Enhanced Log Cleaner with User and Session based Clustering for Effective Log Analysis

Ms Shashi Sahu[1], Mr. Omprakash Dewanagan[2]

Abstract—***Now a day's World Wide Web growing fast in popularity. Internet is changing an important part of our life .Today internet has made people's life dependent on it, nearly everything and anything searched on internet. There are many billion of gif file, images and other multimedia files available on internet and number of information still rising. Web log usually contain large amount of irrelevant inconsistent data that is not completely remove by data cleaning method. There are lots of works in web usage mining, Web Usage Mining (WUM) is one among the most applications of data mining. There are types of issues relating with the present Web log mining. Existing web log mining algorithms suffer from difficulty to extract the useful data, overlapping during page ranking, misspellings during data entry. So in this research paper proposed Enhanced Enterprise Proxy Log cleaner Filtering Method. Enhanced log cleaner can filter the irrelevant inconsistent data, trace web requests from multiple clients to multiple web servers becomes much more effective and faster. Identify unique user and pages, who access the web site and which page is mostly hited. Experimental result of this paper reduces the sizes of data entry and finds valid information from web log. Filtering method can improve the data quality and efficiency of log file.***

Keywords— EP Log Cleaner; Data Pre-processing; Web usage mining; Data cleaning; Web mining; User Identification ;Session Identification.

I. INTRODUCTION

The past few years the internet has become the most important and most popular approach way of communication and data dissemination. It is a platform for exchanging many varieties of knowledge. The quantity of information available on the net is increasing quickly with the explosive growth of the Web, The Web has made the large document collection in the Web and the billions of users are seeking for data processing research. Web mining is that the application of data mining techniques to extract the knowledge from web log data, including Web log documents, hyperlinks between documents, usage logs of web sites, etc. The three kinds of information have to be handled in a web site: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is that to discover useful information from the content of a web page [Losarwar V. et al. 2012]. Enhanced EP Log Cleaner is that filtering the irrelevant items based on different filter. Mining web server log plays an important role for user and server. In log file sizes of data is increasing rapidly, which makes it difficult to identify the "right" or "interesting" information today. Web Usage Mining consists of different four steps:-

- Data Collection
- Data Preprocessing
- Pattern Discovery
- Pattern Analysis

Data Collection : Users log data is collected from various sources like server side, client side, proxy servers and so on.

Preprocessing Pre-processing is a very important step because of the complicated nature of the Web architecture that takes 80% in mining method [Pabarskaite Z. 2002].

Pattern discovery : Application of varied data mining techniques to processed information like statistical analysis, association, clustering, pattern matching and so on.

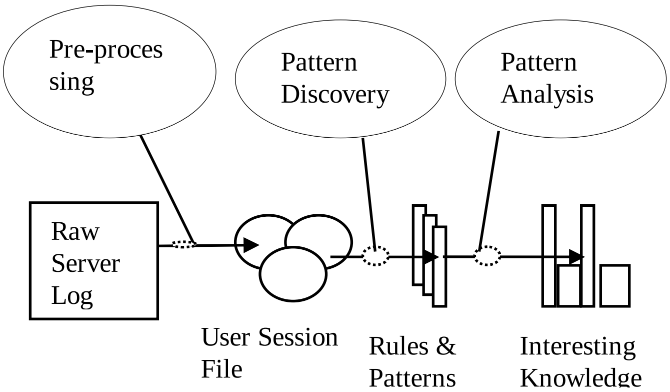

Fig 1.1 Process of Web Usage Mining

The rest of the paper is organized as follows: - In section 2, Different sources to collect the data. Section 3, we are present a review existing techniques. Section 4 Problem identification, Section 5, explains the proposed methodology of filtering the log mining. Section 6 describes the Result and

Manuscript received Feb, 2013.

