International Journal of Advanced Research and Development
ISSN: 2455-4030
Impact Factor: RJIF 5.24
www.advancedjournal.com
Volume 2; Issue 6; November 2017; Page No. 398-403

Efficient web mining using clustering techniques for better page searching by enhancing web log data

[1] Anshu Agarwal, [2] Dr. Akash Saxena, [3] Alex Patel
[1] PhD Scholar, Rai University, Ahmedabad, Gujarat, India
[2] PhD Guide, Rai University, Ahmedabad, Gujarat, India
[3] Research Scholar, Rai University, Ahmedabad, Gujarat, India

Abstract
Web logs are the best repository of data that can be extracted, managed for searching and understanding usage pattern of customers, websites visited in a session, its intensity etc. Log file is a huge source of such data stored in an unformatted manner that can be molded as needed and filtered how required. Needed data can be mined using data mining techniques and then classify it to various groups based on similarity (homogenous groups) or clusters. The classified data then can be used to find out the desired result or make study accordingly. The purpose of this study is to make efficient searching of web pages by clustering techniques and extracting and reformatting data derived from web log to enhance the data extracted using hash key and value structure between the web log data and database storage before applying any clustering techniques applied. Once clustering had been done the desired result will be used to form query string based on search criteria and result will be displayed to the user.

Keywords: weblog, mining, clustering, hash keys

1. Introduction
As the use of internet over time is increasing WWW has huge data related to web site access and usage. The same is maintained in different files like server weblog files, proxy server log files, browser log files etc. These files contains all information regarding the generator of request like session, access time, url, browser, type of request along with the basic information of generator of websites, errors generated, response time etc.
The same data can be used for marking access pattern of users, most visited websites, next visited web sites, preferences, buying patterns, searching patterns and also it can be used for ranking the web pages. This information is extracted from log files using data mining techniques commonly called as web log mining.

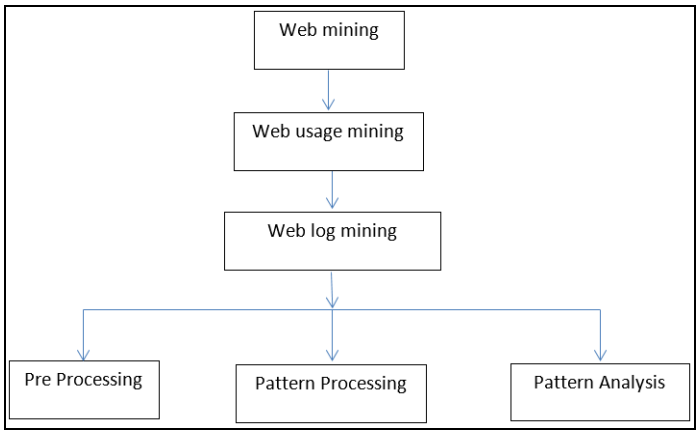

Fig 1

1.1 Web mining
It is technique of extracting data from web by using data mining techniques. It deals with extraction of data from World Wide Web.

1.2 Web Usage Mining
It is used to analyze the data as a result of transaction (request sent and received on web). This data is stored in web log files because of which it leads to web log mining.

1.3 Web Log Mining
It is used to extract data from log files created during transactions. This data is further used to create patterns, analyze the patterns to find interesting collection of data all linked together.
Web log mining is a technique which allows user to extract

data from log files, process and filter it, cluster and generate the required result. Web log mining is a subset of web mining. It includes various sources like web server logs, application server logs. It is done for the collection of information for web page access, which are stored in log files and also CGI scripts allow us to get the data using referrer log, survey logs etc. log files may look like and contain the following details.

