Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Log File Data Clustering Using K-Means and Decision Tree

Supinder Singh *,
*Student of M.Tech Computer Science,
Department of CSE,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Sukhpreet Kaur
Assistant Professor,
Department of CSE,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Abstract- Web mining is used to discover interest patterns which can be applied to many real world problems like improving web sites, better understanding the visitor's behavior, product recommendation etc. Our Paper is focused on web log file format, its type and location. These web log files records information of each user request. Advantage of log files - data is easily available to be analyzed. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log file by the method of K-Means Clustering. Log file used for debugging purpose. Preprocessing web log file is used in data mining techniques, also can be used in intrusion detection system as input to detect intrusion. Hence our approach lies on the efficient retrieval of information from web log file than before.This is resprested through the asscioated Results set which are discussed in last section of paper.

Keywords--

1. INTRODUCTION[1,2]-WEB USAGE MINING

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services (Etzioni,1996). Web mining is categorized into 3 types.

1. Content Mining (Examines the content of web pages as well as results of web Searching)
2. Structure Mining (Exploiting Hyperlink Structure)
3. Usage Mining (analyzing user web navigation)

Web usage mining is a process of picking up information from user how to use web sites. Web content mining is a process of picking up information from texts, images and other contents. Web structure mining is a process of picking up information from linkages of web pages[2]. These 3 approaches attempts to extract knowledge from Web generate some useful result from that knowledge and apply the result to certain real world problems. Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files.

II. WEB LOG FILES[3]- A **log file** is a recording of everything that goes in and out of a particular server. It is a concept much like the black box of an airplane that records everything going on with the plane in the event of a problem. The information is frequently recorded chronologically, and is located in the root directory, or occasionally in a secondary folder, depending on how it is set up with the server. The only person who has regular access to the log files of a server is the server administrator, and a log file is generally password protected, so that the server administrator has a record of everyone and everything that wants to look at the log files for a specific server. Servers are not the only system that use log files. Process control systems, as well as computer operating systems have logging subsystems that work exactly like a log file does. While these are more sophisticated than a simple log file, most times it is the same concept, where a log message is recorded in the file and saved until it is needed. Other forms of log filing use more sophisticated systems, some of which even analyze the logs before they are needed, but it all depends on where the log file is located. The point of a **log file** is to keep track of what is happening with the server. If something should malfunction within a complex system, there may be no other way of identifying the problem. Log files are also used to keep track of complex systems, so that when a problem does occur, it is easy to pinpoint and fix. Log files are also important to keeping track of applications that have little to no human interaction, such as server applications. There are times when log files are too difficult to read or make sense of, and it is then that log file analysis is necessary. Log file analysis is generally performed by some kind of computer program that makes the log file information more concise and readable format. Log files can also be used to correlate data between servers, and find common problems between different systems that might need one major solution to repair them all[3].

III. LOG FILE TYPES -Access Log, Agent Log, Error Log and Referrer Log.

IV.CLUSTERING WITH K MEANS[4]- Classification can be taken as supervised learning process, clustering is another mining technique similar to classification. However clustering is a unsupervised learning process. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters.[4] In classification which record belongs which class is predefined, while in clustering there is no predefined classes. In clustering, objects are grouped together based on their similarities. Similarities between objects are defined by similarity functions, usually similarities are quantitatively specified as distance or other measures by corresponding domain experts. For example, based on the expense, deposit and draw patterns of the customers, a bank can clustering the market into different groups of people[3,4]. For different groups of market, the bank can provide different kinds of loans for houses or cars with different budget plans. In this case the bank can provide a better service, and also make sure that all the loans can be reclaimed.

K MEANS WITH CLUSTERING- This nonhierarchical method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart.[5]

Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.[3]

