Web User Access Pattern Analysis on Proxy Log Data

A. Ann Sinthusha (*annsinthu21@gmail.com*) **& E. Y. A. Charles**
Department of Computer Science, Faculty of Science, University of Jaffna

Introduction

The aim of this study is to test data mining techniques on 'Big data'. In this regards this project was done using the proxy server log entries of an institution to identify user groups based on their Internet access patterns. The proxy log entries specify the detail of each request for a resource on the Internet and the outcome of their request. Big data is a term used to describe the exponential growth and availability of data, both structured and unstructured. These data sets are so large or complex that traditional data processing applications are inadequate.

Motivation

Extract meaningful data from large quantity of proxy log entries and study on applying data mining techniques

Big data analysis is an emerging field and many organisations are exploring its potential. This study would provide a better understanding of this area and would show ways to analyse the data collected from day to day activities of an organization.

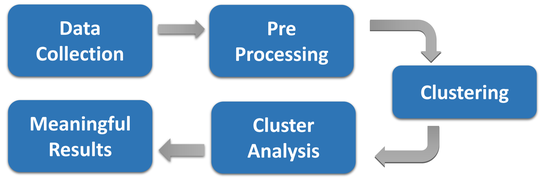
Objective

- ✓ Identify useful web usage patterns using proxy access logs and to enhance the performance of a proxy server to provide a better service to users
- ✓ Providing a case study for the application of Big data analytics and data mining techniques

Major issues

Proxy log files contains vast number of missing values and blank entries, making it difficult to read correctly. Several methods were tested to handle missing and erroneous values. As part of this research work several tools and methods were studied and utilised such as Apache Hadoop framework, Java libraries, H2O predictive analytic tool and Matlab. One major problem in data analysis is handling numerical and categorical data together. Several approaches to handle different data types were tested and k-means and k-mediods clustering techniques were applied to cluster the data.

Methodology



Data Collection

The data have been collected as raw log entries recorded by a squid server for a period of time.

Duration					No. of entries
25/May/2014	4am	-	29/May/2014	4pm	5,108,226
09/Feb/2014	4am	-	16/Feb/2014	4am	6,315,612
23/Mar/2014	4am	-	30/Mar/2014	4am	7,639,287
11/Jan/2015	1pm	-	12/Jan/2015	1pm	621,891
12/Jan/2015	1pm	-	18/Jan/2015	3am	4,177,279
27/Jan/2015	3am	-	28/Jan/2015	3am	7,219,048
28/Jan/2015	3am	-	1/Feb/2015	3am	11,299,938

A sample log entry

Client IP Address	192.168.33.21
Date and time	25/May/2014:04:02:13
GMT time	530.394
Response time	25172
Request status	TCP_MISS/200
Request method	GET
Reply size	304
Request URL	http://185.8.105.27/din.aspx?
Squid status	DIRECT/185.8.105.27
Username	dmse
Content type	application/octet-stream

Pre-Processing

Client IP

Before	After
192.168.33.21	VLAN33-21

According to the institutions LAN design, each client IP addresses are replaced with an easily understandable label – virtual LAN ID, to which that IP address belongs.

Date and Time

Before	After			
25/May/2014:04:02:13	25	May	2014	Slot-4

The selected date_time values are divided as date and time. Then from the date value date, month, year, and day information are extracted and the time value are converted as one hour time slot labels

Response Time

Before	After
25172	Above_AVG

Numerical values of Response Time are replaced with the label Above_AVG for the data points having the value above the average value and Below_AVG for the data points having the value below the average value

Reply Size

Before	After
304	Below_AVG

The numerical values of Reply Size are replaced with the label Above_AVG for the data points having the value above the average value and Below_AVG for the data points having the value below the average value

Request URL

Before	After
http://185.8.105.27/din.aspx?	185.8.105.27

The hostnames are extracted from each URLs.

Squid Status

Before	After
DIRECT/185.8.105.27	DIRECT

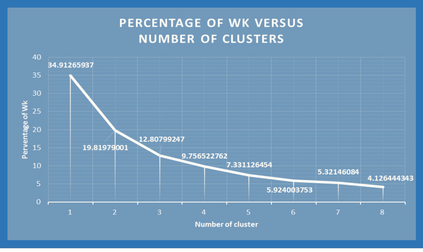
Squid status appeared as 'None' when it accessed from the squid and 'Direct/ <IP address of the site>' when it accessed from the hosted server location. From these data only the 'Direct' and 'None' have been considered as the attribute value

Removal of Error & Missing Values

Missing values and error data which obtained because of the advertisement's pop-ups, have been replaced with "null" string.

Clustering

Three clusters were identified form a sample set of data using the Matlab tool K-Medoids the Elbow metric. Further analysys is necessary to understand the meaning of the clusters.



Discussion & Conclusion

- ✓ There are some tools like H2O and Apache Mahout provide facilities to apply data mining techniques like clustering and classification on big data. But these all limited to numerical data only and failed to identify any cluster.
- ✓ Proposed pre-processing methods were able to retain the knowledge
- ✓ K-medioid was able to identify three clusters on a sample data set.

Future Scope

- ✓ Process the whole data to obtain meaningful clusters rather than processing a sample set of data.

References

- ✓ V.Chitraa and Dr. Antony Selvdoss Davamani , 2010. **A Survey on Preprocessing Methods for Web Usage Data.** (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3
- ✓ en.wikipedia.org
- ✓ hadoop.apache.org
- ✓ 0xdata.com