Date	Time	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	cs-username	c-ip	Cs (User-Agent)	sc-status	sc-substatus	sc-win32-status	time-taken
2013-11-26	13:59:06	191.160.1.20	GET	/app/sheet.aspx	-	80	-	192.168.3.27	8Mozilla/5.0+(Windows+NT+6.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/31.0.1650.57+Safari/537.36	200	0	0	265
2013-11-26	13:59:06	191.160.1.20	GET	/app/images/download.gif	-	80	-	192.168.3.27	Mozilla/5.0+(Windows+NT+6.1)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/31.0.1650.57+Safari/537.36	200	0	0	198

UserIP/ClientIP: it results in IP address of the client that helps us to understand to generator of the request or the source from where the request of a particular website occurred.
Seesion: there is no direct record of session created but we can get the details by comparing and collecting data from the attributes, time, time taken, date and ClientIP. This will help us to gain session time in which a particular user was active on a particular page.
ServerIP: These help we to understand the destination to which request was sent.
URL: this gives detail of web site accessed.
Method: it tells us whether we requested the page by GET method or it was a response from websites using POST method.

1.3.1 Pre Processing
It is the first phase of web mining where raw data is extracted from web logs and unwanted data and wanted data are identified and accordingly they are filtered. It includes removal of bots from log, images and videos from it along with incomplete page request/response or broken links if any. It is actually removal of noisy data. It converts data into wanted format [1, 2].

1.3.2 Pattern Processing
Pattern processing is a technique that applied data mining technique once preprocessing phase is over. It works on cleansed and categorized data. It may have frequent pattern mining, clustering or association rule mining applied to it. In clustering two types of clusters can be applied one for clustering user data and another clustering web page data that can be analyzed for interesting patterns [3].

1.3.3 Pattern Analysis
It is analyzing the pattern identified. There are various methods for it.

1.3.3.1 Sequential
That identifies or discovers similar patterns. Example in e-commerce website mining one can identify the purchase pattern of a customer or most frequently bought product by customers and so on.

1.3.3.2 Association
It is used to find the connectivity or relation between different item sets. For example, products which are frequently bought together, products you bought last or ones which you searched with other products or other people purchase with some combination of buying patterns.
Web logs can be extracted from various sources typically servers which have been listed below.

1.4 Some sources of web log files or such similar files
These files are automatically created and maintained by the servers. They also have a fixed standard format that is maintained by the w3c along with other formats available too. Common log files are an example [4].

1.4.1 Web Servers
These maintain a common log format in the web servers of the hosted websites, or log files of local websites or world wide websites, etc. the problem here is again people would not like share their logs specially for the public domain websites or of the organizations that maintains confidentiality.

1.4.2 Proxy Servers
These are the servers similar to web servers but they are available as a substitute server whenever DNS server is not available either due to blocking of the websites or large network congestion traffic. These servers also maintain huge amount of data stored in log files.

1.4.3 Fire wall Servers
It stores access data for the users applied for security at organization levels to handle to and fro of data. It also contains users ip, target url, session active time etc.

1.4.4 Client browsers
The data in this collected form users browsers data stored in cache or java scripts or any other client side scripting. Problem in this is on clients machine JavaScript should be enabled, cache and cookies not cleared.

2. Related work
They have logged data from weblog to a database and then

developed an interface to retrieve data from the database. They have also defined data ware architecture for clustering the data. Ad hoc analysis has been done to generate the result [5].

It can have a transactional d/b that first populates database and then the fact table. No check for unauthorized access of data from weblog. Their main parameter was time from session log which can be extended in combination of other parameters.

The authors have used web log mining to find out the information regarding DOS attacks, and other threats. They have also proposed an approach to bring together different log files and identify the usage patterns and implemented decision rules for the data filtration [6].

Although they have used web mining for various purposes to improve working in an organization, yet no security measures have been defined how to block the suspicious files from accessing the web, and how these mining can be used for security of an organization. It is for an organization only so we can also extend work for larger organizations.

Authors talks about importance of preprocessing the log files before actually taking it for clustering and analysis. It talks about elimination of local and global noise, videos, graphics, failed http status code, and robot cleaning [7].

