INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)

ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online)
Volume 4, Issue 3, May-June (2013), pp. 101-112
© IAEME: www.iaeme.com/ijcet.asp
Journal Impact Factor (2013): 6.1302 (Calculated by GISI)
www.jifactor.com

IJCET
©IAEME

# DEVELOPMENT OF PATTERN KNOWLEDGE DISCOVERY FRAMEWORK USING CLUSTERING DATA MINING ALGORITHM

**Mr. Rinal H. Doshi**
Research Scholar
School of Engineering,
R.K. University, Rajkot
L.C. I.T., Bhandu, India

**Dr. Harshad B. Bhadka**
Director- MCA
C. U. Shah University,
Surendranagar, India

**Ms. Richa Mehta**
Research Scholar,
Lecturer, SKPIMCS
KSV University,
Gandhinagar, India

## ABSTRACT

The prime objective of this research work is to identify a pattern of clustering and extend to improve the use of Web Data Mining. This extension helps to sensitize Knowledge Discovery and Business Improvement Intelligence. The motivation is to analyze user access patterns and improve the access privileges of the users. These improved access privileges helps to channelize the analysis for optimized selection of objects. This work obtained secondary dataset of CPU processors from the web data repository UCI. The dataset was subjected to the application of the techniques of statistics, machines learning, and clustering data mining. The selection of appropriate Web Mining Technique is a challenge for Knowledge Discovery. The selection primarily depends upon the nature of a dataset. To obtain optimized solution to the challenge, proposed work uses the Clustering Data Mining techniques. The results obtained there at are analyzed to optimum outcome of data. To achieve optimum result this work has proposed a framework Pattern Knowledge Discovery for macro level and micro level pattern evaluation and test the proposed framework with its implementation for the purpose of justification of the work.

**KEYWORDS:** Web Mining, Data Mining, Knowledge Discovery Framework, Machine Learning, Business Intelligence

## 1. INTRODUCTION

For this research work the focus is Data Mining and within the Data Mining the sub area is Web Data Mining. The areas of Web Mining have three dimensions: Content, Structure and Usage. The ultimate the focus work is Web Usage Data Mining and in that Clustering Techniques are used to analyze user access pattern from web dataset repositories and find hidden predictive knowledge.

## 2. LITERATURE REVIEW ANALYSIS AND REVIEW FINDINGS

The authors mainly focus on to get knowledge regarding improvement and implement web development path by suggesting the sources of Web Usage Data like Web Server, Proxy Server and Web client. They propose preparation, user identification and session identification algorithm. They use Data Mining and Knowledge Discovery technique for achieve targeted information [1]. While identify the session, universal variable like time is not considered.

The work focus by authors is to generate data for web developers to optimize web development scenario. They analyze statistical data of web browsing including parallel browsing behavior of user to extract knowledge with the help of Graph Mining, onPageLoad, onPageFocus, OnSessionEnd and preprocess algorithm [2]. Here they can consider time variable while calculating the user session. But not consider the firewall and antivirus overheads while calculating user session timings.

The authors focus on to reduce Field Extraction time by improving cleaning process of Web Usage Mining. They use Preprocess, Field Extraction, Data cleaning, and Session information for achieving targeted objective [3]. But in this research Machine learning approach (only talk about CLF file) is not adopted.

The authors have tried to improve Customer oriented approach in enhancement of website development design and to formulate more customised web development. To achieve this they apply pre- processing and clustering data mining technique on log file and establishment web browsing path by using rate and time matrices of visited page [4]. In this work they cannot fill the gap of untrained customers who are monitored and analysed.

Authors of the paper analyse the comparative study of Web Mining's today's structure and tomorrow view [5]. There is no clarity between web mining and data mining in their survey work.

Author introduces Web Mining and Web Mining Techniques to achieve targeted information related to e- commerce. They used Pattern Discovery and Analysis tool to find user access pattern targeted to website optimization, website design customization, customer relationship management, e-commerce security

enhancement, operating cost reduction to improve overall competitiveness of an enterprise [6].