***Ms. Shashi Sahu,** Computer Science & Engineering, RCET Bhilai (Chhattisgarh) India.*
***Mr. Omprakash Dewangan,**, Computer Science & Engineering, RCET Bhilai,(Chhattisgarh) India.*

discussion of methodology. Section 7 conclusion and future work of the paper.

II. Data Collection

This section discusses the types of web data that is use for web analysis. Data source will be collected at the server-side, client-side, proxy servers, or acquired from an organization's database, that contains from the business information or consolidated Web information [Tyagi A. et al. 2010].

- Server side
- Client side
- Proxy servers

2.1 Server Side Collection

Server log files are in relationship of many to one. Many users might visit one particular web site and user behavior about that specific web site can be captured accordingly. Web servers are richest and the most typical source of data. They will collect large amounts of knowledge in log files. These logs sometimes contain basic information e.g.: name and Information, Processing of the remote host, date and time of the request, the request line came from the client, etc. Web logs files are computer files contains requests addressed to web servers. These are recorded in a chronological order. The most popular log file reformats the original file logged.

```
222.126.107.93 - - [14/Dec/2007:21:54:30 -0800] "GET
/img/tm_sep.gif HTTP/1.1" 200 412
```

Fig.1.2: Common Server log entries

Fig 1.2 shows that a user IP address 222.126.107.93 successfully requested the page on Dec 14, 2007 at 21:54:30. The HTTP method the remote user used "GET". There are many of bytes returned to the user that is 412. This line consist the following fields. Client IP address, User id Access time, HTTP request Method, Path of the resource on the Web server, Protocol used for the transmission.

2.2 Client Side Collection

It is in the form of one-to-several relationships of client and web sites visited by the specific user. All pages are of not same importance and number of significant pages for user profile can be taken by formula. Eq. 1 [Hussain T et al. 2010].

$$S = 1 + 2 * \log(n) \qquad (1)$$

2.3 Proxy Server side collection

Proxy server log files are most complex and more vulnerable to user access data in log file. Proxy server logs contain the Hyper Transfer Text Protocol requests from multiple clients to many Web servers. This may serve as a knowledge supply to find the usage pattern of a group.

III. Literature Review

Literature review is the most important step in development process. Different method has removing the inconsistency irrelevant data of a server log in web page mining.

Reddy K. et al (2013) has proposed several data preparation techniques of access stream even before the mining process can be started and these are used to improve the performance of the data preprocessing to identify the unique sessions and unique users. The paper is concluded by proposing the future research directions in this space.

Shaa H. et al (2013) has proposed EP Log Cleaner Method. EP Log Cleaner that can filter out plenty of irrelevant, unwanted data based on the common prefix of their URLs. This method is improving data quality by removing the irrelevant items such as jpeg, gif files, sound files and web request with a wrong http status code.

Tyagi N.et al (2010) provides an algorithmic approach to data preprocessing in web usage mining. They take requests for graphical page content, or any other file which may be induced into web page, or navigation sessions performed by robots. Though they discussed the importance of proxy log, the data cleaning method they used is quite simple.

Zheng L.et al (2010) has proposed Optimized User Identification, Optimized Session Identification. The strategy based on the referred web page is adopted at the stage of user identification. Experiments have proved that advanced data preprocessing technology can enhance the quality of data pre-processing results.

Theint (2011) has proposed data mining techniques to discover user access patterns from web log. This paper mainly focus on data preprocessing stage of the first phase of web usage mining with activities like field extraction and data cleaning algorithms. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data.

Hui Lee C.et al (2011) has proposed an efficient prediction model, two-level prediction model (TLPM). The experiment results prove that TLPM can highly enhance the performance of prediction when the number of web pages is increasing.

Losarwar V.et al (2012) has discussed the importance of data preprocessing methods and various steps involved in getting the required content effectively. A complete preprocessing technique is being proposed to preprocess the web log for extraction of user patterns.

