Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 327 – 334

Procedia
Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Neuro-Fuzzy Based Hybrid Model for Web Usage Mining

G. Shivaprasad*, N. V. Subba Reddy, U. Dinesh Acharya and Prakash K. Aithal

Department of CSE, Manipal Institute of Technology, Manipal University, Manipal 576 104, India

Abstract

Web Usage mining consists of three main steps: Pre-processing, Knowledge Discovery and Pattern Analysis. The information gained from the analysis can then be used by the website administrators for efficient administration and personalization of their websites and thus the specific needs of specific communities of users can be fulfilled and profit can be increased. Also, Web Usage Mining uncovers the hidden patterns underlying the Web Log Data. These patterns represent user browsing behaviours which can be employed in detecting deviations in user browsing behaviour in web based banking and other applications where data privacy and security is of utmost importance. Proposed work pre-process, discovers and analyses the Web Log Data of Dr. T.M.A.PAI polytechnic website. A neuro-fuzzy based hybrid model is employed for Knowledge Discovery from web logs.

Keywords: Feed forward neural network; Fuzzy *C* means clustering; User identification; User session identification; Web log data.

1. Introduction

The increasing number of web based applications has resulted in collection of massive amount of data in web server logs. This has resulted in Knowledge Discovery by the application of Web Usage Mining techniques. In Web Usage Mining browsing patterns of users are analyzed to extract useful information. Business communities make use of the discovered knowledge to increase the profit by personalizing the web sites for the customer thereby improved customer satisfaction. With the goal of understanding the user behaviors and preferences and thereby increasing the profit of the web based applications, Web Usage Mining discovers interesting usage patterns hidden in web data[1]. Usually, Web Usage Mining consists of Pre-Processing, Knowledge Discovery and Pattern Analysis.

Web Server Logs serves as the input to the Web Usage Mining process. The Web Log Data is unstructured and noisy and ambiguous. In order to extract useful patterns from Web Logs, it must be preprocessed to remove noisy and irrelevant data, thereby reducing the bulk of data to be processed. Data mining algorithms can then be applied to the pre-processed web logs to extract useful patterns.

The uncertain and fuzzy characteristic of the browsing behavior of user is very difficult to model. Generally, a Web site attracts many different groups of users. For example, a college web site visitor group may include prospective students, parents, bankers, civil contractors, book publishers and book shop owners etc along with other anti social

*Corresponding author. Tel.: +91-0820-2924514.
E-mail address: shiva.prasad@manipal.edu

doi:10.1016/j.procs.2015.06.038

elements. Each group can have some specific need or goal. Further, a user within a group can have different intention during his visit to the web site. For example, a banker may visit a college web site and access links pertaining to admission dates so as to campaign his bank for educational loans. Same banker may visit the college web site to know about faculty details to campaign their Car or House Loan Scheme. Hard Clustering algorithms fail to capture such overlapping behaviors or interests of users as these algorithms push every object exclusively to a single cluster. Hence, fuzzy clustering algorithms are more suitable for Web Usage Mining domain. Hence, in the earlier work, the Fuzzy C-Means Clustering algorithm was employed for clustering the web user sessions.

In this work, a hybrid model based on neuro – fuzzy clustering is implemented to efficiently cluster the users of polytechnic website based on similar browsing patterns. The Web Log was preprocessed using Dimensionality reduction techniques and combined methodologies.

The rest of the paper is organized as follows. Section 2 presents the related work; Section 3 presents the proposed method; Section 4 deals with the experimental set up and result analysis and finally, Section 5 concludes the work.

2. Related Work

A. Bhargav, *et al.*[2] proposes a framework for Web Usage Mining consisting of Pre-processing, Pattern Discovery and Users classification. This framework classifies the users based on country, site entry and access time. M. A. Eltahir, *et al.*[3] and Sanjay Kumar Malik, *et al.*[4] explores and discusses Information extraction from user navigation history using Web Usage Mining. A detailed survey on data collection and pre-processing stage of web usage mining is discussed by Varnagar C. R.[5]. K. Sudheer Reddy, *et al.*[6] proposes several data preparation techniques of access stream to identify the unique sessions and unique users. To analyze the learners' behaviour to help in learning evaluation and to enhance the structure of a given course, Educational data mining techniques are employed[7]. B. U. Maheswari, *et al.*[8] proposes a new algorithm for pre-processing and clustering of web log.