Authors have developed algorithm for mining learner interest and offering personalized design for e-learning. The mining patterns provides user specific correlated pages, learner's interest in pages, etc. which can help to enhance the effectiveness of teaching website improving the design [7]. To predict personalized need of student is challenging task.

Authors used Clustering Web Usage Mining on secondary data in order to find user access pattern and order of their visits of hyperlink. This can lead to cluster user with similar access pattern to collaborate [8]. The challenge of their work is to decide initial cluster and improve the cluster algorithm.

The author applied text mining on a set of 763 documents from Web databases and analysed word frequency, keywords etc. to draw conclusions on topic/keyword of interest [9][10]. For getting optimum outcome this work need to apply text mining on more documents from various locations.

Authors applied Web Mining Techniques to analyze security data on e-commerce websites based on user behaviour. The analysis involved Clustering Techniques of K-means to partition observation into clusters with related security issues [11]. For getting more accurate outcome more clustering technique must be needed.

The authors have Clustering and Association Mining Techniques in inter-mix form to Improve Web Personalization [12].

## 3. PROPOSED FRAMEWORK

The analysis of literature review is to be the foundation to design knowledge discovery framework. The proposed framework is to be implemented with utilization of diverse dimension of application of dataset to provide diversified services under neat the propose framework. The plan of phases to be involved in the implementation data collection, data pre-processing, and tabulate the result in all phases with their individual analysis. The selection of appropriate Web Mining technique is the only challenge for Knowledge Discovery and the selection of technique is based on the nature of a dataset. The work is extended to discover knowledge by using Web Data Mining. Proposed work focus on Clustering Data Mining Technique and analyzing data to find out optimum outcome of data. Extracted result will generate knowledge. To fulfill the proposed work this work focus on implementing pattern knowledge discovery framework.