2. Motivation

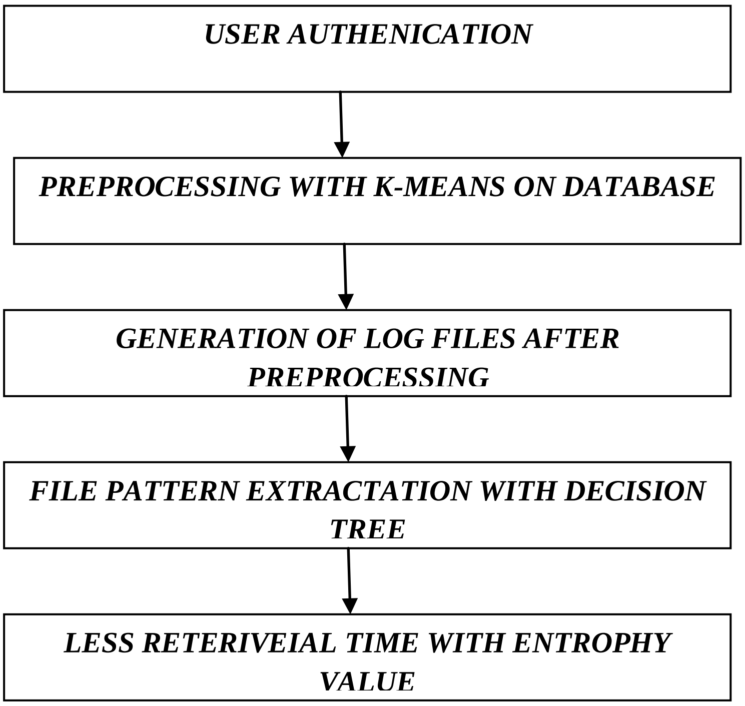
Last papers explain that growing trend among companies, organizations and individuals alike to gather information through web mining to utilize that information in their best interest. But it is a challenging task for them to fulfill the user needs and keep their attention in their website. Web usage mining has valuable uses to the marketing of businesses and a direct impact to the success of their promotional strategies and internet traffic. This information is gathered on a daily basis and continues to be analyzed consistently. Analysis of this pertinent information will help companies to develop promotions that are more effective, internet accessibility, inter-company communication and structure, and productive marketing skills through web usage mining. If we will be able to propose an efficient algorithm for the pattern extraction than it will help in the business of the website owners to understand their customer's behavior properly so that they can fulfill their requirements.[1,2]

3. Problem Statemnt with Purposed Work

Our problem is concerated about to improve the Web log content from web log files by efficient k means clustering methods.

The proposed work represents K-means clustering algorithm and its accuracy in clustering the data of web log file. The web log file contains the noise and un-usable contents like tags of html and attributes hence we need to cluster the scattered useful information from log file and extract it pattern wise for the good visibility of the user. In this we have done analysis with K-means clustering algorithm , one is the existing K-means clustering approach which gives the better results in clustering of data. As scattered (useful data) by this algorithm will be collected and clustered efficiently then displayed pattern wise.

4. Methodlogy of Purposed Work


USER AUTHENICATION
PREPROCESSING WITH K-MEANS ON DATABASE
GENERATION OF LOG FILES AFTER PREPROCESSING
FILE PATTERN EXTRACTATION WITH DECISION TREE
LESS RETERIVEIAL TIME WITH ENTROPHY VALUE

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 (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid.

A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

Decision Tree

This article is about decision trees in decision analysis. For the use of the term in machine learning, see Decision tree learning. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. c trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Decision Tree is a flow-chart like structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label (decision taken after computing all attributes). A path from root to leaf represents classification rules.

In decision analysis a decision tree and the closely related influence diagram is used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of 3 types of nodes:

1. Decision nodes - commonly represented by squares
2. Chance nodes - represented by circles
3. End nodes - represented by triangles

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

5. Results and Results Parameters

We have calculated our results set on different type web site data that are stored in our database as log files after pattern extraction with decision tree and note down the entrophy value and compare our result set on basis of k means and our purposed method on 5 numbers of clusters and displayed the comparsion on graph paper.

Entropy -Entropy is the sum of the probability of each label times the log probability of that same label.

Calculated Entropy of Work : 0.15

Execution time: The time in which a single instruction is executed. It makes up the last half of the instruction cycle.

Total Execution time: Total Execution time means the time that will take by the one task for completion. Like from the starting, when the task entered for execution and till the end when the result is display on the screen. This time from starting to end is called total execution time.