In this paper the authors have focused to improve website by identifying the reasons for failure access to the websites. Which are as follows, 403 forbidden, 404 not found and 503 – is out of resources. They have used weblog expert program to analyze and filter the data. Main focus is on errors and types of errors. Not used the data mining techniques like sequential, association clustering etc. instead used weblog expert program. They have considered data from only one source.

The author have used frequency count to predict the next page a visitor will be surfing based on data retrieved from web logs. They have collected data based on session of a user. Cluster had been formed for unique and repeating URLs. Mathematical proof had been given to show the reliability of proposed work. Hash had been used to store the data of websites surfed. Only session object is used to predict the next Url, it could use Data Structures and other fields from web log to make the prediction more efficient. It could be used for web page recommendation also.

The author paper talks about classification of webpages and their efficient searching by using IP address, url shortening, domains and multi sub domains are separated by (-) and etc. three different JRR techniques have been compared namely Random forest, Random Tree and J.48, among which Random Forest has the best result than other techniques [10].

Web usage pattern of users is identified by accessing and using weblog files of organization and analyzed using weblog expert tools. It works by extracting the set of visitors and their IP address, compared. It is used to provide a better navigation and website design. It talks about educational domain.

The domain can be enlarged and techniques for collecting and filtering data had not been specified [11].

It is to help the users and the developers to take care of web page surfing of their interest. They have proposed algorithms to data preprocessing and preparation. Also they have used modified GSP algorithm and modified Prefix Span algorithm [12].

The authors have discussed existing k-means clustering techniques and its drawbacks and have proposed changes in its working for efficient mining [13].

Dynamic pattern cluster to be created and we can find a way to actually manage unmanaged web log file.

3. Proposed work

The purpose of this work is to enhance web page searching or we can say it aims at improvement towards better search result using the data that had been accessed earlier.

The other way of doing it is use of static cache but it requires huge amount of space for cache to be maintained, although Ida Mele had proposed a graph based approach to optimize the space utilization that allows recommendation of news articles on surfing any other relevant data. Other studies had been done by analyzing query log data by [14].

This proposed work aims at modifying the data structure extracted from the web log before input it into the data base. We aim at introducing a hash key and value structure that will work like an indexing kind of concept. What we are doing is before directly entering the data into the database we define the main attributes or the main category under which the data can be sorted. Based on this sub category is formed which is again used for forming clusters to retrieve the desired result.

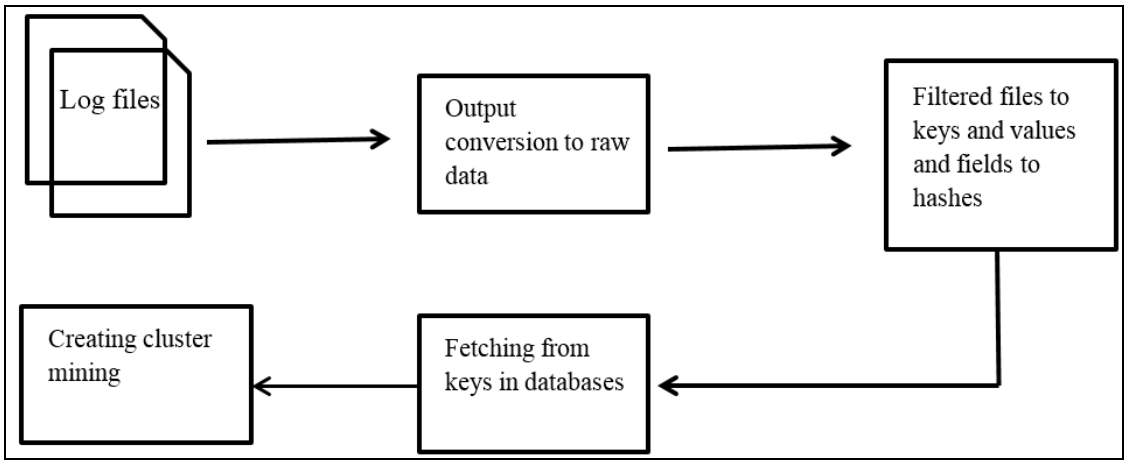

Fig 2

Elements from the filtered files are processed and identified, keys are identified to categorize the data from log file are stored in below format using database and hash structure.