### 3.1 Graphical View Of Proposed Framework

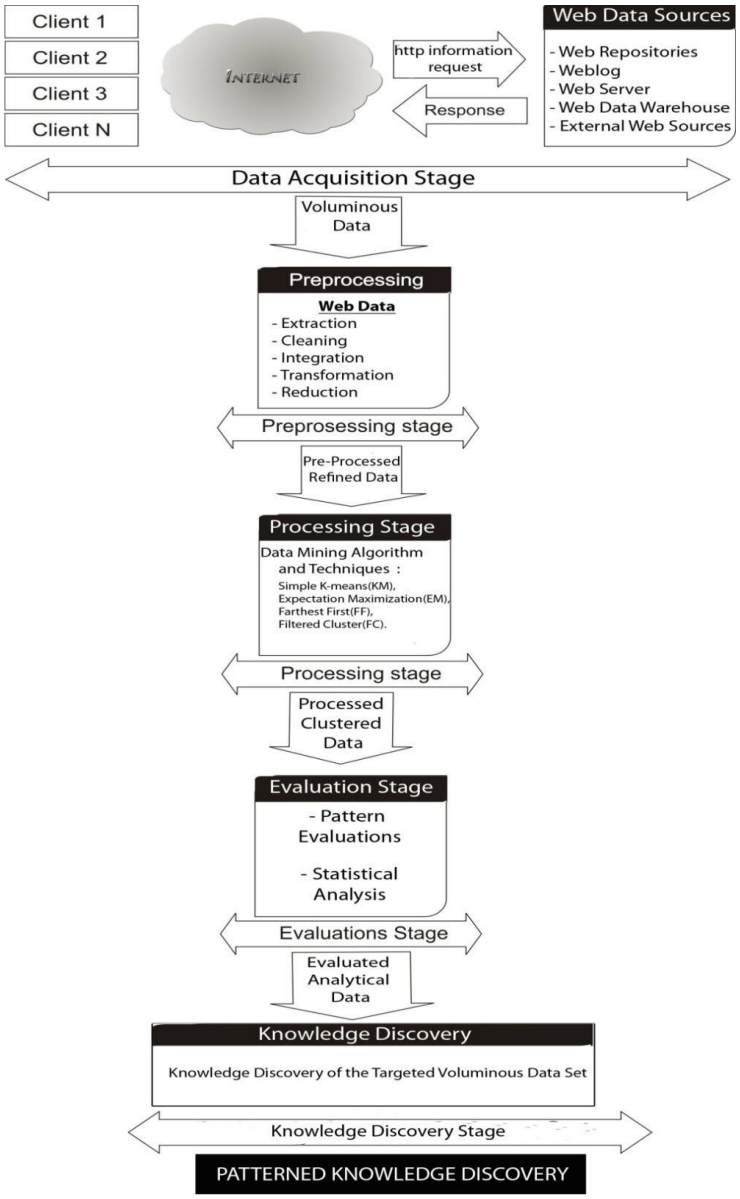


**Fig. 1:** GRAPHICAL VIEW OF PROPOSED FRAMEWORK
(PATTERN KNOWLEDGE DISCOVERY FRAMEWORK)

### 3.2 Logical view of Proposed Framework

**3.2.2** Data Acquisition Stage

Input: http information request

Processing: searching for the requested information in the Web Sources like web repositories, weblog, Web Server, Web Data warehouse, Web Databases, External Web Sources etc.

Output: Response of the requested information in the form of voluminous data

**3.2.3** Pre-processing Stage

Input: Voluminous data acquired from data acquisition stage

Processing: Voluminous data pre-processing is done at five levels-Extraction (Level 1), Cleaning (Level 2), Integration (Level 3), Transformation (Level 4) and finally reduction (Level 5).

Output: Pre-processed refined data.

**3.2.4** Processing Stage

Input: Refined data after the five level pre- processing stages.

Processing: Processing of the refined pre- processed data using Data Mining open source tool WEKA. Processing is done through various clustering Data Mining algorithm and techniques- Simple K-Means (KM), Expectation Maximization (EM), Farthest First (FF) and Filtered Cluster (FC).

Output: Processed Clustered data

**3.2.4** Evaluation and analysis Stage

Input: Clustered data from the processing stage.

Processing: Macro level pattern evaluation and statistical analysis is performed on the various integrated User Access Pattern obtained in the form of different types of clusters after applying the processing stage cluster Data Mining Techniques.

Output: Evaluated analytical data.

**3.2.5** Pattern Knowledge Discovery Stage

Input: The Evaluated analytical data after pattern evaluation and statistical analysis.

Processing: The output obtain from evaluation stage is studied at micro level to get patterns which results in the Knowledge Discovery of the targeted voluminous data set.

Output: Pattern Knowledge Discovery

## 4. IMPLEMENTATION

The process uses a set of clustering techniques for mining hidden patterns in WEKA open source data mining tool.

### 4.1 Implementation using Simple K- Means Algorithm

This technique takes input from CPU_PERFORMACE pre-processed data set and then applies K- means clustering technique.

The step involved in K- Means algorithm are: [13]

To divide items or objects into groups of subsets or sub-subsets.

To obtain central values of the clusters.

To allocate each item or object to a relevant cluster with the next-door points or nearest mean points.

To obtain the central values of the clusters Repeat step 2 to step 4 till the conversions has been reached. The Cluster central values are tabulated in Table 1.

**TABLE 1:** CLUSTER CENTRAL VALUES

| Attribute | Full data (213) | Cluster 0 (163) | Cluster 1 (50) |
|---|---|---|---|
| Supplier | ALL* | amd | Zilog |
| CycleTimeNS | 207.0469 | 250.7485 | 64.58 |
| RAMMin | 2876.8732 | 1845.8282 | 6238.08 |
| RAMMax | 12025.3333 | 8191.5092 | 24523.6 |
| CacheMem | 25.9343 | 11.9509 | 71.52 |
| ChnlMin | 4.8732 | 3.0613 | 10.78 |
| ChnlMax | 18.3099 | 12.7117 | 36.56 |
| P_RPP | 99.8732 | 51.6687 | 257.02 |

ALL* (213) = amd (163) + zilog (50)

This process took 12 iterations and to build the clustered datasets it took 0.02 second. Number of clusters selected by cross validation is 2 (0 through 1).