IV. Problem Identification

After going through many research paper which have been done on the field of web log mining found problems. Web log are generally noisy and ambiguous. Web log file is saved as text (.txt) file. Due to large amount of "irrelevant information" within the sever log, the original log file cannot be directly usage in the web usage mining (WUM) procedure. World day by day variety of web sites and its users are increasing quickly. Problem to finding the unique user and unique pages, which pages is mostly hited. There are a lot of works on data cleaning of web logs, it consist inconsistent irrelevant items and useless data that can not completely remove. So problem are arises, difficulty in specifying the "right" and "interesting" data from the log file with unbounded accesses to websites, web requested by many clients to many web servers and also overlapping during page ranking, When multiple knowledge sources need to be integrated, data quality issues are present in single data collections, like files and database of log, e.g., due to misspellings during records entry.

V. METHODOLOGY

4.1 Proposed Enhanced EP Log Cleaner Filtering Methods

Enhanced EPLogCleaner that can filter out irrelevant, inconsistent items based standard filter and common prefixes . Mining web server log plays an important role for user and server. The size of web log is increasing rapidly, which makes it difficult to find the "right" or "interesting" information today. We build an evaluation of Enhanced EPLogCleaner with web log data. Filtering method can improve data quality [Shaa H. et al. 2013]; After Filtering method reduces the records of data. So file size are also resizes. It uses different filter for removing the multimedia system dat. It is also work for how many unique users are available in huge amount of data set and unique pages are mostly hited in web site. There are different steps consist in filtering method for removing the irrelevant data of records.

- Standard Filter
- Date Filter
- Status Filter
- Unique User's and Pages
- Session Filter
- Prefix Filter

4.1.1 Uses Regular Expression matching algorithm

Uses of Regular Expression matching algorithm for filtering of irrelevant data from web log. Regular expression may be a string of characters that defines an extract pattern. You commonly use a regular expression to search text for a group of words that matches the pattern. Regular expressions are a notation for describing patterns of text, in result a special-purpose language for pattern matching.

```
StartIndex = regexp(str,expression)
[startIndex,endIndex] = regexp(str,expression)
```

//Algorithm: Filtering method//

Input- Original log file from web server

Output- Filtered log

Step 1 Input original log file

Step 2 Call Standard filter algorithms for eliminating the multimedia entries.

Step 3 Input this reduced log file into Day Standard filter algorithm and eliminate date and time is not in the range.

Step 4 Input the reduce data in Day_Standard log in Session filter algorithm

Step 5 Compute the number of unique user versus unique pages

Step 6 Compute the session per user matrix

Step 7 Input the process file in prefix _URL filtration

Step 8 Log file as a final compressed and filtered log file.

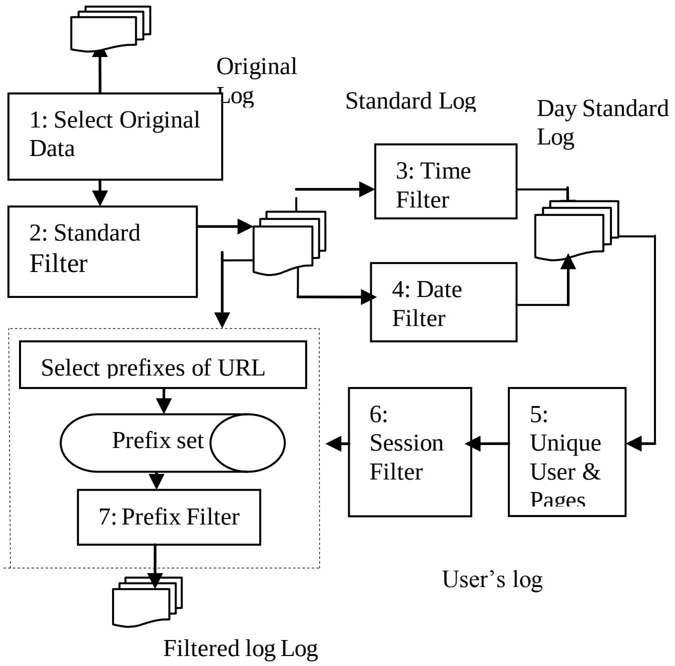

Fig 5.1 Processing of Proposed Filtering Method

4.1.2 Steps by Steps Processing of Filtering Method

There are different steps perform in this method. Filtering method filters the noisy data. Noisy data are available on server log file. After filtering the records reduces the data and also log sizes. In this method there are different steps are applied for valid information of data.

