Data Extraction from Web Server Logs using Web Usage Mining

B.Harika[1], Prof .Thatimakula Sudha[2]
[1]*Research Scholar, Dept. of Computer Science, SPMVV, Tirupati*
[2]*HOD & Department of Computer Science, SPMVV, Tirupati*

Abstract- **The growth of the Internet is growing rapidly, and the use of Web sites and Web-based systems has become very common. The main problem facing website administrators or web application systems is increasing the number of data types and formats stored in server log files in seconds, protecting users, future needs, and structure. Website or web service content depending on previous data. Web Usage Mining aims to discover useful information or information about usage data recorded in log files based on the main data types used in the mining process. The data extracted from the web service registry and the browsing behaviour of the user are used to get the same user generated web access order. In this document, we will use one of the web mining techniques to learn about web server log files that record the entire user's browsing history using the web mining technique.**

Index terms- **Web mining; web usage mining; navigation patterns; log file.**

INTRODUCTION

Web Usage Mining (WUM) is an active research area and has the potential to generate new information in Internet-based work. WUM applications are used by several well-known websites to understand customer profiles and their performance regarding the strengths and weaknesses of websites. This article provides a brief introduction to the WUM life cycle, data mining technology, and WUM implementation [1]. The main problem facing website administrators or web applications is the increase in data stored per second in server log files in various types and formats. Learn about users, predict or predict future needs, and protect the structure and content of your website or web service based on previous data. Paper issues, the large number of users on a Web site, large amounts of data about users, and the need for webmaster and site management technologies can help users make quick decisions. To apply the WUM method, a dataset of server log files was selected in this study. The main purpose of using WUM is to gather information about the user navigation model. This information can be used to improve your website from a user perspective. The results from blog mining can be used for a variety of purposes. For example, you can customize the display of web content or pre-cache and cache web design and customer satisfaction to improve user browsing.

RELATED WORK

Recently, WUM has become one of the best research areas. WUM techniques are widely used to detect user navigation patterns in web server logs. DNS monitors user interest by applying a keyword matching approach to the corresponding domain or by working with a search engine. Identify online users on the network and their behaviour or interests [2]. Analysis of website errors that helps system administrators and web designers improve their systems by identifying system errors, broken links, and broken links using WUM [3,4]. Another approach is to use data warehouse data and web data to improve marketing efforts [5]. Surveys were conducted in various categories of Web mining such as Web content mining, Web structure mining, Web usage mining, etc. [6].

SOURCE OF DATA FOR WUM

Data Sources: There are many data sources some are below:

A. Web server logs

Record information about log files, user request history (eg, information about requests such as client IP address, request date and time, request page, HTTP code, provided bytes, user agent, referrer, etc.) are usually recorded. This data can be combined into a single file or split into individual logs, such as access logs, error logs, and referrer logs. Normally, server logs do not collect user-specific information. These files are generally not accessible to ordinary Internet users, but only to webmasters or other administrative organizations [7].

B. Proxy server logs

Web proxy is a caching mechanism between the client browser and the web server. This helps reduce the load time of web pages and the load of network traffic on both sides (server and client). Proxy server logs include HTTP requests from multiple clients to multiple web servers. It detects usage patterns of anonymous user groups and serves as a data source for sharing a common proxy server.

C. Browser logs

JavaScript and Java applets used to collect client-side data. This implementation of client-side data collection requires user cooperation in enabling JavaScript or Java applets in modified browsers [8].

LOG FILE FORMAT

The most popular log file formats are the Common Log Format [W3C] and an extended version. The W3C maintains standard format (Common Log Format) for web server log files (see Fig 1). Information obtained from log file is explained as follows:

Number of Hits: The number of times any resource is accessed in a Website. When a web page is uploaded from a server the number of "hits" or "page hits" is equal to the number of files requested.

Number of Visitors: The users who navigates to website and browses one or more pages.

Visitor Referring Website: The referring website gives the information or URL of the website which referred the particular website in consideration.

Visitor Referral Website: The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

Time and Duration: The time and duration for how long the website was accessed by a particular user.

Path Analysis: Path analysis gives the analysis of the path a particular user has followed in accessing contents of a website.

