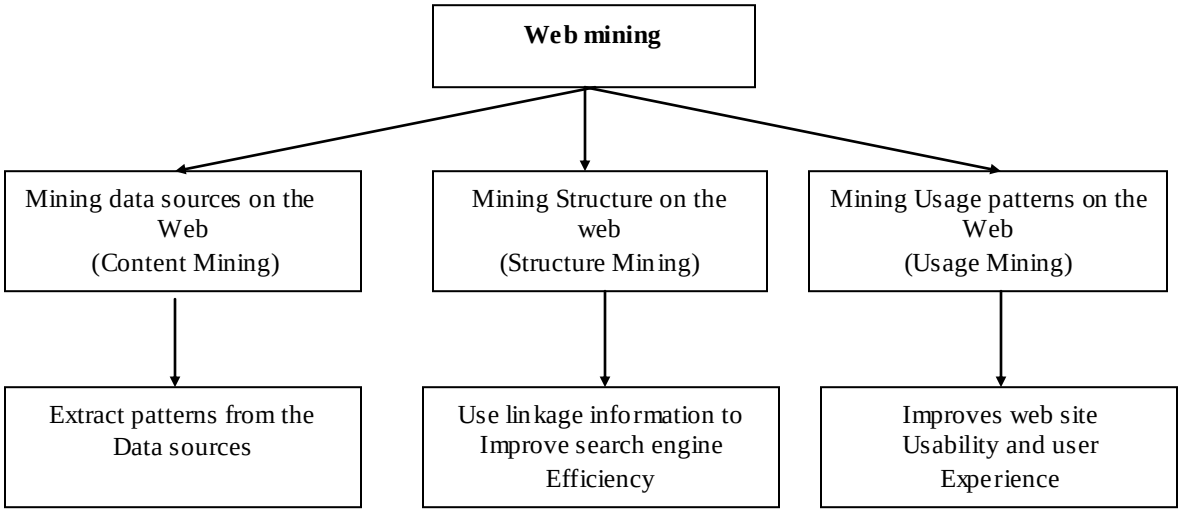


Fig 2.1 Web mining Taxonomy [1]

#### 2.1.1 Web Content Mining[1][3]

Web content mining is the mining, extracting, analyzing and integrating the useful data, information and knowledge of web pages contents. The various heterogeneity and the structure of that allow much of the ever-expanding information from the World Wide Web like organization hypertext document and search and indexing tools of the internet like web crawler. But they do not generally provide structural information nor the categories in current years these factors have highlighted the research to developed more intelligent tools for getting the relevant information from the available on the web. Web content mining is also known as the text mining it is probably the second step into the Web data mining Content mining is the scanning and mining of the content like text, images , graphs of a Web page to determine the relevance of the content according to the user search. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. Following is the diagram showing the web content mining process; it shows that how web crawler or web spider will go through all the content put on the web by anyone. It will analyzed that web content and store into the web database accordingly.

 

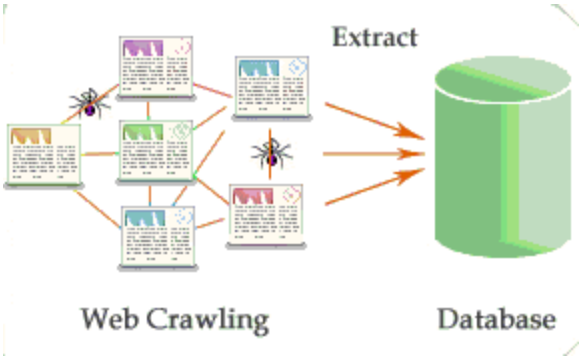


Fig 2.2 Web content mining

Content data is the accumulation of facts a web page is designed to contain. It may consist of text, pictures, videos, audios or structured records such as lists of tables. Application of text mining to web content has been the most extensively researched. Text mining include topic discovery and tracking, extracting combining patterns, clustering of dataset of web documents and classification of web pages.

**2.1.2 Web Stucture Mining[1][3]**

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site[wiki]. Web structure mining is one of the mining categories of the web mining of data, which is used to differentiate the relationship amongst the Web pages linked different connections or directly connected by link. This structure can be identified by the web structure schema for database techniques used in web pages. This techniques used spider scanning the web sites, retrieving the relevant data from the from the home page, links into home pages and different reference into the pages [4]. Web structure mining can be categories on the basis of web structure data as:-

- Extracting patterns from the hyperlinks from web.
- Mining of the different document structures: analysis of tree like structure of different web pages of html and xml formats.