Step 1 Input the original log from web server log file

Firstly select the original log file from web log. The log file contains the irrelevant data and also included different column for like IP address, Date and time column, method, page column and sizes of files.

Step 2 Standard Filter

The first step is named as Standard Filter filters where some types of irrelevant items, such as multimedia data files and error access knowledge, http request method of log file. For example jpg, jpeg or "gif file, access URLs not using GET method, Hyper Text Transfer Protocol error standing code 400 which indicates bad request. The remaining items are stored into the "Standard Log" File.

// Algorithm: Standard Filter//

```
Input- Original log file from web server log
Output - Standard log
Step 1 Input the log file
Step 2 Matching Expression == 'html', 'jpg', 'gif' or expression == 'php' 'jpeg' from the log column entries
Step 3 Declare the matching column as page column
Step 4 Matching Expression == 'GET' from the log column entries
Step 5 Declare the matching column as method column
Step 6 From the page column calling String finding algorithm for multimedia entries and making find indices as true
Step 7 if index of row found true making entries as avoid entry
Step 8 end
```

. In log file there are different irrelevant multimedia data are available. So reduce the noisy records of multimedia data from log file. We match the expression of data like expression= "html", "gif" file, after filtering valid data is in Standard log.

Step 3 Time Filter

Time Filter, aims to discard the untraceable requests that are generated by computers, without human operations in a long time of midnight and also Software updating automatic requests at any time

Step 4 Date Filter

Date filter work for filtering date is not match in particular range of a date. After this step, all the automatic requests occurring at night in the "Standard Log" are filtered out, and the remaining items are stored into the "Day Standard Log" File.

Step 5 Unique User's and Pages

This step identifies the how many unique users and unique pages are in log file entries and which page is mostly hited. It is finding by session filter. User interesting pages are finding in these steps and find who visited the web site.

Step 6 Session Filter

If session in the range for each unique IP address, then delete the entering otherwise not. A session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period

// Algorithm Session Filter//

```
Input = Log file
Output= Unique user and pages in User's log
Step 1 Finding the ip column, date column, Time column using matching expression from the log entry
Step 2   Find the each unique user extracting unique pages
Step 3   Counting the time differences among the entries
Step 4   if  Time differences ≥Threshold time
               Session = session+1
           else
               Session =1
Step 5   end
```

In this step we will perform user session page matrix. We will consider the IP column, page column date column and time column through regular expression matching algorithm. Find the user and session at a particular time period.

Step 7 Prefix Filter

To address problem, we use the prefixes of URLs accessed at night as the filtering rules in this step, because there are usually the same type of resources with the same prefix. Firstly, we choose the path which is the part before the last slash in a URL as its prefix [Shaa H. et al. 2013]. It is because that there are always multiple URLs under one path

VI. Results and discussion

In this paper, we used the data generated by the NASA log file. Selecting useful data is an important task in filtering method. We have evaluated the performance of proposed algorithm by using the same synthetic dataset. We take log file of NASA space center from web server. The log data was collected from 00:00:01 July 1, to 00:06:02 July 4, 1995. Firstly we need to analysis the server log data file, we take some entries of server log for filtering the log records.

6.1 Process and Results

To achieve relevant information and improve data quality of web logs. Enhanced EPLogCleaner is improving the quality of log data by further filtering URL requests. Removing the records of irrelevant data apply Regular Expression matching algorithm.