The model and evaluation of clustered instance are tabulated in Table 2 and visualized as Figure 2.

**TABLE 3:** CLUSTERED INSTANCES

| Cluster no | Instances | Percentage % |
|---|---|---|
| **0** | 163 | **77** |
| **1** | **50** | **23** |

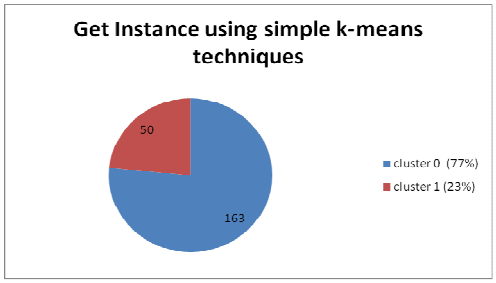


**Fig. 2:** INSTANCE USING K-MEANS TECHNIQUES

This Simple K-Means technique resulted in 2 clusters comprising of amd and zilog suppliers for CPU performance.

### 4.2 Implementation using Expectation Maximization Algorithm

The EM (Expectation Maximization) algorithm is a technique which reflecting the abstraction of clustering K-Means algorithm. It tracks an iterative loom, sub-optimal, which seeks to get the constraints of the probability distribution that can say maximum probability of its characteristics. EM Clustering is basically model based clusters [14].

Number of clusters selected by cross validation is 6 (0 through 5).

This research work finds the means and standard deviation of different parameter like CycleTimeNS, RAMMin, RAMMax, CacheMem, ChnlMin, ChnlMax, P_RPP.

This process took 11.22 second to build to model the clustered the datasets. The model and evaluation of clustered instance are tabulated in Table 4 and visualized in Figure 3.

**TABLE 4:** CLUSTERED INSTANCES FOR CPU_PERFORMANCE

| Cluster no | results | Supplier |
|---|---|---|
| 0 | 36% | motorola |
| 1 | 9% | sony |
| 2 | 9% | amtel |
| 3 | 2% | zilog |
| 4 | 20% | freescale |
| 5 | 24% | amd |

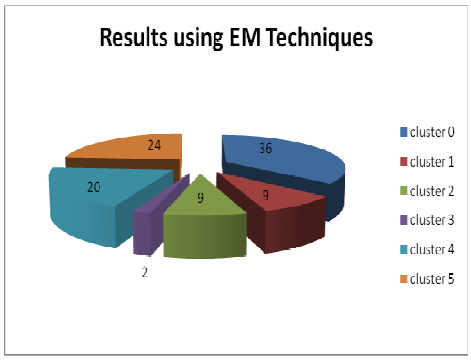


**Fig. 3:** INSTANCE USING EM

This EM technique resulted in 6 clusters comprising of Motorola, amd, freescale, Sony, amtel, zilog suppliers for CPU performance.

### 4.3 Implementation using Farthest First Algorithm

Farthest First clustering technique offers a very fast analysis when we carry on in brain point constrain, except builds reasonably high error rate to other clustering technique. Farthest first is an alternative of K means that places each cluster center in turn at the point furthest from the existing cluster centers. This point must laze contained by the data part. These significantly speed up the clustering in mainly cases since a lesser amount of relocation and modification is desirable [15].

This research work used farthest fast method for discovering pattern form datasets using WEKA open source tools.

In this technique incorrect instances are 175. Cluster centroids obtain by weka tools are tabulated in Table 5.

**TABLE 5:** CENTROIDS USING FF CLUSTERING TECHNIQUES

| Cluster | Centroids |
|---|---|
| 0 | 125.0 512.0 1000.0 0.0 8.0 20.0 19.0 |
| 1 | 23.0 32000.0 64000.0 128.0 32.0 64.0 1238.0 |

The model and evaluation of clustered instance are tabulated Table 6 and visualized in Figure 4.

