Mining Access Patterns Using Clustering

Mrs. Kiruthika M
Asst. Professor,Computer Dept.
Fr.C.R.I.T,Vashi.

Mrs. Dipa Dixit
Lecturer, Computer Dept.
Fr.C.R.I.T, Vashi

ABSTRACT

Web usage mining is an application of data mining techniques to discover usage patterns from web data, in order to understand and better serve the needs of web based application. The aim of this paper is to discuss about a system proposed which would perform clustering of user sessions extracted from the web logs.HTML links are extracted from these web logs for each user which constitutes the dataset. Clustering is then performed on these datasets based on the key attributes to partition the users into several homogenous groups such that similar user access patterns belong to the same cluster. Implementation and the results are also discussed.

Keywords

Web Usage mining, Access Patterns, classification.

INTRODUCTION

Overview of Data Mining

Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. Data mining can be used to uncover patterns in data but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. Data mining cannot discover patterns that may be present in the larger body of data if those patterns are not present in the sample being "mined". Inability to find patterns may become a cause for some disputes between customers and service providers. Therefore data mining is not foolproof but may be useful if sufficiently representative data samples are collected. The discovery of a particular pattern in a particular set of data does not necessarily mean that a pattern is found elsewhere in the larger data from which that sample was drawn. An important part of the process is the verification and validation of patterns on other samples of data.

Problem Definition

Web is the single largest data source in the world. Today there are several billions of HTML documents, pictures and other multimedia files, as a result of which, discovery and analysis of useful information becomes a practical necessity. Due to heterogeneity and lack of structure of the World Wide Web, this discovery and analysis gets difficult. User visits the websites of his/her own interest. At such times it becomes feasible to create user profiles according to navigation path of the user. But creating user profiles is a problem as it involves complex processes as modeling and predicting a user’s access. Many types of software are readily available in the market, with increased size and complexity. Thus, more intelligent mining techniques are necessary. We would like to propose a system that:

a. Analyze access patterns of users by clustering user sessions extracted from web logs.
b. Partitions these users into several homogeneous groups with similar activities.
c. Extracts user profiles from each of these groups.

The aim of the system is to analyze access patterns of users by clustering user sessions extracted from web logs.Web logs can be server side (where in the access made by a user are recorded on the server) or client side (where the accesses are recorded on the client's workstation).This system has essential importance in fields like Pattern recognition, Spatial data analysis, Image processing, Market research, Document classification.

CLUSTERING CONCEPTS

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. For our system, we have chosen the data mining technique as Clustering.

Design Steps

Design steps for clustering are given below:

1. Collection of datasets

Web logs are log entries that contain details about the users logging on to a website. Web servers register a web log entry for every single access they get, in which important pieces of information about accessing are recorded, including the URL requested, the IP address from which the request originated, and a timestamp.

A log is a set of triples <u_i, p_i, t_i>
Where u_i : Set of users
p_i :Set of pages visited
t_i :Timestamp associated

2. Conversion into Database Tables

The weblog is converted from unstructured format (stream of data) into a structured format (tables).

3. Identifying Attributes for Clustering

After analyzing the record in the database, attributes which are relevant and which can give necessary information can be chosen.

4. Applying Clustering

Cluster or segment the database according to the attributes.

5.Visualization

Once the data is analysed, it needs to be abstracted in some form so that useful information is extracted. Main goal of data visualization is to communicate information clearly and effectively through graphical means.

IMPLEMENTATION

The above design has been proposed for a general dataset but the implementation of the above design has been done for a particular set of web logs.

Step 1:

Raw web logs, to be used as dataset, were downloaded from the web.(www.cs.depaul.edu)

Figure 1: Raw Web Log Records

Step 2:
Now these web logs were converted to database tables using the following step 1 shown below in Figure no 2.

a. Log into mySQL command line client.
b. Create a mySQL database.
c. Create table into DB with attributes and primary key. Import the log files into database.