Visitor IP Address: This information gives the Internet Protocol (IP) address of the visitors who visited the website in consideration.

Browser Type: This gives the information of the type of browser that was used for accessing the website.

Cookies: A message given to a web browser by a web server. The browser stores the message in a text file called cookie. The main purpose of cookies is to identify users and possibly prepare customized web pages for them [7].

Platform: This information gives the type of operating system used to access the website [8].

WUM PROCESS

WUM is a process similar to investigating data mining using various data source types and tools used. The WUM process is a series of steps that can be summarized as follows in Figure 1:

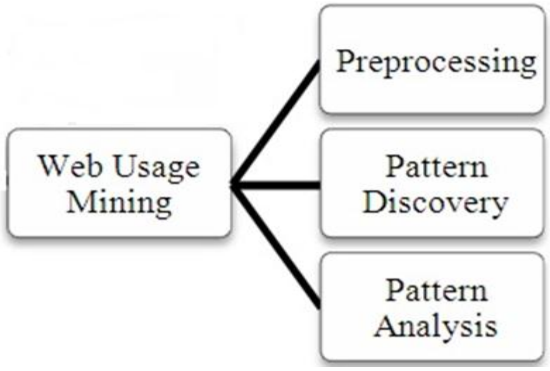


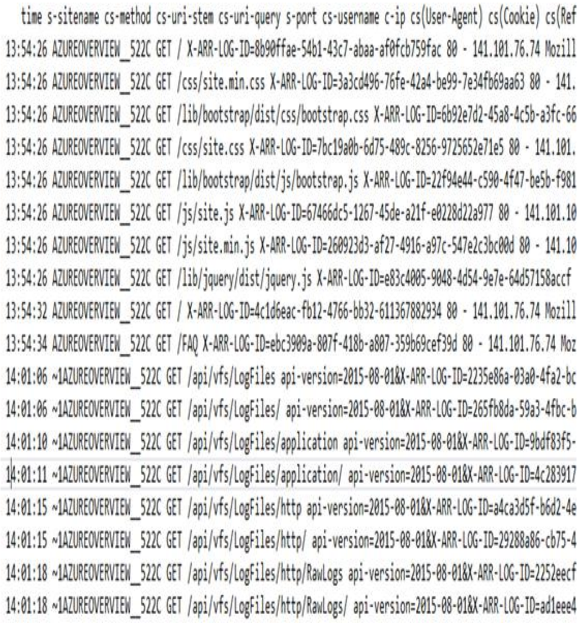
Figure 1. Web usage mining process

a) Data Collection

Within this stage, usage data from various sources are collected from web servers, clients connected to a server, or from middle sources such as proxy servers and packet sniffers.

Data of a typical web server is shown in figure 2, where there is a sample data of the first raw of the log file:

1.2012-10-12 00:01:21 W3SVC4045 C27384-57916 70.87.39.67 GET /robots.txt - 80 - 24.232.136.71 HTTP/1.1 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT +5.1;+S V1;+MSIECrawler) - - www.interactivegt.com 304 0 0 213 320 281.



```
   time s-sitename cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) cs(Cookie) cs(Ref
 13:54:26 AZUREOVERVIEW_522C GET / X-ARR-LOG-ID=8b90ffae-54b1-43c7-abaa-af0fcb759fac 80 - 141.101.76.74 Mozill
 13:54:26 AZUREOVERVIEW_522C GET /css/site.min.css X-ARR-LOG-ID=3a3cd496-76fe-42a4-be99-7e34fb69aa63 80 - 141.
 13:54:26 AZUREOVERVIEW_522C GET /lib/bootstrap/dist/css/bootstrap.css X-ARR-LOG-ID=6b92e7d2-45a8-4c5b-a3fc-66
 13:54:26 AZUREOVERVIEW_522C GET /css/site.css X-ARR-LOG-ID=7bc19a0b-6d75-489c-8256-9725652e71e5 80 - 141.101.
 13:54:26 AZUREOVERVIEW_522C GET /lib/bootstrap/dist/js/bootstrap.js X-ARR-LOG-ID=22f94e44-c590-4f47-be5b-f981
 13:54:26 AZUREOVERVIEW_522C GET /js/site.js X-ARR-LOG-ID=67466dc5-1267-45de-a21f-e0228d22a977 80 - 141.101.10
 13:54:26 AZUREOVERVIEW_522C GET /js/site.min.js X-ARR-LOG-ID=260923d3-af27-4916-a97c-547e2c3bc00d 80 - 141.10
 13:54:26 AZUREOVERVIEW_522C GET /lib/jquery/dist/jquery.js X-ARR-LOG-ID=e83c4005-9048-4d54-9e7e-64d57158accf
 13:54:32 AZUREOVERVIEW_522C GET / X-ARR-LOG-ID=4c1d6eac-fb12-4766-bb32-611367882934 80 - 141.101.76.74 Mozill
 13:54:34 AZUREOVERVIEW_522C GET /FAQ X-ARR-LOG-ID=ebc3909a-807f-418b-a807-359b69cef39d 80 - 141.101.76.74 Moz
 14:01:06 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles api-version=2015-08-01&X-ARR-LOG-ID=2235e86a-03a0-4fa2-bc
 14:01:06 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles/ api-version=2015-08-01&X-ARR-LOG-ID=265fb8da-59a3-4fbc-b
 14:01:10 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles/application api-version=2015-08-01&X-ARR-LOG-ID=9bdf83f5-
 14:01:11 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles/application/ api-version=2015-08-01&X-ARR-LOG-ID=4c283917
 14:01:15 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles/http api-version=2015-08-01&X-ARR-LOG-ID=a4ca3d5f-b6d2-4e
 14:01:15 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles/http/ api-version=2015-08-01&X-ARR-LOG-ID=29288a86-cb75-4
 14:01:18 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles/http/RawLogs api-version=2015-08-01&X-ARR-LOG-ID=2252eecf
 14:01:18 ~1AZUREOVERVIEW_522C GET /api/vfs/LogFiles/http/RawLogs/ api-version=2015-08-01&X-ARR-LOG-ID=ad1eee4
```

Figure 2. An example of raw log file

The log file is a customizable ASCII text-based format. The field prefixes in the file are defined as follows:

s : Server actions.

c Client actions

sc Server-to-Client actions.

cs Client-to-Server actions.

Variable fields appear as:

Date, time, s-sitename, s-computername, s-ip, cs-method, cs-uristem, cs-uriquery, s-port, cs-username, c-ip, cs-version, cs-useragent, cs-cookie, cs-referer, cs-host, sc-status, sc-substatus, sc-win32status, sc-bytes, and cs-bytes [9].

b) Data Preprocessing

Data preprocessing describes any type of processing that is performed on the raw data and prepares it for another processing step. Data preprocessing, commonly used as a preliminary data mining practice, transforms data into a format that is easier and more efficient to meet the user's goals. This is the step of removing the data from noise, resolving inconsistencies, and integrating for use as input to the next stage of pattern detection. The technology used here can provide details of the client data [10].

The different tasks of data preprocessing are:

Data Cleaning: The first step in data preprocessing is to erase raw web data. In this step, available data is examined and extraneous or redundant items are removed from the dataset. Extraneous records are deleted during data cleansing. The target of WUM is to acquire the scanning pattern [10]. Below are two types of unnecessary records that need to be deleted.

1. The records with filenames extension of GIF, JPEG, CSS.
2. The records with noisy data or uncompleted queries are removed.

User Identification: Identifying the individual users who visit your website is one of the most important issues for a successful personalized website. The simplest approach is to assign a different user for each different IP identified in the log file. Cookies also help identify website visitors by storing an ID generated by the web server for each user visiting the website [11].

Session Identification: User session identification also received a great deal of attention in the WUM process, as sessions encode the user's navigation behaviour and are most important for pattern detection. A user session is a delimited set of pages accessed by the same user during a particular visit to a website [12].

c) Pattern Discovery

At this stage, knowledge is discovered by classifying users according to navigation activity. The purpose of classification is to identify characteristics of a predefined class based on a set of instances with users of each class [13]. Classification is a technique for mapping data items into one of several predefined classes. This involves extracting and selecting features that best describe the properties of a particular class or category [14].