**TABLE 6:** CLUSTER INSTANCES FOR CPUPERFORMANCE

| Cluster | instances | Percentage | Supplier |
|---|---|---|---|
| 0 | 208 | 98% | amd |
| 1 | 5 | 2% | amtel |

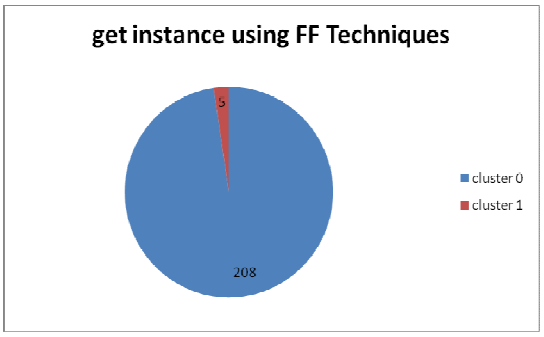


**Fig. 4:** INSTANCE USING FF

This FF technique resulted in 2 clusters comprising of amd and amtel suppliers for CPU performance.

### 4.4 Implementation using Filtered Cluster Algorithm

In FilteredClusterer instances will be procedure by the filter not including altering its structure. FilteredClusterer class is in the weka. clusterers. FilteredClusterer package. Details of methods are in this package [16].

In FilteredClusterer Data Mining Techniques number of iteration is seven. In this techniques cluster centroids are tabulated in Table 7.

**TABLE 7:** CLUSTER CENTRAL VALUE

| Attribute | Full Data (213) | Cluster 0 (186) | Cluster 1 (27) |
|---|---|---|---|
| CycleTimeNS | 207.0469 | 231.4462 | 38.963 |
| RAMMin | 2876.8732 | 1865.1505 | 9846.5185 |
| RAMMax | 12025.3333 | 8674.172 | 35111.1111 |
| CacheMem | 25.9343 | 16.3656 | 91.8519 |
| ChnlMin | 4.8732 | 3.2043 | 16.3704 |
| ChnlMax | 18.3099 | 13.7957 | 49.4074 |
| P_RPP | 99.8732 | 56.129 | 401.2222 |

Incorrect instance are 173. The model and evaluation of clustered instance are tabulated in Table 8 and visualized in Figure 5.

**TABLE 8:** CLUSTER INSTANCE OF CPU_PERFORMANCE

| cluster | instances | % | Supplier |
|---|---|---|---|
| 0 | 186 | 87% | amd |
| 1 | 27 | 13% | amtel |

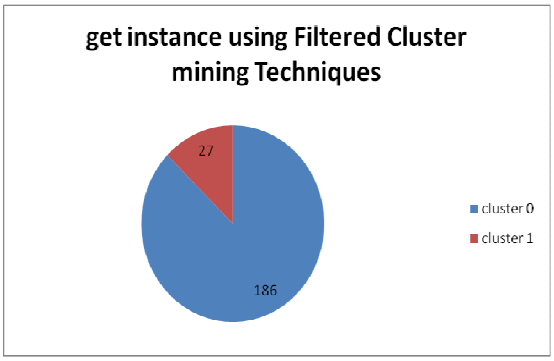


**Fig. 5:** INSTANCE USING FILTERED CLUSTERERS

This Filtered cluster technique resulted in 2 clusters comprising of amd and amtel suppliers for CPU performance.

## 5. RESULT ANALYSIS WITH JUSTIFICATION

The outputs generated by the application of four Data Mining Techniques on to the dataset of CPU_PERFORMCE are tabulated in Table 2 for the purpose of Result Analysis as Table 9.

**TABLE 9:** ANALYSIS TABLE

| Supplier / Algorithm | amd | amtel | motorola | frescale | Sony | zilog |
|---|---|---|---|---|---|---|
| **Simple K- means** | 77% | 23% | | | | |
| **EM** | 24% | 9% | 36% | 20% | 9% | 20% |
| **FF** | 98% | 2% | | | | |
| **Filtered Cluster** | 87% | 13% | | | | |

The results are obtained by subjecting the dataset of CPU_PERFORMANCE to a Data Mining Techniques. The data volume is 213 instances in the dataset. Each instance is a group of seven attribute and one class. The result obtained on applying K-Means clustering techniques. This technique generated two clusters classifying the cluster named amd (77%) and zilog (23%) though there are 29 suppliers. This research uses K-Means clustering

technique because it produces computationally faster and tighter cluster as compare to hierarchical clustering technique. K-Means algorithm is able to identify the non linear structure. It is best suit for the real life dataset or web log data.