Steps to import text files to database:
a. Upload log file('/logfile.txt') to local directory accessible by mySQL
b. Load this in file or LOCAL file into mySQL table by appropriately specifying the path and delimiters.
c. load data local infile '/logfile.txt' into table tablename fields terminated by ' ' lines terminated by '\n' (table schema)

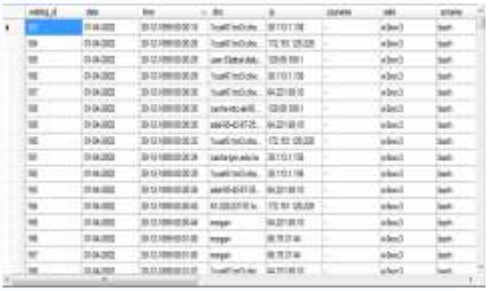


Figure 2: Web log converted into tabular format

Step 3:
After conversion, the attributes which can help extracting information from web logs are identified from the table. These attributes also help in identifying the user.

Step 4:
Clustering or segmentation of the database was carried out based on the identified clustering attributes.

RESULTS

For clustering two cases were considered i.e; based on IP address and timestamp

Case1: Clustering Based On IP Address

To perform clustering based on IP address certain steps were followed which are shown below in the form of results.

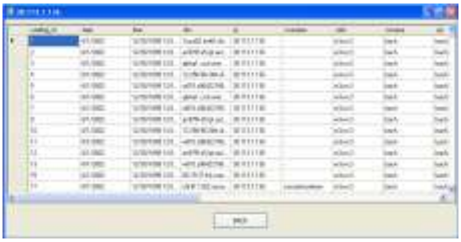
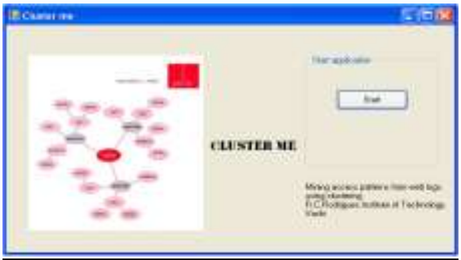


Figure 3 : Raw web log records

Step 1: Start screen:



Cluster me

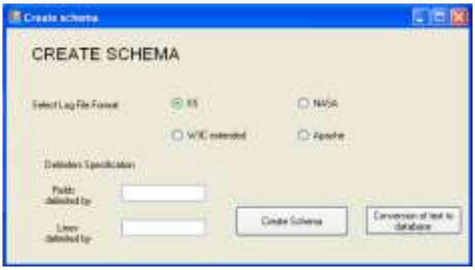
CLUSTER ME

Start application

Mining access patterns from web logs using clustering
Fr.C.Rodrigues Institute of Technology Vashi

Figure 4: start screen for clustering

Step 2: Schema creation:



Create schema

CREATE SCHEMA

Select Log File Format ● IIS ○ NASA ○ W3C extended ○ Apache

Delimiters Specification

Fields delimited by

Lines delimited by

Figure 5:IIS schema created for log records

Step 3: Selecting relevant attributes:

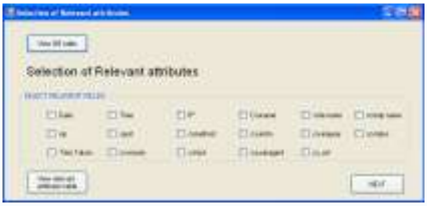
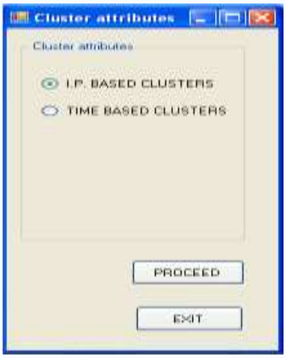


Figure 6 :Relevant attributes selected

Step 4: Database sample:

Figure 7:Relevant attributes selected shown in form of table

Step 5: Selection to view clusters based on which attributes:



Cluster attributes

Cluster attributes

● I.P. BASED CLUSTERS
○ TIME BASED CLUSTERS

Figure 8: Creation of IP based clusters

Step 6: IP based clusters with line graph to compare the result:

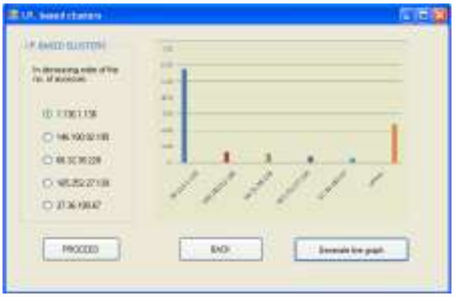
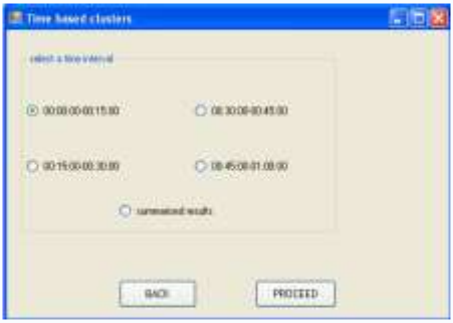


Figure 9:IP based clusters in form of graph

Case 2: Clustering based on Time stamp

To perform clustering based on IP address certain steps were followed which are shown below in the form of results.

Step 1: select particular timestamp for which cluster needs to be identified.



Time based clusters

select a time interval

● 00:00:00-00:15:00 ○ 00:30:00-00:45:00
○ 00:15:00-00:30:00 ○ 00:45:00-01:00:00
○ summarized results

Figure 10:Different time stamps from web log

Step 2: Sample database with clustering performed based on timestamp (date and time) is shown below

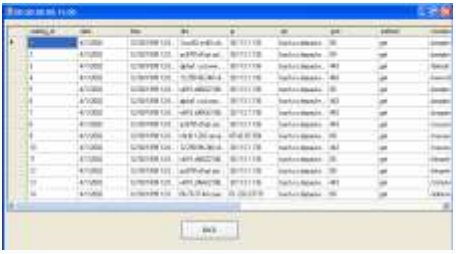


Figure 11 : Cluster for timestamp (00:00:00:00:15:00)

Step 3: Summarized result for time based clustering is shown in the form of pie chart for different time stamps present in web log records.

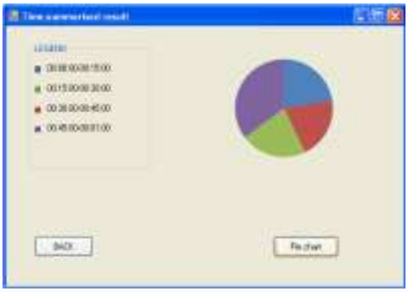


Figure 12 : Pie chart for clusters of time stamp

CONCLUSION

The above paper discusses the design steps and results applied to a dataset and how it was clustered into one of the homogenous groups with similar activities. Segmentation/ Clustering was done based on IP address and time in our system. Several other attributes can also be chosen for clustering. After segmentation has been done on the web logs based on a defining attributes, this can be utilized in variety of ways. If clustering is done on the basis of time, it can be used for the purpose of website modification and if it is done on the basis of IP addresses, it can be used to identify users with similar access patterns and can be used for the purpose of marketing.

REFERENCES

[1] Cooley, R. Mobasher, B. and Srivastave, J. “Web Mining: Information and Pattern Discovery on the World Wide Web” (1997) In Proceedings of the 9th IEEE ICTAI Conference, pp. 558-567, Newport Beach, CA, USA.

[2] I-Hsien Ting, Chris Kimble,Daniel Kudenko, “Applying Web Usage Mining Techniques to Discover Potential Browsing Problems of Users”, Department of Computer Science, the University of York, Seventh IEEE International Conference on Advanced Learning Technologies,2007.

[3] Margaret H Dunham, “Data mining introductory and advanced topics” 5th Ed.

[4] Jiawei Han and Micheline Kamber, “Data mining: concepts and techniques”.

[5] Jian Pei,Jiawi, HanBehzad Mortazavi-asl and Hua Zhu Simon Fraser ,“Mining access patterns efficiently from web logs” by University,Canada.

[6] Yan LIa,b, Boqin FENGa, Qinjiao MAOa, Xi' an Jiaotong, Shaanxi, “Research on Path Completion Technique in Web Usage Mining” University of Technology, China.