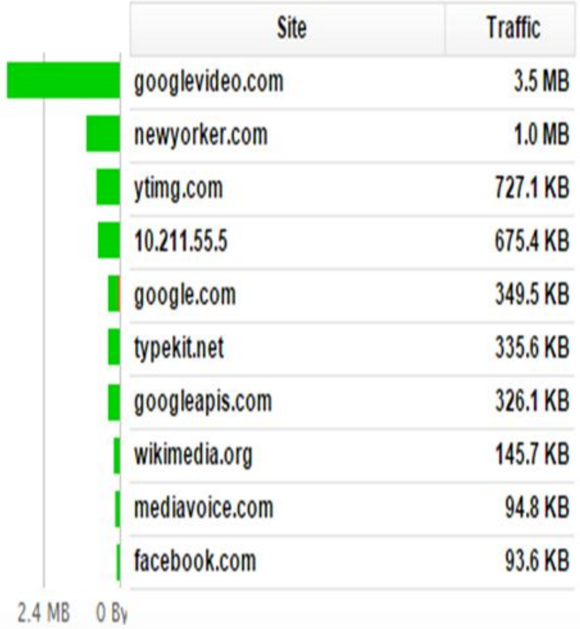
d) Pattern Analysis

This is the last step in the WUM process. After preprocessing and pattern detection, the captured usage patterns are analyzed to filter out non-essential information and extract useful information. You can use methods such as SQL (Structured Query

Language) processing and OLAP (Online Analysis Processing). Using the relational query language (SQL), you can create queries to retrieve data [15]. Develop a high-level data mining query language to enable users to describe data mining tasks by facilitating the specification of relevant datasets, domain knowledge, types of knowledge to mine, and conditions and constraints for analysis must be applied to the discovered pattern. Pattern analysis can automatically detect patterns in data from the same source and make predictions for new data from the same source.

DISCUSSION

This study analyses the web server logs of (www.interactivegt .com) is done with the help of Deep log Analyzer program. The log file consisted of 14-days data about user queries from 12/02/2020 at 12:01:21PM to 25/02/2020 at 11:59:50AM, with size storage of 36.0 MB and content fields are 142252 records. The results of the analysis can be viewed as knowledge to answer these questions. How to extract knowledge from incomplete data structures? Is the log data collected about the user enough to understand the user? What is the optimal structure and content of a website to attract the most interest of visitors? What do users want to do? What is the right method and method of web mining to extract knowledge?



Site	Traffic
googlevideo.com	3.5 MB
newyorker.com	1.0 MB
ytimg.com	727.1 KB
10.211.55.5	675.4 KB
google.com	349.5 KB
typekit.net	335.6 KB
googleapis.com	326.1 KB
wikimedia.org	145.7 KB
mediavoice.com	94.8 KB
facebook.com	93.6 KB

Figure 3. Top visited pages

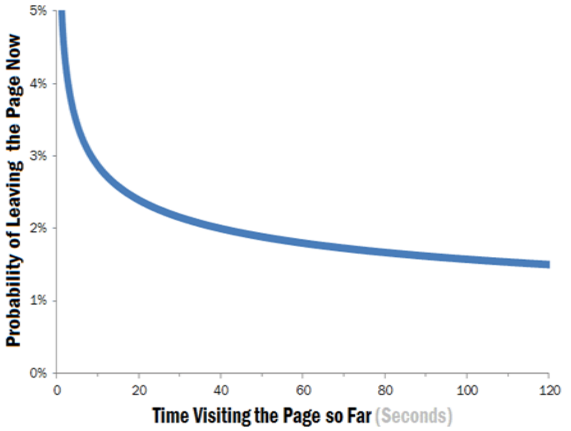
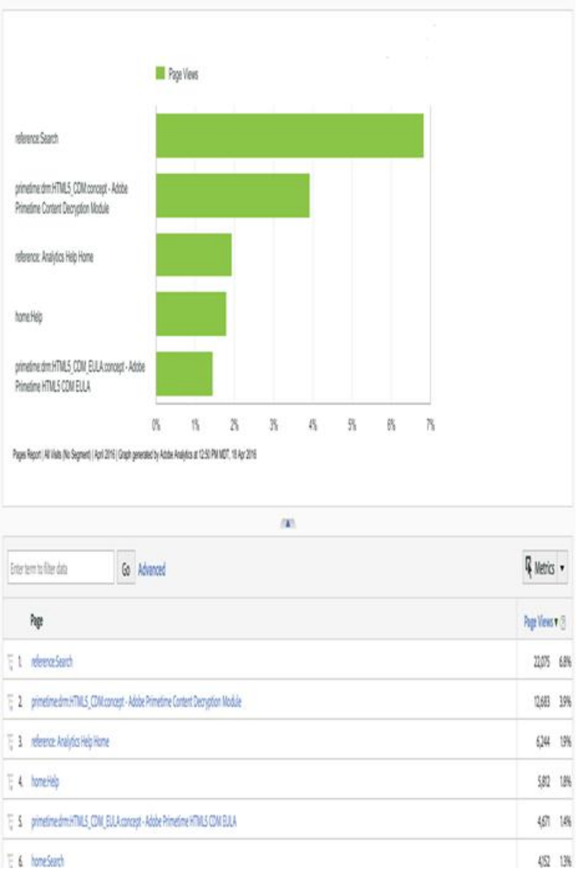