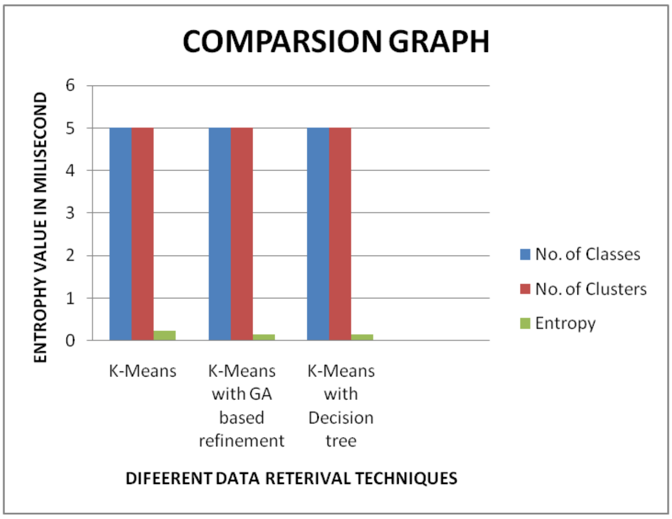
Execution Time : 78 milliseconds

Total Execution Time: 7410 milliseconds

Comparison

The following table presents the results:-

	K-Means	**K-Means with GA based refinement**	**K-Means with Decision tree**
No. of Classes	5	5	5
No. of Clusters	5	5	5
Entropy	0.2373	0.1502	0.15



6. Conclusion and Future Scope

Our Work is focusing on web log file format, its type and location. These web log files records information of each user request. Advantage of log files - data is easily available to be analyzed. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log file by the method of K-Means Clustering. Log file used for debugging purpose. Data preprocessing is an important steps to filter and organize appropriate information before using to web mining algorithm. This paper present an efficient k means algorithm for field extraction and data cleaning. Preprocessing web log file is used in data mining techniques, also can be used in intrusion detection system as input to detect intrusion. Hence our approach lies on the efficient retrieval of information from web log file than before.Future work will be focusing on the research of more efficient clustering and noise removal techniques. Further the comparison of the proposed algorithm with other clustering algorithms on web-usage mining can also undergo under development to enhance the robustness of the Log file data detection.

References

[1] Sachin Pardeshi and Ujwala Patil "*Central web mining services – public and free access log files*" Proceedings of National Conference on Emerging Trends in Computer Technology (NCETCT-2012) Held at R.C.Patel Institute of Technology, Shirpur, Dist. Dhule, Maharashtra,India. April 21, 2012

[2] Priyanka Patil and Ujwala Patil "*Preprocessing of web server log file for web mining*" Proceedings of National Conference on Emerging Trends in Computer Technology (NCETCT-2012) Held at R.C.Patel Institute of Technology, Shirpur, Dist. Dhule, Maharashtra,India. April 21, 2012

[3] Shiv Pratap Singh Kushwah, Keshav Rawat, Pradeep Gupta "*Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining*" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-3, August 2012

[4] Liu Kewen, "*Analysis of Preprocessing Methods for Web Usage Data",* 1ntemational Conference on Measurement, Information and Control (MIC),2012

[5] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi. *"The Survey of Data Mining Applications and Feature Scope"*, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[6] Amit Sharma, S.N. Panda and Ashu Gupta, "*Data Mining Techniques and their role in Intrusion Detection Systems*", J. Acad. Indus. Res. Vol. 1(4) September 2012

[7] K. Gloswan & Zorawan "*Hallucination depends deep web statistics withdrawal for web document clustering*" Global Journal of Computer Science and Technology, Volume 12 issue 5 version 1.0 March 2012.

[8] Sonia Sharma, ShikhaRai, "*Genetic K-Means Algorithm – Implementation and Analysis*" International Journal of Recent Technology and Engineering (IJRTE)ISSN: 2277-3878, Volume-1, Issue-2, June 2012.

[9] S.K.Pani, L.Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Pandhi "*Web Usage Mining: A Survey on Pattern Extraction from Web Logs*" International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011

[10] Mrs. G. Sudhamathy, Dr. C. Jothi Venkateswaran, "*Web Log Clustering Approaches – A Survey",* G. Sudhamathy et al. International Journal on Computer Science and Engineering (IJCSE) 7 July 2011

[11] Theint Theint Aye, "*Web Log Cleaning for Mining of Web Usage Patterns*" **,**IEEE 2011

[12] Raval.H.yen" *International Journal of Forward Research in Software Engineering*" Volume 2, Issue 10, October 2011

[13] C.Rekha, N.Sujatha, K.Iyakutti, "*Algorithm to Improve the Cluster Quality using Genetic Algorithm*" Technical Journals Vol 02, Issue 04; July-September 2011

[14] Gary M. Weiss, Brian D. Davison, "*Data Mining*" To appear in the Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, 2010