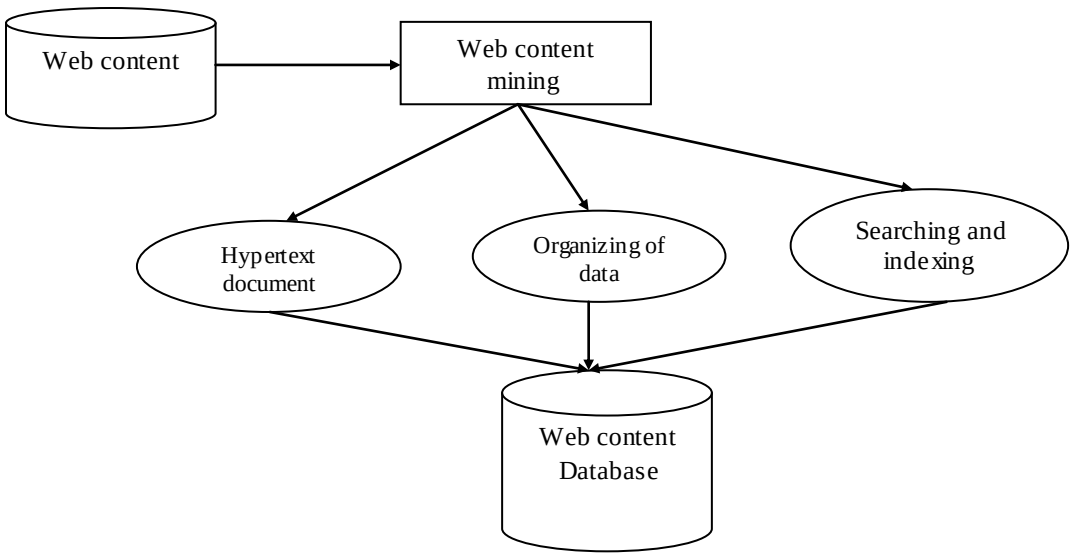


Fig 2.3 Web Content mining

**2.1.3 Web Usage Mining[1]**

Web usage mining is the process of extracting useful information from web server logs e.g. What the user have searched, at what time and from which IP all these is include in usage mining process [5]. The information often store into the web logs , user subscription and the user surf on the internet logs. As the different users have different need according to their work some of the users need text documents, some others are interested into the multimedia content like images, videos, flash etc. The different web usage techniques are used to find the interesting patterns from the weblogs and the web data in order to understand the need of the users searched. So the server can serve the user in better way and provide their relevant data according to the behavior of the users. According to the different patterns identified

 

by which are frequently used by the user and give the information about the user behavior and type of searching content. Following is the figure of the web usage mining into the web mining.

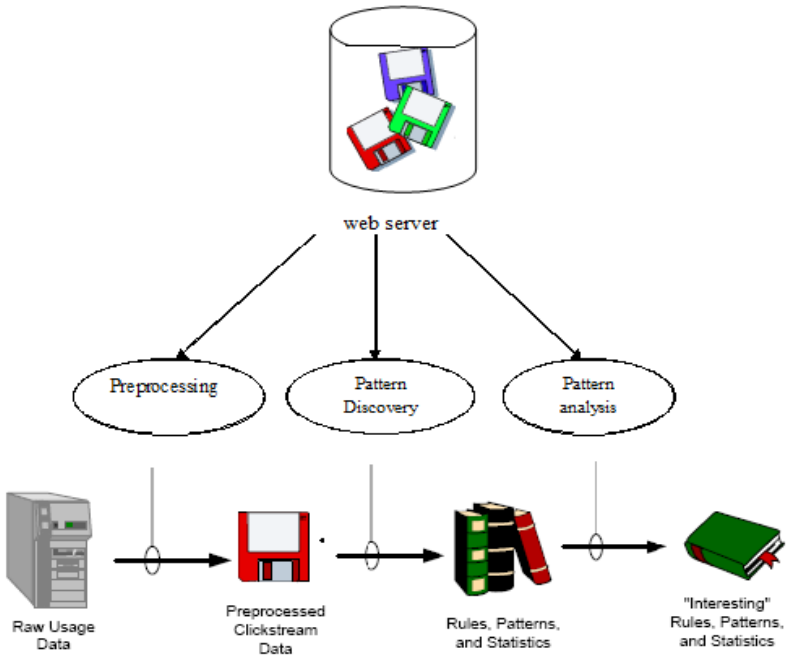


Fig 2.4 Web usage mining

**2.2 Clustering**

A group of similar thing, people or events related closely together is called as a cluster of that thing, people or events. Similarly computer cluster is a set of loosely or tightly coupled computers which are going to work together as a single system then such system is called as computer cluster. It mainly used in the exploring the data mining and a common technique for data analysis, used in many field like including machine learning, pattern reorganization , image analysis , information extracting or retiring. Clustering can be considered as the most important unsupervised learning technique. So it will deal with the structure in the collection of the unstructured data. A normal definition of the cluster can be defined as "the process of contaminating or organizing the different objects into groups where there are similarities in the objects in one or some way"

**2.2.1 Need of clustering**

The main needs of the different lustering algorithm are to divide the data into small groups to differentiate them, easy to revive particular type and less expense of the resource. Different clustering algorithm are applies for clustering to satisfy some of the following result[2].

