Improved Approach to Predict user Future Sessions using Classification and Clustering

Akshay Kansara[1], **Swati Patel**[2]

[1]Gujarat Technological University, L.D. College of Engineering,
Ahmedabad 38009, Gujarat, India

[2]Department of Computer Engineering, Gujarat technological University,
L.D. College of Engineering, Ahmedabad 38009, Gujarat, India

Abstract: ***With the huge quantity of information which is available on internet that makes it challenging for providers to make such relevant information available to users in a fast and personalized manner. One way to tackle with this challenging issue is to use a recommendation technique, which can make visitors to discover further offerings. Web usage mining is the key process of extracting knowledge of user access pattern from web servers. This paper presents the combination of the classification and clustering techniques to predict user future movements.***

Keywords: Web usage mining, Web log, Classification, Clustering

1. Introduction

When we relate data mining to the web data then it is known as web mining. Web mining gains its importance with the increasing amount of web information that is becoming much larger day by day. It focuses on the web page link structure, their content and their usage. Web mining is the process of discovering useful information from the web data. The web mining field encompasses a wide array of issues, primarily aimed at delivering actionable knowledge from the web, and includes researchers from information retrieval database technologies, and artificial intelligence. Web mining can be defined by two ways. The very first, known as Web Content mining which is related to the procedure of information discovery from resources across the WWW. Second one, known as web usage mining, the process of mining user browsing access patterns.Web mining is the significance of data mining techniques to mine knowledge from web data, i.e. web content, web structure, and web usage data mining. Web content mining extracts or mines useful information or knowledge's from web page contents. WSM aims to discover useful knowledge from hyperlinks, which represent the structure of the Web. Hyperlink is a link that exists in a web page and refers to another region in the same web page or another web page. Finally, web usage mining, also acknowledged as Web Log Mining, aims to capture and model behavioral patterns and profiles of users who interact with a web site.

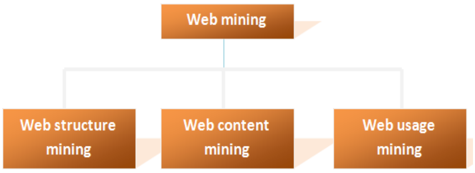


Figure 1: Taxonomy of Web Mining

2. Fundamental of WUM

WUM be able to be described as the finding along with study of user access patterns, through the log files and related data from the Web site. It defines what users are searching on internet. WUM applies data mining procedures to analyze user access of web sites. As with any KDD (knowledge discovery and data mining) process, it contains three main steps:

- Pre-processing
- Pattern discovery
- Pattern analysis

a. Preprocessing

It is generally used as groundwork of data mining practice. The preprocessing task within the WUM process involves cleaning and structuring data to prepare it for the pattern discovery task. It can be classified into three parts: Usage preprocessing, Content preprocessing and Structure preprocessing.

b. Pattern Discovery

In this, WUM can be able to unearth patterns in server logs and carried out only on samples of data. Interpretation and evaluation of results be done on samples of data. The various pattern discovery methods are; Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns, and Dependency Modelling.

c. Pattern Analysis

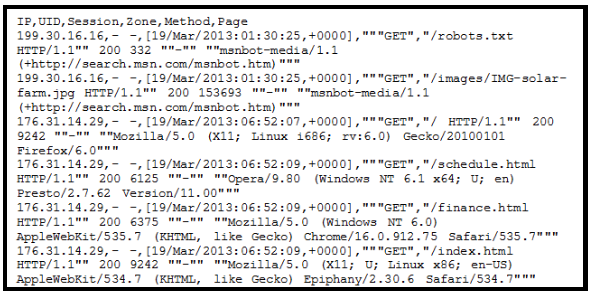
The requirement behind analysis is to sort out irrelevant data or patterns from the data sets which are found in the pattern discovery stage. Most frequent type of analysis contains a query mechanism like SQL etc. Both content as well as structure information could be used to filter patterns containing pages that match a certain hyperlink structure.

3. Web Log

Web data can be stored in log file. Once web data is obtained, it might be combined with other databases, in which the techniques are implemented. Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent.[5] By analyzing these log files gives a neat idea about the user. A Web log is a file to which the Web server writes information each time a user requests a website from that particular server. A log file can be located in three different places:

• Web Servers
• Web proxy Servers
• Client browsers

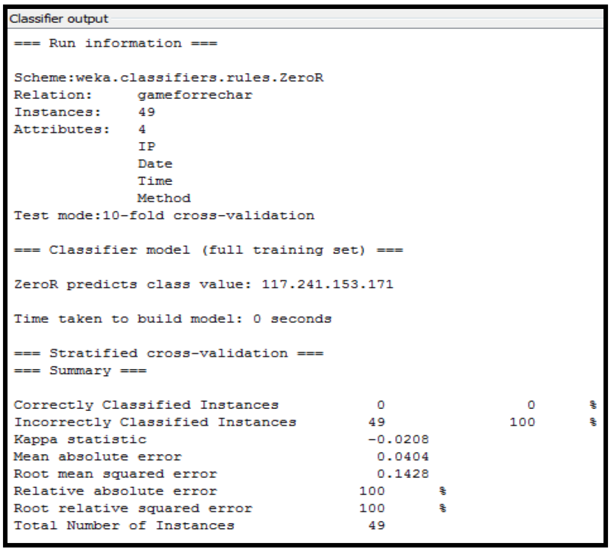
The most popular log file formats are the Common Log Format (CLF) and the extended CLF. [6][11] A common log format file is created by the web server to keep track of the requests that occur on a web site. A standard log file has the following format [12] [13].



```
IP,UID,Session,Zone,Method,Page
199.30.16.16,- -,[19/Mar/2013:01:30:25,+0000],"""GET","/robots.txt
HTTP/1.1"" 200 332 ""-"" ""msnbot-media/1.1
(+http://search.msn.com/msnbot.htm)"""
199.30.16.16,- -,[19/Mar/2013:01:30:25,+0000],"""GET","/images/IMG-solar-
farm.jpg HTTP/1.1"" 200 153693 ""-"" ""msnbot-media/1.1
(+http://search.msn.com/msnbot.htm)"""
176.31.14.29,- -,[19/Mar/2013:06:52:07,+0000],"""GET","/ HTTP/1.1"" 200
9242 ""-"" ""Mozilla/5.0 (X11; Linux i686; rv:6.0) Gecko/20100101
Firefox/6.0"""
176.31.14.29,- -,[19/Mar/2013:06:52:09,+0000],"""GET","/schedule.html
HTTP/1.1"" 200 6125 ""-"" ""Opera/9.80 (Windows NT 6.1 x64; U; en)
Presto/2.7.62 Version/11.00"""
176.31.14.29,- -,[19/Mar/2013:06:52:09,+0000],"""GET","/finance.html
HTTP/1.1"" 200 6375 ""-"" ""Mozilla/5.0 (Windows NT 6.0)
AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.75 Safari/535.7"""
176.31.14.29,- -,[19/Mar/2013:06:52:09,+0000],"""GET","/index.html
HTTP/1.1"" 200 9242 ""-"" ""Mozilla/5.0 (X11; U; Linux x86; en-US)
AppleWebKit/534.7 (KHTML, like Gecko) Epiphany/2.30.6 Safari/534.7"""
```

4. Classification

The main objective of Predicting User navigation patterns using Clustering and Classification from web log data is to predict user navigation patterns using knowledge from the classification process that identifies potential users from web log data and a clustering process that groups potential users with similar interest and using the results of classification and clustering, predict future user requests.



```
=== Run information ===

Scheme:weka.classifiers.rules.ZeroR
Relation:     gameforrechar
Instances:    49
Attributes:   4
              IP
              Date
              Time
              Method
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: 117.241.153.171

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          0                 0      %
Incorrectly Classified Instances       49               100      %
Kappa statistic                        -0.0208
Mean absolute error                     0.0404
Root mean squared error                 0.1428
Relative absolute error               100      %
Root relative squared error           100      %
Total Number of Instances              49
```

5. Clustering

Clustering aims at dividing the data set into groups where the inter-cluster similarities are minimized while the similarities within each cluster are maximized. Clustering Web sessions can be achieved through page clustering or user clustering. Web page clustering is performed by grouping pages having similar content. Page clustering can be simple if the Web site is structured hierarchically. In this case, clustering is obtained by choosing a higher level of the tree structure of the Web site.

Central clustering algorithms [4] are often more efficient than similarity-based clustering algorithms. We choose centroid -based clustering over similarity-based clustering. We could not efficiently get a desired number of clusters, e.g., 100 as set by users. Similarity-based algorithms usually have a complexity of at least O (N2) (for computing the data pair wise proximity measures), where N is the number of data instances. Web sessions are first identified and grouped according to functionality and using meaningful features. Then, the Web sessions are grouped into a number of categories. *K*-means clustering algorithm is based on Web session categories identified and is carried out according to some distance metrics.

A. Simple k-means

The K-Means algorithm is one of the partitioning clustering algorithms. It is based on distance, unconfirmed and partition based. K- Means clustering algorithm is the simplest and most commonly used clustering algorithm, especially with large data sets. It involves following steps:

- Define a set of data sets
- Define total number of clusters (*k*).
- Arbitrary allocate a number of items to each cluster.