The result obtained on applying EM clustering techniques. This technique generated six clusters classifying the cluster named Motorola (36%), amd (24%), freescale (20%), Sony(9%), amtel(9%), zilog(2%) though there are 29 suppliers. EM clustering algorithm used because Exception Maximization (EM) clustering algorithm is distance-based algorithm whose dataset modeling assumption is linear combination of multi variant normal distributions. This algorithm finds distribution parameters that maximize loglikelyhood.

This algorithm is chosen to cluster data because it can handle high dimensionality because of its linearity and converges fast having given appropriate initialization. There are certain cluster analysis situation with real world dataset where the need we to perform cluster analysis a small region of interest. EM clustering algorithm gives optimized results even compare to result given by K-Means algorithm.

The result obtained on applying FF clustering techniques. This technique generated two clusters classifying the cluster named amd (98%) and amtel (2%) though there are 29 suppliers. This technique is used because Farthest First is K means clustering algorithm variant that places each cluster centre in turn at the point furthest from the existing cluster centers. This point must be positioned inside the data part. This precisely means the steps in sequences as: Pick first center randomly, next is the point furthest from the first center, next is the point furthest from both previous centers and so on till convergence. This greatly speeds up the clustering because of the less need of relocation and tuning.

The result obtained on applying Filtered r clustering techniques. This technique generated two clusters classifying the cluster named amd(87%) and amtel(13%) though there are 29 suppliers.

With the help of weka tool, dataset of CPUPerformace are passed through different clustering Data Mining Techniques like simple K-Means, EM, Farthest First and FilteredClusterer methods.

In above section this research discuss about different supplier who gives best CPU performance based up on different criteria like CycleTimeNS, RAMMin, RAMMax, CacheMem, ChnlMin, ChnlMax, P_RPP.

For this analysis found that in simple K-Means techniques amd and zilog supplier gives best performance. EM methods Motorola, amd, freescale, sony, amtel, zilog suppliers gives best cpu performance. Farthest First method amd and amtel supplier gives best performance. FilteredClusterer method amd and amtel supplier gives best performance.

All the method's results are studied and this research gives amd is best supplier in terms of the performance. And second supplier is amtel who gives best performance.

## 6. CONCLUSION

The dimensions of Web Usage Mining are to extract and analyze pattern for the purpose of Knowledge discovery and ultimately pattern knowledge discovery.

Appropriate Web Mining Technique, Clustering Algorithm and dataset is prime challenge of survey findings. So this work proposes a Pattern Knowledge Discovery framework to solve the challenges in literature review.

The designed pattern Knowledge Discovery (PKD) framework which involves five stages of pattern Knowledge Discovery process- data acquisition (stage 1), five level pre-

processing (stage 2), processing (Stage 3), Evaluation and analysis (Stage 4) and Pattern Knowledge Discovery (Stage 5) can be used in variegated areas of applications of knowledge data discovery (KDD) process.

From the bundle of Data Mining Techniques – Classification, Regression, Association Rule Mining, Clustering etc. the suggested Cluster Data Mining algorithms and techniques can be availed to get the expected precise outcome from which various inferences are derived to perform the comparative evaluation and statistical analysis to obtain the mined pattern knowledge.

The implementation of the PKD framework and the results obtains after applying the various Data Mining Clustering Techniques using the open source Web Mining tool WEKA, illustrates that the above Web Usage Mining category of Web Data Mining, the five staged PKD framework and the proposed clustering techniques fulfils the criteria of the targeted Data Mining Knowledge Discovery process.

## 7. RESEARCH EXTENSION

The research work has used Web Usage Mining category of Web Mining which has a wide areas of applications on the web and which can be implemented in the other live and user specific domains. The dataset acquired from the Web Sources can also be obtained from live dataset sources from various industries, institutes , research and development organizations.