- Improve efficiency of the resources.
- Affecting the different attributes of objects.
- Differentiate cluster of arbitrary shape.
- Ability to handle noise and outliers.

**2.2.2 Some Clustering Methods[3]**

- Hierarchical methods :
  This method use recursive partitioning the instances in either top down manner or bottom up manner. These methods can be sub divides as :
  i. Agglomerative hierarchical clustering.
  ii. Divisive hierarchical clustering.
- Partitioning methods:
  i. Error minimizing algorithm.
  ii. Graph theoretic clustering.
- Density based clustering.
- Model based clustering;
  i. Decision tree.
  ii. Neural network.
- Grid based method

 

## III. LIETRATURE SERVEY

The main problem in web usage mining is getting the relevant data from the whole database while searching, which is seen and approve by many other users while searching the same content. The purpose of the extracting information form the web log using rough set theory is to build better searching and better design of the websites [4].The factor affecting the cost and time is searching and computation in it. Accumulate the of the user interaction with cluster technique. the cluster are form on a particular basis having some similarities between the interaction of the user or a pattern into user interaction. On the basis of that we can definitely identify the user nature and behavior of it. So by analyzing it the websites can be making more users friendly and interactive to users. Pattern based clustering can help the analyst to provide best result of their need [5]. Clustering by different algorithm to get the efficient clustering of the data .One of the algorithm is k mean clustering algorithm . In the proposed paper there are two ways to find the better in itial centroid point .first to find better initial centroid, so to reduce the computation and cost. Another method is to assign the better data point to kmeans algorithm to reduce time and computation [6]. The preprocessing of the weblog data by the means of USIA(User and Session Identification) algorithm it find the user and session detail . If the user from the same computer will access the web again then it can be identify by their ip. So on the basis of that we can find the particular interest of the different users using session and user [7]. In this proposed work enhancing methods of the assigning the center point to the k means algorithm is discussed. As in the basic k mean algorithm after each iteration the distance is calculated between the data and the center point which is set . these will lead to increase in the computational time of the k mean depend upon the data sets. If the data set is more than more comparison and more computation is needed. So to reduce this computation an improvement in the k mean algorithm is done by selecting the initial centroid in such a way that that after that the Kmeans will provide efficient result of the clustering [8]. Zhang chen et al discuss the initial centroid k mean algorithm having voided alternative randomness in finding the initial centroid. so to get the efficient result of the clustering the data set [9]. In selecting the initial centroid another approach by Fang Yuan state that if in the original algorithm of kmeans k –objects are randomly selected as the initial centroid from the data set then select the centroid the object according to the output of the basic k mean algorithm [10]. One of another approach of improve the clustering of the data is by combining two algorithms. The combination of two Kmeans and the hierarchical algorithms. The result of this combination will help in finding the more efficient centroid in the k mean algorithm so to reduce the computation and cost[11]. One of the proposed novel algorithm for clustering known as Divisive Correlation clustering algorithms (DCCA). This algorithm also takes in consideration the initial centroid. But the computation of the algorithm is higher and the cost is also increase [12]. The search result also affected by the rank of the page in web .It is one of the factors in getting the relevant and trusted result from the search topic. Page rank algorithm was discovering by the Larry page. It is used by Google internet search engine. In this algorithm the a numeric value is given to every page and on the basis of that rank assign to every link they are arrange when the user search some related content of the the pages. On the basis of the hits that are view and click of the pages the numeric value increase of that particular page. More the page click and view on the basis of that the value of the page rank increase [13]. In extend to the page rank another algorithm of page rank is weighted page rank algorithm. In this it is proposed that assigning a grater rank to the pages instead of dividing the page rank by the total out coming links. The algorithm is more efficient than the basic page rank algorithm because this algorithm use two parameters. It considers both forward link and backlink into consideration. So on the basis of both in link and out link the page ranking is done [14]. Another algorithm named Weighted Page Content Rank Algorithm (WPCR) is there for ranking the pages. This algorithmic gives sorted order or the searched web pages from the search engine according to the user query. Weighted Page Content Rank Algorithm assign a numeric value same as the page ranking and Weighted Page Rank Algorithm, but it will use web usage mining technique as well. Weighted Page Content Rank Algorithm (WPCR) first find the ranking according to the pages. The importance of a page can be find by how many pages are referring to a particular page so it will be hits more and relevance of the page depend upon the search query of the user and matching result into the content of the page. This algorithm is better than both the previous page rank algorithms [15].