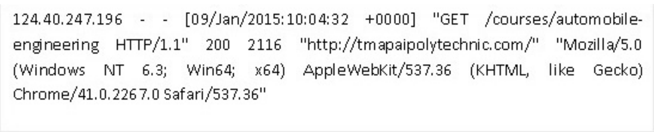
Data Mining and Analysis techniques based on regular expressions on the data generated by University HTTP Server Logs has been proposed by Adamov A.[9]. M. Joshi, *et al.*[10] implements and analyzes the performance of different soft – computing algorithms over an educational site. CLIQUE (CLUstering in QUEst) algorithm for clustering web sessions for web personalization has been adopted by K. Santhisree, *et al.*[11]. S. Nadi, *et al.*[12] proposes a model for dynamic recommendation based on fuzzy clustering techniques, applicable to currently on-line users. The fuzzy clustering approach, in this study, provides the possibility of capturing the uncertainty among Web user's behaviours. A fuzzy clustering around medoids approach is adopted by Pierpaolo D'Urso *et al.*[13] to classify ordered sequences (paths) representing patterns of individual behaviour in an actual or virtual space – time domain. Z. Ansari, *et al.*[14] employs a Fuzzy Membership Function to assign weights to sessions based on the number of URLs accessed by the sessions followed by application of Fuzzy *c*-Mean Clustering algorithm to discover the clusters of user profiles. A neuro – fuzzy model for data clustering is introduced by Farhat Roohi[15].

3. Methodology

3.1 Web log data collection

Each user request to the server will be recorded in a web server log. Each line in a log represents the request made to the server and usually in Extended Log Format contains the following fields:

i. Host/The Remote IP address: Identifies who had visited the web site.
ii. User Authentication: Username and password if the server requires user authentication (generally "-").
iii. Date and time of the request: Used to determine how long a visitor has spent on a given page.
iv. The HTTP Request: The Method (GET, POST, HEAD, etc.) used for information transfer is noted along with the Requested Resource Name (an HTML page, an image file, or a script) and Protocol Version (HTTP protocol being used).
v. The Request Status: HTTP status code returned to the client (200, 404 etc.).
vi. The Page size: Content – length of the document transferred.



```
124.40.247.196 - - [09/Jan/2015:10:04:32 +0000] "GET /courses/automobile-
engineering HTTP/1.1" 200 2116 "http://tmapaipolytechnic.com/" "Mozilla/5.0
(Windows NT 6.3; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/41.0.2267.0 Safari/537.36"
```

Fig. 1. Sample web log.

vii. The Referrer Field: Lists the URL of the previous site visited by the client, which is linked to the current page (-if this information is missing).

viii. The User Agent Field: Provides information about the client's browser, the browser version, and the client's operating system (-if this information is missing).

The web access log was collected from the web server of Dr. T. M. A. Pai Polytechnic web site[16] from 31st Dec. 2014 12:09:56 through 11:18:07 15th Jan. 2015, a total of 15 days. The sample of web log record is given in Fig. 1.

The above log entry indicates that user with IP Address 124.40.247.196 requested the link automobile – engineering under courses on 9th Jan 2015 at 10:04:32 AM and he traversed from the link http://tmapaipolytechnic.com. The request was successful and a total of 2116 bytes have been downloaded. Also, it tells that Mozilla (compatible) 5.0 was the browser and Windows NT 6.3 was the operating system used.

3.2 *Web log pre-processing*

Web log consists of all interactions with the web server. These consist of actual user requests and requests that are initiated due to the image files present in the web page. In addition to user requests, the log also consists of automated requests initiated by robots or bots. In order to discover useful patterns from web logs, the web log should be preprocessed to identify the users and user sessions or user activities in the web site. The web log preprocessing consists of Web Log Cleaning, User Identification and Session Identification. Prior to web log cleaning, the features in the web log need to be extracted and time stamp, an integer representing the time elapsed from baseline date need to be computed for further computational use.

The web log pre-processing using different dimensionality reduction techniques and combined methodologies is employed to efficiently identify the underlying users and user sessions.