The proposed Pattern Knowledge Discovery (PKD) framework which focusing on clustering Data Mining Techniques can be further extended to design a versatile Data Mining Techniques implemented robust approach in a single integrated pattern Knowledge Discovery framework.

The Evaluation and analysis Stage can be extended to get a user friendly visual output leading to visual Data Mining and Knowledge Discovery for much better understanding of derived knowledge quantum.

The undertaken work has used WEKA tool, the set of similar and relevant tools can be used for achieving more optimized results comparatively.

The research work undertaken in timeframe in predefine objective and expected outcome will have left out issue and challenges encounter in research work. This issue and challenges will provide opportunity to extend this work beyond its summed up research.

## REFERENCES

1. Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, Khandakar Entenam Unayes Ahmed, “Pattern Discovery of Web Usage Mining”, 2009 International Conference on Computer Technology and Development, IEEE Computer society.
2. Mehdi Heydari, Raed Ali Helal, Khairil Imran Ghauth, “A Graph-Based Web Usage Mining Method Considering Client Side Data”, IEEE 2009 International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia.
3. Theint Aye, “Web log cleaning for mining of web usage patterns”, IEEE 2011.
4. Shuyan Bai, Qingtian Han, Qiming Liu, Xiaoyan Gao, “Research of an algorithm based on Web Usage Mining”, IEEE 2009.

5. Kavita Sharma, Gulshan Shrivastava, Vikash Kumar, "Web Mining: Today and Tomorrow", IEEE 2011.
6. Zhang Haiyang, "The Research of Web Mining in E-commerce", IEEE 2011.
7. FeiQiong Rong, " The Application of Web Usage Mining in Personalized Network Education", IEEE 2011
8. K. Suresh, R. MadanaMohana, A. RamaMohanReddy, A. Subrmanyam," Improved FCM Algorithm for Clustering on Web Usage Mining ", IEEE 2011
9. Lanjie Chen, Li Wei, "The Hot Research Topics and the Research Fronts in the Field of Web Data Mining (WDM) Based on Web of Science ", The 5th international conference on computer science & education Hefei, China. August 24-27 2010, IEEE 2010.
10. R. Manjusha, Dr. R. Ramachandran, "Web Mining Framework for Security in E-commerce", ICRTIT 2011, MIT, Anna University, Chennai, IEEE 2011
11. Yi Dong, Huiying Zhang, Linnan Jiao, "Research on Application of User Navigation Pattern Mining Recommendation", Proceeding of the 6th world congress on intelligent control and automation, June 21-23 2006, Dalian, China, IEEE 2006
12. Ching-Ming Chao, Shih-Yang Yaang, Po-Zung Chen, Chu-Hao Sun, "An Online Web Usage Mining System Using Stochastic Timed Petri Nets", Fourth international conference on Ubi-media computing, IEEE 2011
13. Kehar singh, Evolving limitations in K-Means algorithm in Data Mining and their removal, IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011, ISSN (Online): 2230-7893
14. Norwati Mustapha, Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems, European Journal of Scientific Research, ISSN 1450-216X Vol.32 No.4 (2009), pp.467-476
15. Reuben Evans, "Clustering for Clasification" From http://www.cs.waikato.ac.nz/~geoff/Thesis.pdf [access on 12/12/2012]
16. WEKA, "The University of Waikato", machine learning group, weka documentation, clusterers, From http://weka.sourceforge.net/doc.dev/weka/clusterers/FilteredClusterer.html [access on 22/12/2012]
17. M. Karthikeyan, M. Suriya Kumar and Dr. S. Karthikeyan, "A Literature Review on The Data Mining and Information Security", International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 1, 2012, pp. 141 - 146, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375
18. R. Manickam, D. Boominath and V. Bhuvaneswari,, "An Analysis of Data Mining: Past, Present and Future", International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 1, 2012, pp. 1 - 9, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375
19. R. Lakshman Naik, D. Ramesh and B. Manjula, "Instances Selection using Advance Data Mining Techniques", International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 2, 2012, pp. 47 - 53, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375