## IV. PROPOSED WORK AND LIMITATATION

As we have discuss different algorithm on clustering. In which we discuss that which algorithm is better and why according to the different reach work on those algorithms. The different survey papers shows that kmeans with initial centroid is better algorithm on an average for doing the clustering. Similarly we have gone through the page ranking algorithms by which the the user will get the most relevant and accurate searching result of it's query.

So by combining to technique the result can be more effective and more accurate. In the proposed work when a user will search the query on web server it's web log is created from instead of only rank relevancy we consider time , IP and depending upon the rank relevancy report, more efficient results may be obtained by checking the popularity of that web page or considering the particular time slot which will provide current relevant data by forming the clusters on that basis

 

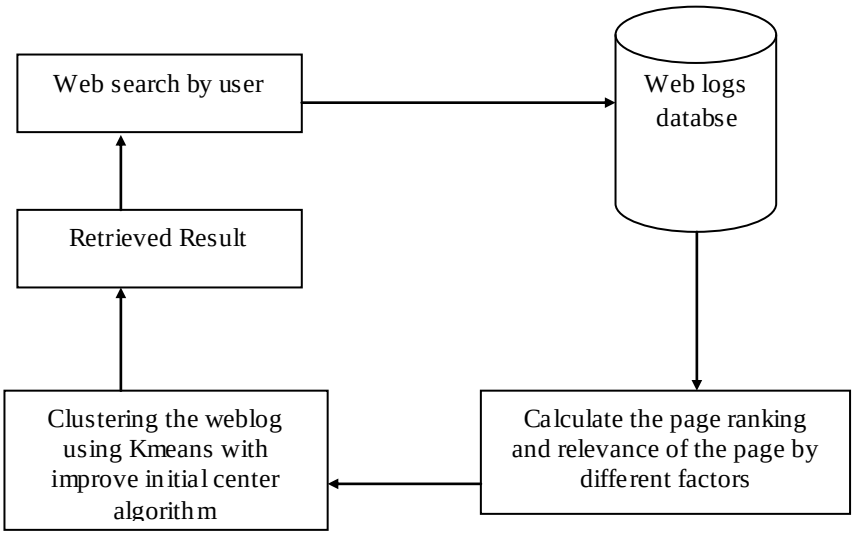


Fig 6.1 proposed work flow

## V. CONCLUSION

Hence as we are going to consider the both the factors of revival of efficient and accurate result of the search query by clustering and page ranking algorithms , the output will be more efficient than alone both of the algorithms .The proposed work there is less computation and cost , so the result will be fast and most relevant.

## REFERENCES

[1] Neeraj Raheja and V.K.Katiyar ," Efficient Web Data Extraction Using Clusteringapproach In Web Usage Mining", IJCSI International Journal of Computer Science Issues ,January 2014

[2] Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa," Enhancing K-Means Clustering Algorithm Withimproved Initial Center", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010..

[3] "Data Mining: Concepts and Techniques", Second Edition ,Jiawei Han and Micheline

[4] Jeeva Jose and P. Sojan Lal(2013)" Extracting Extended Web Logs to Identify the Origin of Visits and Search Keywords ", Intelligent Informatics Advances in Intelligent Systems and Computing Volume 182, pp 435-441.

[5] Yinghui Yang and Balaji Padmanabhan. (2005)." *GHIC: A Hierarchical Pattern-Based Clustering Algorithm for Grouping Web*" Transactions. IEEE Transactions on Knowledge and Data Engineering, Vol 17, No. 9.

[6] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009),

[7] Vol 1, July 2009, London, UK.Zhang Huiying, Liang Wei.An (2004). "Intelligent Algorithm of Data Pre-processing in Web Usage Mining". In Proceeding of the 5th World Congress on Intelligent Control and Automation. pp. 15-19. Hangzhou, P.R. China.

[8] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced k-means clustering algorithm," journal of Zhejiang University, 10(7): 16261633, 2006.

[9] Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[10] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.

[11] Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for Centroids initialization for k-means," department of IT and Electrical Engineering Politechnique in Surabaya, Faculty of Sciece and Engineering, Saga University, Vol. 36, 2007.

[12] A. Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," bioinformatics, Vol. 24, pp. 1359-1366, 2008.

[13] Balamurugan C ,Munibalaji T (2012), "Analysis of Link Algorithms for Web Mining", International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 2(2014).

[14] Wenpu Xing and Ali Ghorbani, "Weighted Pagerank Algorithm", In an Annual Conference on Communication Networks & Services Research, 2004.

[15] Bhatia Tamanna (2011), "Link analysis algorithm for web mining", IJCSE(0976-8491).

 