3.2.1 *Web log cleaning*

A general algorithm for web log cleaning is presented below. The algorithm takes as input the web log file (WebLogFile) and produces as output, a cleaned log file (New – LogFile), which is free from irrelevant and redundant entries. The algorithm calls routines for checking whether the current request is an image request, a robot request, a successful GET request, which are boolean functions returning TRUE or FALSE.

```
Step 1. Repeat the steps 2–3 until end-of-file (WebLogFile) is not reached.
Step 2. Read the next record of WebLogFile into Log – Record.
Step 3. If the Log – Record is not a robot request and not an image request and is a successful GET request write
        Log – Record to New – LogFile.
Step 4. Stop
```

3.2.2 *User identification*

User Identification based on IP address, User Agent and Referrer Field has been implemented to uniquely identify the users. The algorithm for User Identification is given below which take as input the cleaned log file and identifies each distinct user.

```
Step 1. Set User – Count ← 0.
Step 2. Repeat the steps 3–8 while end of file(New – LogFile) is not reached.
Step 3.  Read the current record of New – LogFile into Cur – Record.
Step 4.  Read the next record of New – LogFile into Next – Record.
Step 5.  Let IPcur and IPnext be the IP Address in Cur – Record and Next – Record.
Step 6.  Let UAcur and UAnext be the User Agent in Cur – Record and Next – Record.
Step 7.  Let RUcur and RUnext be the Referrer URL in Cur – Record and Next – Record.
Step 8.    If IPcur <> IPnext [IP addresses are different]
             Identify both entries as belonging to different user.
             Increment the User-Count by 1.
           Otherwise
             If UAcur <> UAnext [Both browser and OS are unique]
                  Identify both entries as belonging to different user.
                  Increment the User-Count by 1.
             Otherwise If RUcur<>"-"and RUnext = "-"
                       Identify both entries as belonging to different user.
                       Increment the User-Count by 1.
                  otherwise
                       Assume as same user.
Step 9.  Stop
```

3.2.3 *Session identification*

Session Identification aims at identifying the user activities in a web site. The algorithm for Session Identification based on both navigation and time oriented heuristics is presented below.

```
Step 1.  Set PageStayTime ← 10 minutes.
Step 2.  Let SessionSet = { }
Step 3.  Set K ⇓ 0 and J ⇓ 0.
Step 4.  Let Lj, URIj, tj , RUj and Uj denote log entry, URI, time stamp, Referrer URL and user respectively.
Step 5.  For each unique user Uj do the steps 6–8
Step 6.  Let I ⇓ 0
Step 7.  For each log entry Lj do the step 8.
Step 8.    If referring URL, RUj = " –" and (tj- tj-1)>PageStayTime
                K←K+1
                Sk ←URIj
                SessionSet ←SessionSet U Sk
             Otherwise
                If RUj is Present in any Si of Uj
                     Si ←URIj
                Else
                     K←K+1
                     Sk ← URIj
                     SessionSet = SessionSet U Sk
Step 9.  Stop
```

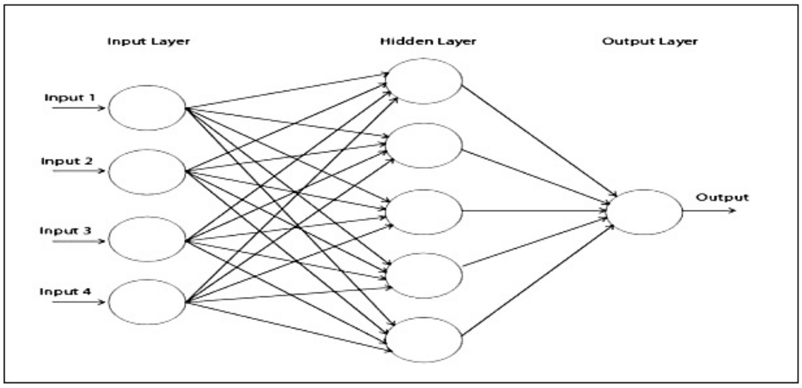


Fig. 2. Multi – layer feed forward neural network.

3.3 Knowledge discovery using neuro – fuzzy clustering algorithm

Once the user sessions have been identified, a clustering process is applied in order to group similar sessions in the same cluster. Each cluster includes users exhibiting a common browsing behaviour and hence similar interests. In the earlier work, the Fuzzy C-Means Clustering algorithm was employed to cluster the user sessions.