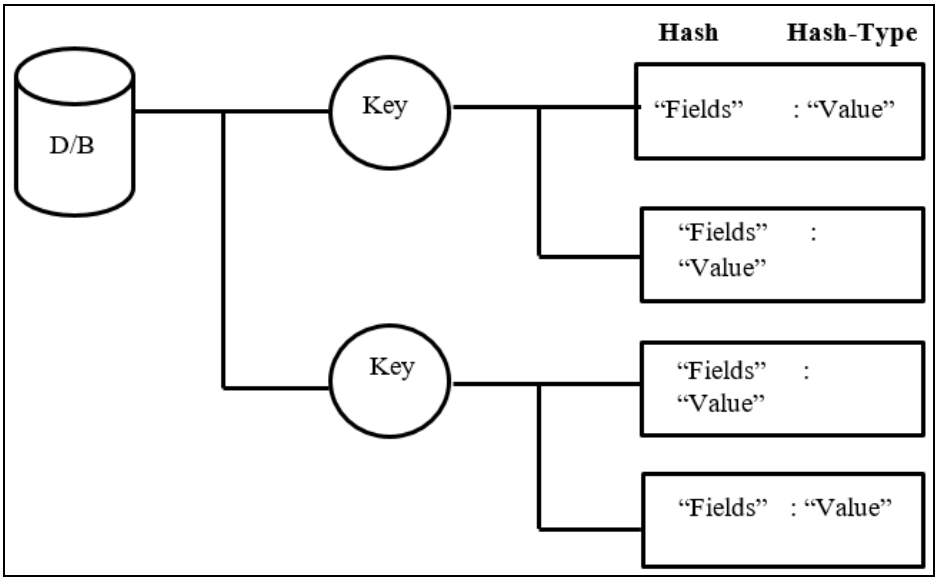

Fig 3

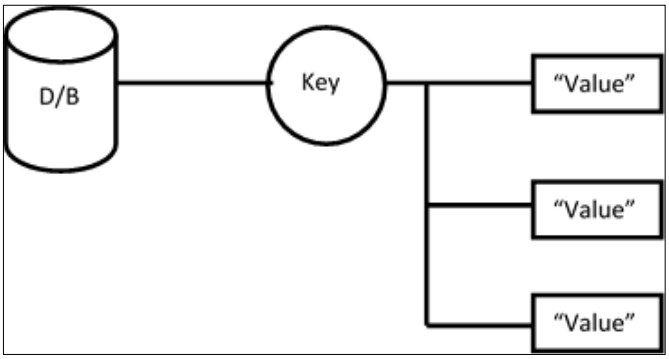

Fig 4

In fig.3 the parent structure is defined that stores the main attributes which define our data. It has a key that is used to identify the key words for data storage which again has hash and hash type that defines the child values. It can be of any data type supported by hash data type. Hash is data structure that stores data in key and value format which implements an associative array data type. It can be user defined data type. It uses hash function which is used to map value for the keys. Hash function acts as index for values. This structure will store immense amount of data which can be again clubbed based on fig.4

In fig.4 we are storing the main key which will In turn refer to the keys in fig.3. It will have block of keys of similar interest from the above figure. It forms a block of commonly needed keys from fig.4. It is used as either indexing or as an identifier. Indexing will be used to store an id which in turn will have collection of sub keys based on similar interest and identifier as a static and predefined value stored based on which we want to bring the data collectively. Say for example in case of firewall we define MAC as an identifier and its sub nodes can be other common attributes like IP Address, Session ID, URL visited, Username, etc.