Figure 4. Visitors stay length



Page Views
reference:Search
primetime:drm:HTML5_CDM:concept - Adobe Primetime Content Decryption Module
reference: Analytics Help Home
home:Help
primetime:drm:HTML5_CDM_EULA:concept - Adobe Primetime HTML5 CDM EULA
0%
1%
2%
3%
4%
5%
6%
7%
Pages Report | All Visits (No Segment) | April 2016 | Graph generated by Adobe Analytics at 12:50 PM MDT, 18 Apr 2016
Enter term to filter data
Go
Advanced
Metrics ▾

	Page	Page Views ▾	
1.	reference:Search	22,075	6.8%
2.	primetime:drm:HTML5_CDM:concept - Adobe Primetime Content Decryption Module	12,683	3.9%
3.	reference: Analytics Help Home	6,244	1.9%
4.	home:Help	5,812	1.8%
5.	primetime:drm:HTML5_CDM_EULA:concept - Adobe Primetime HTML5 CDM EULA	4,671	1.4%
6.	home:Search	4,152	1.3%

Figure 5. Popular paths through site

Many different types of results occur according to WUM technique used. Visualization of WUM results should be expressed in high-level languages. The (visual) knowledge can be easily understandable and usable by humans. Implementation of WUM steps displays an example results below. The analysis accessed resources to know user preferences, top visited pages is shown in Figure 3. The analysis visitor activities to know of visit is shown in Figure 4. For an analysis site navigation to understand user behavior, see Figure 5. For an analysis of website popularity, search engines, and phrases used by visitors, see figure 6 which shows visits by hour of day.

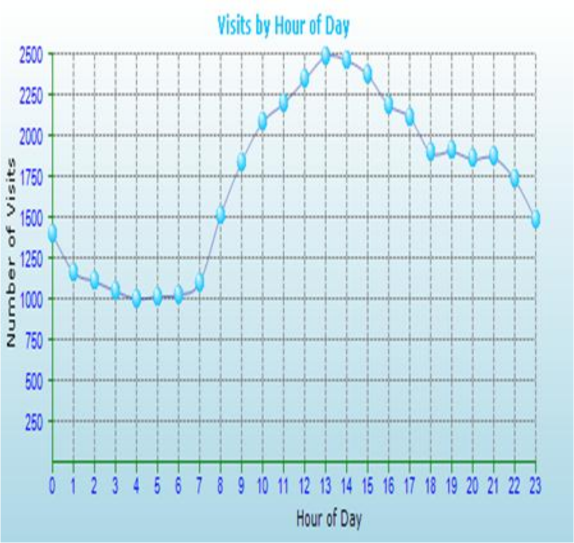


Figure 6. Visits by hou r of day

CONCLUSION

To make a website popular among visitors, website administrators and web designers increase their effectiveness by understanding users and learning about them from usage information stored in log files. The WUM method is very good at extracting knowledge from unstructured data. The obtained WUM results can be used by web administrators or web designers to help organize websites by determining system errors, user settings, technical information about users, broken and broken links. It also solves the problem of exploring ways in which large amounts of information stored in unstructured sources can add ontology-encoded knowledge to unstructured data, and more meaningful ways to leverage knowledge. Finally, this paper has an important aspect of exploring and analyzing user activity and preference data, and despite the importance of building strong relationships between web administrators and users.