Artificial neural network seeks to emulate the architecture and information representation patterns of the human brain. Artificial neural networks are designed as per the target to be accomplished. Patterns are presented at the input, which are associated with the output nodes with differential weights. An iterative process is followed to adjust the weights between the input nodes and the output nodes until a termination criterion is satisfied. This process of weight adjustment, called learning, provide, continuous learning or artificial learning capability to the system, which can be either supervised or unsupervised learning in artificial neural networks. The supervised learning demands an output class declaration for each of the inputs.

In the current work, we create a hybrid model based on neural networks and fuzzy clustering to cluster users according to the browsing patterns. First, a sample set of pre-processed web log data is clustered using Fuzzy C Means Clustering algorithm. Then, the input of clustering algorithm is given as input to neural network and the output of clustering algorithm is given as target output and the training of neural network is done. Mean Square Error – the average squared error between the network outputs and the target outputs is used as the performance measure.

We employ a Multi – layer Feed forward neural network, trained with a back – propagation learning algorithm as shown in Fig. 2.

4. Experimental Setup and Results Analysis

The web access log was collected from the web server of Dr. T. M. A. Pai Polytechnic web site from 31st Dec. 2014 12:09:56 through 11:18:07 15th Jan. 2015, a total of 15 days. The web log recorded a total of 5817 requests during this period.

4.1 Web log pre-processing

The Web Log Cleaning algorithm eliminated all irrelevant records from the web logs, retaining only 1169 records, approximately 20% of the records for further processing. The statistics of irrelevant requests recorded under different categories is shown in Table 1.

It can be observed that around 50% request of the web log constitute image downloads, which is eliminated by the web log cleaning algorithm. Table 2 shows the aggregate results of web log cleaning algorithm.

The application of User Identification algorithm over the cleaned web log identified a total of 235 unique users. This is followed by the application of Session Identification algorithm. The algorithm identified one session per user with maximum number of pages in a session = 25.

Table 1. Statistics of individual request (irrelevant) category.

Request category	Number of records	Percentage
PNG	408	7
JPEG	2501	43
CSS	291	5
JS	582	10
NOT GET	58	< 1
NOT 200	1164	20
Robot	931	16

Table 2. Aggregate results of web log cleaning.

Statistics	Number of records
Original size	5817
Corrupt requests	–
Failed requests (other than 200)	1164
Multimedia objects	3782
Cleaned log size	1169
Percentage in reduction	80

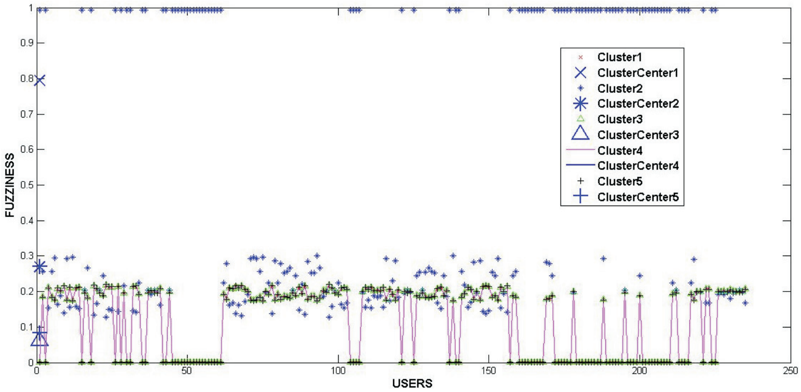


Fig. 3. Final fuzzy cluster centers.

4.2 Knowledge discovery phase

The Fuzzy C-Means Clustering algorithm was implemented using MATLAB fuzzy tool. The MATLAB function fcm performs Fuzzy C-Means clustering. The function fcm takes a data set and a desired number of clusters and returns optimal cluster centers and membership grades for each data point. It starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Next, fcm assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, fcm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade. Fuzzy C-Means clustering is an iterative process. The process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified. We begin with the value 2 for number of clusters, 100 for number of iterations and 1e-5 for minimum improvement. The Fig. 3 shows the final fuzzy cluster centers.

Results show that sessions were divided into 5 clusters. User – sessions are distributed among these clusters. Most of the users in cluster 5 have high degree of membership.