The objective function is

$$\min_{\mu_1 \ldots\ldots \mu_k} \sum_{h=1} \sum \left\{ \left| x - \mu_h \right|^2 \right\} \text{where } x \text{ € } x_h$$

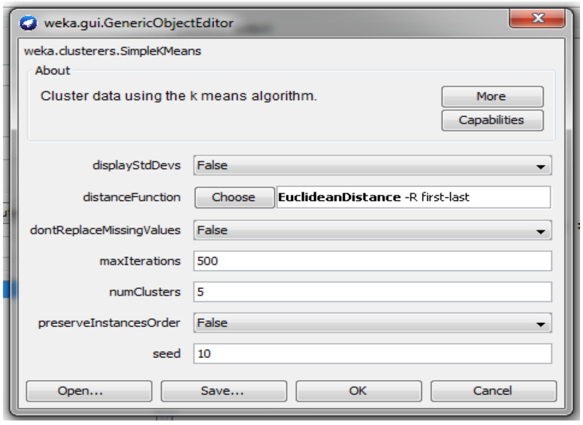
The complexity of this algorithm is O (n^{kd+1} log n).
Here k and d are fixed where n are no. of clustered. This algorithm repeatedly performs the following steps:

1. Compute mean vector from all items in each cluster.
2. Relocate the items whose midpoint is nearby.

This algorithm creates first random cluster for that it runs different times and each time it starts from a different point and computes different results. All the clusters are compared using the distances within clusters and the least sum of distances is considered. as a result, *k*-means algorithm deals with the total number of clusters (*k*), the total number of runs and the distance measured. Output defines total number of clusters with actual number of items in each cluster. Distance measured between all data sets in each cluster plays a crucial role in the clusters. Here, we used Euclidean distance:

$$\text{Euclidean}(x,y) = \sqrt{\sum (x_i - y_i)^2}$$

It defines the actual arithmetical distance.



weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

displayStdDevs: False

distanceFunction: Choose **EuclideanDistance** -R first-last

dontReplaceMissingValues: False

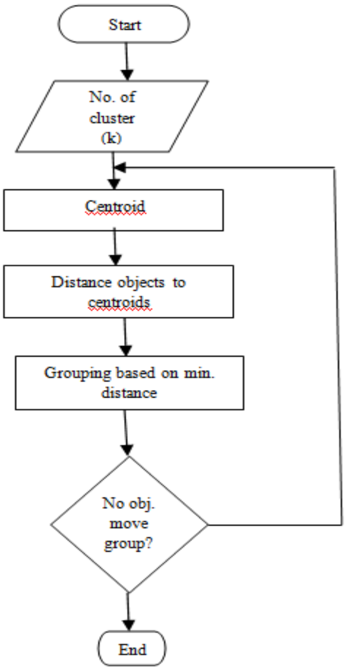
maxIterations: 500

numClusters: 5

preserveInstancesOrder: False

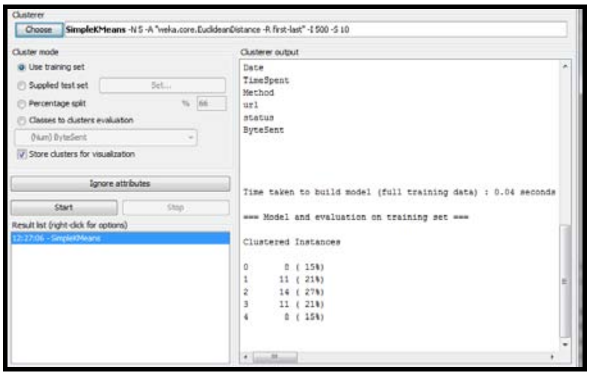
seed: 10

This paper presents combining approach of classification and clustering using the weka tool. We defined k=5(number of clusters). Basically, K-means algorithm is totally related to how clusters are primarily created. It is always needed to use different values or parameters and calculate the results.


Start
No. of cluster (k)
Centroid
Distance objects to centroids
Grouping based on min. distance
No obj. move group?
End

K-means clustering process

The result demonstrates that each cluster as well as total number and fraction of instances allocated to different clusters. The midpoints of clusters are the mean factor for each cluster (so, each cluster value in the cluster represents the mean value). Therefore, centroids could be used to differentiate the clusters.