With the explanation of both the figure we can say that the structure we obtain to store data and retrieval of data from this will be more efficient and faster in terms of query execution time and response time. This data then will be sent to a database where we can apply the mining techniques. Once clusters are formed the same result can be used to form URLs from the database data sets. We here propose that url retrieval or formation will be quick as compared to retrieve the same data from a normal database and hence forth the page results will be quicker.

4. Clustering

Clustering is used to refine data from a huge collection of data based on similar group and in a certain manner. It is finding similarities in data according to its characteristics or can be called as special type of classification which is not predefined. Clustering is similar to data base segmentation.

Sample Table

Table 1

Id	Name	Gender	Marital status	Qualification	Age
1	Ahmed	Male	Married	PG	27
2	Ashima	Female	Married	12^{th}	19
3	Meera	Female	Single	UG	21
4	Sumit	Male	Single	PG	25

From the above table we can try to study the age group which is married and single, we can also find that does qualification effects marriage effect? And many more such results can be predicted from such classified data.

Clustering is done on different attributes not necessarily on one attribute only. it can be used in various domains and for

various applications to find some pattern and classifications like use of e-learning and user preferences and search results for the education sector, can be used in e-commerce or m-commerce for getting most bought products or visited web sites etc. clustering includes set of key words and phrases, author of the document, creation date, size etc. the similar patterns can be found using cosine similarity.
Similarity for 2 documents d1, d2 can be given as

Similarity $(d1, d2) = \frac{d1.d2}{|d1||d2|}$

4.1 Clustering techniques used for web usage mining

The author talks about enhanced k-means algorithms, which follows two steps, First phase divides data set into two sub sets and initializes cluster points and then clusters are identified using k-means clustering technique [1].
Author have discussed about improved k means algorithm that minimizes the total of squares of the gaps [15].
There are many other native clustering techniques used for web usage mining but here we consider the ones which are frequently used as discussed below.

4.1.1 k-means algorithm

This method classifies data set into certain number of clusters, k clusters in priori. Every cluster has a central value called as centroid. It is advised to keep these centroids far from each other as different locations results different values. In next step we take point for each set and associate it to the nearest centroid. Then we need to do this till all the possible centroid had been explored and data [14, 16].
Algorithm is as follows.

1) Form clusters and assign centroids K to it that forms the initial group of centroids.
2) Assign each group to the closest centroid based on their similarity or association.
3) Once done reassign the centroids that is formed because of relocation of it.
4) Repeat step 3 and 4 till no explored centroids exists.

4.1.2 Hierarchical clustering technique

This method deals with decomposition of data. It has different hierarchical clustering algorithms like agglomerative clustering that uses one point clusters and it merges more similar pattern clusters, divisive clustering that starts with a point that has all objects in it and goes down dividing till it forms individual object, chameleon hierarchical method uses dynamic modeling to find the similarity between cluster sets. It combines the clusters that are clustered if they have high relation and are close to each other. Some more clustering techniques under this category are CURE and BIRCH that is graph based and starts with a matrix. It captures the geometry shapes of clusters whereas later is used cluster large amount of number data by using micro and macro clustering at later stage [17, 4].

4.1.3 Fuzzy clustering techniques

It is a prototype based clustering. The data is grouped based on similarity for example users can be clustered for similar purchase or searching behavior or access to similar websites. It allows objects to belong to more than one group with some weight. The clusters are characterized by statistical parameters. (Means and Variance). Constraint are applied on relationships of neighboring clusters and their degree [17].
The degree of certainty in fuzzy logic is between 0 and 1. Fuzzy clusters can be can be formed as follows: data point X={x1,…….,xn} where each point xi is m dimensional point ie. Xi=(xi1,…..xin).
Collection of clusters can be represented as C1, C2,…Cn that is subset of X.
Fuzzy partitions are formed using the following:

1. All the weights for given point Xi sums up to 1.

$\sum$ wij=1, wij is degrees of relations.