After the clustering algorithm has been applied and clusters have been formed, the next step is training the Neural Network to cluster the input web logs automatically into identified clusters. For this purpose, Multilayer Feed Forward Neural Network was implemented using MATLAB Neural Network Tool as shown in Fig. 4.

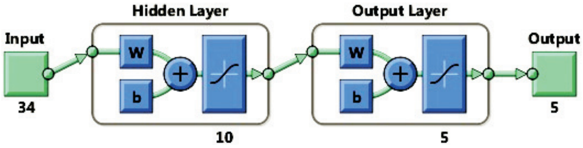


Fig. 4. Multilayer feed forward neural network using MATLAB neural network tool.

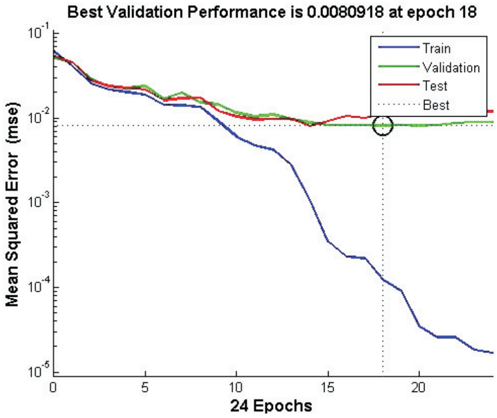


Fig. 5. Performance of neural network.

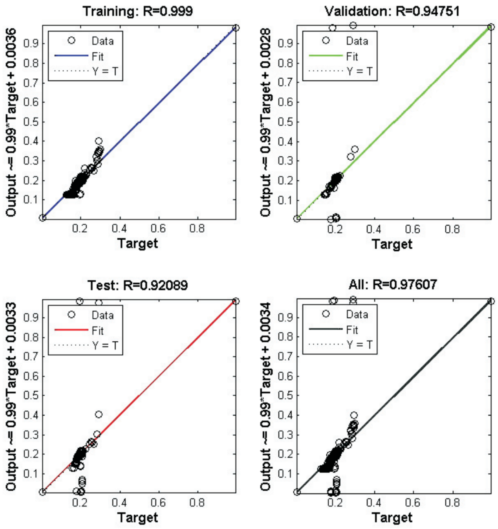


Fig. 6. Regression plots of the network.

The neural network attained the best validation performance = 0.0080918 at epoch 18 as shown in Fig. 5.

The regression plots displaying the network outputs with respect to targets for training, validation, test and all sets is shown in Fig. 6.

For a perfect fit, the data should fall along a 45 degree line, where the network outputs are equal to the targets. For this problem, the fit is reasonably good for all data sets, with R value of 0.92 for test data and a value of 0.97 for overall data. Neural Network accurately classified the users into different clusters as per their browsing patterns.

5. Conclusion

Web Log Data is a repository of patterns of user activities in a web site. Knowledge Discovery from Web Log Data has an important role in serving the needs of web based applications. In this work, neuro – fuzzy based hybrid model is implemented to discover hidden patterns in the Web Log of polytechnic web site. Web Log Pre-processing techniques based on dimensionality reduction techniques and combined methodologies were employed. The preprocessing step eliminates all irrelevant and noisy data, with a resulting Web Log size of 20% of the original log size. The neuro – fuzzy model combines the neural networks and the fuzzy set theory. Clustering is a subjective process, which means that the same set of data items repeatedly need to be partitioned differently for various applications. It makes clustering difficult as a single algorithm or approach will be inadequate to solve all the clustering problems. This problem is taken care of by the neuro – fuzzy system as it is a self learning system and generates patterns and rules automatically. The neuro – fuzzy clustering grouped the users having similar browsing patterns into clusters. Also, the information gained from the analysis can then be used by the website administrators for efficient administration and personalization of their websites.

Acknowledgements

Authors are thankful to the Managing Committee of the Dr. T. M. A. Pai Polytechnic, Manipal for facilitating collection of Web Log Data and making available the collected data for the research purposes.

References

[1] G. Neelima and Sireesha Rodda, An Overview on Web Usage Mining, *Emerging ICT for Bridging the Future – Proceedings of the 49th Annual 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 Congress on 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 8th International 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 10th International 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, *Proceedings of 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).

[16] Dr. T. M. A. Pai Polytechnic Web Site URL http://www.tmapaipolytechnic.com/