REFERENCES

[1] G. Neelima and Sireesha Rodda, An Overview on Web Usage Mining, Emerging ICT for Bridging the Future – Proceedings of the 49thAnnual Convention of the Computer Society of India, vol. 2, Advances in Intelligent Systems and Computing, vol. 338, Springer International Publishing, pp. 647–655, (2015).

[2] A. Bhargav and M. Bhargav, Pattern Discovery and Users Classification Through Web Usage Mining, Proceedings of the International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), IEEE, pp. 632–635, (2014).

[3] M. A. Eltahir and A. F. A. Dafa-Alla, Extracting Knowledge from Web Server Logs using Web Usage Mining, Proceedings of the International Conference on Computing, Electrical and Electronics Engineering (ICCEEE), IEEE, pp. 413–417, (2013).

[4] Sanjay Kumar Malik and SAM Rizvi, Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation, Proceedings of the International Conference on Computational Intelligence and Communication Systems, IEEE, pp. 465–469, (2011).

[5] C. R. Varnagar, N. N. Madhak, T. M. Kodinariya and J. N. Rathod, Web Usage Mining: A Review on Process, Methods and Techniques, Proceedings of the International Conference on Information Communication and Embedded Systems (ICICES), IEEE, pp. 40–46, (2013).

[6] K. Sudheer Reddy, M. Kantha Reddy and V. Sitaramulu, An Effective Data Preprocessing Method for Web Usage Mining, Proceedings of the International Conference on Information Communication and Embedded Systems (ICICES), IEEE, pp. 7–10, (2013).

[7] N. Sael, A. Marzak and H. Behja, Web Usage Mining Data Preprocessing and Multi Level Analysis on Moodle, Proceedings of the ACS International Conference on Computer Systems and Applications (AICCSA), IEEE, pp. 1–7, (2013).

[8] B. U. Maheswari and P. Sumathi, A New Clustering and Preprocessing for Web Log Mining, Proceedings of the World Congresson Computing and Communication Technologies (WCCCT), IEEE, pp. 25–29, (2014).

[9] A. Adamov, Data Mining and Analysis in Depth, Case Study of qafqaz University http Server Log

Analysis, Proceedings of the 8thInternational Conference on Application of Information and Communication Technologies (AICT), IEEE, pp. 1–4, (2014).

[10] M. Joshi, P. Lingras, Yiyu Yao and C. B. Virendrakumar, Rough, fuzzy, Interval Clustering for Web Usage Mining, Proceedings of the 10thInternational Conference on Intelligent Systems Design and Applications (ISDA), IEEE, pp. 397–402, (2010).

[11] K. Santhisree and A. Damodaram, CLIQUE: Clustering Based on Density on Web Usage Data: Experiments and Test Results, Proceedingsof the 3rd International Conference on Electronics Computer Technology (ICECT), IEEE, vol. 4, pp. 233–236, (2011).

[12] S. Nadi, M. Saraee and M. Davarpanah-Jazi, A Fuzzy Recommender System for Dynamic Prediction of User's Behaviour, Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST), IEEE, pp. 1–5, (2010).

[13] Pierpaolo D'Urso and Riccardo Massari, Fuzzy Clustering of Human Activity Patterns, Fuzzy Sets and Systems, vol. 215, Science Direct,pp. 29–54, (2013).

[14] Z. Ansari, A. V. Babuy, W. Ahmed and M. F. Azeemz, A Fuzzy Set Theoretic Approach to Discover User Sessions from Web Navigational Data, Recent Advances in Intelligent Computational Systems (RAICS), IEEE, pp. 879–884, (2011).

[15] Farhat Roohi, Neuro Fuzzy Approach to Data Clustering: A Framework for Analysis,European Scientific Journal March 2013 Edition,vol. 9, no. 9, pp. 183–192, (2013).