2. Every cluster C that has non zero weight but weight<1
3. $0 < \sum_{i=1}^{n} wij < n$, wij <1, wij≠0.

Fuzzy c means algorithm

1. Create and select initial fuzzy partition
2. Repeat
3. Find the centroid using fuzzy pseudo partition
4. Once done re assign partition for all possible centroids.
5. Repeat step 3 and 4 until centroids don't change

4.1.4 Grid based clustering

This type of clustering is applied to very large dataset which is multi-dimensional. The denser region forms clusters. It deals not with centroids but value space around those centroids.

1. Partition the data space into finite number of cells.
2. Calculate the density for each cell.
3. Sort the cluster based on their density.
4. Identify centroids.
5. Traverse the neighborhood cells.

There are many other algorithms for grid based clustering: Sting, Clique, Wave Cluster, O Cluster etc [17].

5. Acknowledgements

I would like to thank my guide Dr. Akash Saxena, my husband Alex Patel, my colleagues and my IT team of the university who have helped me directly or indirectly to achieve my goals and motivate me at every phase.

6. References

1. Chitraa V, DAS Thanamani. An enhanced clustering techniques for web usage mining, international journal of engineering research and technology, 2012; 1(4):1-5.
2. Ivancy R, Vajk I. Frequent Pattern Mining in web log data, Acta Polytechnica Hungarica, 2006; 3(1):77-90.
3. Ivancsy R, Kaovacs F. Clustering techniques utilized in web uasage mining, in 5th WSEAS Int. COnference Artificial Intelligence, Knowledge Engineering and Databases, Madrid, 2006.
4. Gupta AN, Karndikar PA. A Review: Study of Various Clustering Techniques in Web Usage Mining, international journal of advanced research in computer and communication engineering, 2014; 3:5888-5890.
5. Joshi KP, Joshi A, Yelena Y. Warehousing and mining web logs.
6. siddiqui S, Qadri I. Mining log files for web analytics and usage patterns to improve web organisation, international

journal of advanced research in computer science and software engineering, 2014; 4(6):794-802.
7. Dhawan DS, Lathwal M. Study of Preprocessing Methods in Web Server Logs, International Journal of Advanced Research in Computer Science and Software Engineering. 2013; 3(5):430-433.
8. Tyagi NK, Solanki AK, Wadhwa M. Analysis of Server Log by Web Usage Mining for Website Improvement, Internationla journal of computer science, 2010; 7(4):17-21.
9. PG, CS, PS, Raghavendra. Predicting model for fetching web page based on the usage pattern, internationla journal of control theory and applications, 2017; 10(14).
10. Vara SD, Prasad, DKR Rao. A new approach for webpages classification using JRR rechnique, international journal of computational intelligence research, 2017; 13(3):463-472.
11. Dharmarajan K, Dorairangaswamy MA. Discovering User Pattern Analysis from Web Log Data using Weblog Expert, Indian Journal of Science ans Technology, 2016, 9.
12. Patil SS, Khandagale HP. Enhancing Web Navigation Usability using Web Usage Mining Techniques, internationla research journal of engineering and technology, 2016; 4(6):2828-2834.
13. Patel R, tiwari R. An Efficient Algorithm to Generate Dynamic User Pattern Cluster Using K-means CLustring Techniques, IJEDR, 2016; 4(4):606-607.
14. Zhu J, Jun Hong, John G, Hughes. PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation, ACM Transactions on Internet Technology, 2004; 4(2):185-208.
15. Padmaja S, Sheshasaayee DA. Clustering of user behavior based on web log data using improved k-means clustring algorithms, 2016; 8(1):305-310.
16. linHuaXu, HongLiu. Web User Clustering Analysis based on KMeans Algorithm, IEEE, 2010.
17. Kaur R, Kaur S. A Review: Techniques for CLustering of Web Usage Mining, International Journal of Science and Research, 2014; 3(5):1541-1545.