Clusterer

Choose **SimpleKMeans** -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

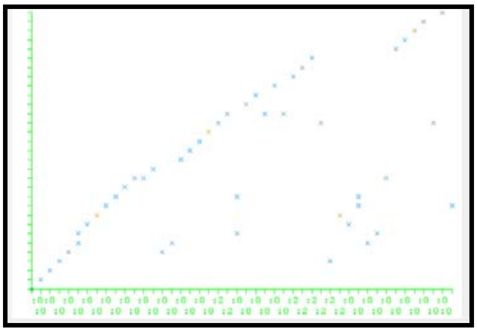
Cluster mode

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) ByteSent
☑ Store clusters for visualization

Clusterer output

```
Date
TimeSpent
Method
url
status
ByteSent


Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 15%)
1      11 ( 21%)
2      14 ( 27%)
3      11 ( 21%)
4       8 ( 15%)
```

Following graph visualize the clustered data.



Clustering results provide with various forms of knowledge extracted from the log data. These include number of visits made to a single webpage, webpage traffic, most frequently viewed page and navigation behavior of the users. The web log data contained 20 unique web pages which are assigned codes for clarity. The performance of the prediction engine was evaluated using three performance parameters, namely, accuracy, coverage and F1 Measure. The navigation patterns are identified from the clusters generated from the previous step and each pattern is divided into two sets. The first set is used for generating prediction and the second set is used to evaluate the predictions.

6. Conclusion

In this paper, a usage navigation pattern prediction system was presented. The system consists of four stages. The main objective of the proposed system is to predict user navigation patterns using knowledge from (i) a Classification process that identifies potential users from web log data and (ii) a clustering process that groups potential users with similar interest and (iii) Using the results of classification and clustering, predict future user requests. The result was then segmented to identify potential users. From the potential user, a clustering algorithm was used to discover the navigation pattern. The experimental results prove that the proposed amalgamation of techniques is efficient both in terms of clustering and classification. In future, the proposed work will be compared with existing systems to analyze its performance efficient. Plans in the direction of using association rules for prediction engine are also under consideration.

References

[1] Jiawei Han, Micheline Kamber, “Data mining concepts and techniques”, Elsevier Inc., Second Edition, San Francisco, 2006

[2] Sheetal A. Raiyani, Shailendra Jain, “*Enhance Preprocessing Technique Distinct User Identification using Web Log Usage data*”, International Journal of Computer Science & Communication Networks, Vol 2(4), 526-530

[3] Vijayashri Losarwar, Dr. Madhuri Joshi, “*Data Preprocessing in Web Usage Mining”*, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

[4] J. Vellingiri and S. Chenthur Pandian, “*A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification*”, Journal of Computer Science, pp. 683-689, 2011.

[5] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai, *“Analysis of Web Logs and Web User in Web Mining”*, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011

[6] Navin Kumar Tyagi, A. K. Solanki and Manoj Wadhwa, “*Analysis of Server Log by Web Usage Mining for Website Improvement”*, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010

[7] C.P. SUMATHI, R. PADMAJA VALLI, T. SANTHANAM, *“An Overview of Preprocessing of Web Log Files for Web Usage Mining”*, Journal of Theoretical and Applied Information Technology 15th December 2011. Vol. 34 No.1

[8] Navin Kumar Tyagi, A.K. Solanki and Sanjay Tyagi, "*An Algorithmic Approach To Data Preprocessing in Web Usage Mining*", International Journal of Information Technology and Knowledge Management, 2010, 279-283

[9] Claudia Elena DINUCĂ, *“An Application for Data Preprocessing and Models Extractions in Web Usage Mining”*, International Conference “Risk in Contemporary Economy”, ISSN 2067-0532 XIIth Edition, 2011

[10] Doru Tanasa and Brigitte Trousse, *“Advanced Data Preprocessing for Intersites Web Usage Mining”*, Published by the IEEE Computer Society, 1094-7167/04/$20.00 © 2004 IEEE

[11] Priyanka Patil, Ujwala Patil, “*Preprocessing of web server log file for web mining*”, World Journal of Science and Technology 2012, 2(3):14-18

[12] G. Castellano, A. M. Fanelli, M. A. Torsello, “*Log Data Preparation for Mining Web Usage Patterns*”, IADIS International Conference Applied Computing 2007

[13] Thanakorn Pamutha, Siriporn Chimphlee, Chom Kimpan1, and Parinya Sanguansat, “*Data Preprocessing on Web Server Log Files for Mining Users Access Patterns”*, International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, June 2012, ISSN: 2046-6447

[14] V.Chitraa, A.S.D., "*A Survey on Preprocessing Methods for Web Usage Data,*" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010

Author Profile

Akshay Kansara received the B. Tech. in Information Technology from Ganpat University, Gujarat in 2009 and pursuing M.E. degrees in Information Technology from L.D. College of Engineering, Gujarat Technological University, Ahmedabad, Gujarat.

Prof. Swati Patel M.E. in Computer Science and Engineering, Asst. Prof at L.D. College of engineering, Ahmedabad, Gujarat, India