- There are resize the size of original log file so reduce the time and cost and also gets which page is mostly hited by user and also identifies who access the page.
- Original log versus filtered log of result
- Original log versus number of frequent pages

6.2 Performance and Evaluation

We have evaluated the performance of proposed algorithm by using the data set. The dataset is collected from NASA log server. Dividing the log of records in different log file where number of data set collected the records. We take the data set 1, data set 2 consists of 35000, 55000 records in the log file. Firstly we select the records of data from original log file. The entries of log file are 55000, after then apply standard filtering process. Standard filter remove the multimedia entries of data like gif file, images, 20487 in standard log. We take input as standard log. This log file contains 20487 entries of data after filtering, filtered data are 7736 in day standard log. After filtering we get unique user are 1842 and unique pages 1288 in user's log and After this method gets the result high accuracy and more reliable. It is clearly that the number of data can be reduced by more than 85% in the last step.

Table 6.1 Number of Records resulted after Filtering Method

SR.	Steps of Filtering Method	Reduces the log of records	
		Data Set 1	Data Set 2
1	Original log	35000	55000
2	Standard log	13952	20487
3	Day Standard log	4435	7736
4	Unique user	972	1842
5	Prefix log entries	950	1288

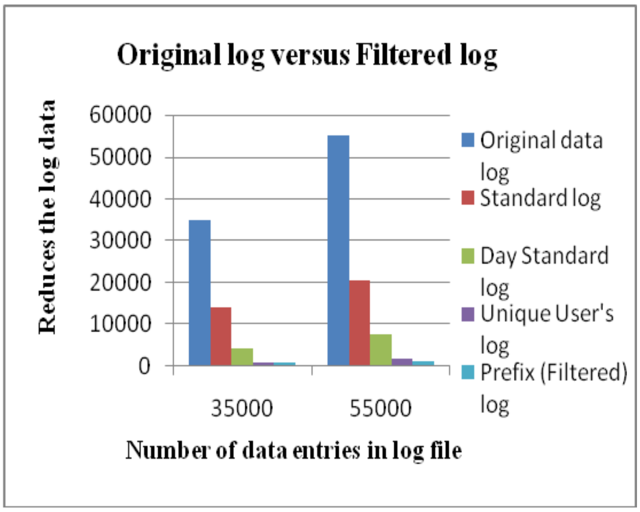

Fig.6.1: Reduces the Original log data

Table 6.2 Number of reduces the log size of log file after filtering method

SR.	Steps of Filtering Method	Reduces the size of data			
		Data Set 1	Compres_ratio (%)	Data Set 2	Compres_ratio (%0
1	Original log in KB	2184	------	3366	------
2	Standard log	870.60	60.13	1253.80	62.75
3	Day Standard log	276.74	87.37	473.44	85.93
4	Unique user	60.65	97.22	112.73	96.65
5	Prefix log entries	59.28	97.28	78.82	97.65

Table 6.2 show reduces the data sizes after flitering the log data. Where log data are reduced then log sizes are also reduced.

6.4 Comparision between methods

EP Log Cleaner filters the plenty irrelevant data of URL request. This method basically focuses on prefix of URL request of night that is automatic generate. Experimental result of this paper is filtered the 30% URL request. This method is not performing who access the page, which page is mostly hited. But Enhanced EP Log Cleaner Enhanced EP Log Cleaner basically work for removing the multimedia data, automatic request of night without human operation is filtered by day standard filter. This method is identifying unique user and unique pages with session filter, which page is mostly hited and generates the valid information and also performs the URL request. This method mainly works for efficiently reduces the sizes and gets valid information. Experimental result shows that more than 85% compresses the data at last step. Results are more accurate and reliable. Compression ratio is high. So all this observation this method is best than existing method.

VII. Conclusion and Future Work

Web log data is collection of large amount of irrelevant, incorrect data. Many interesting knowledge are obtained in the web log. But it is very complicated to extract the information without filtering phase. Log data is very irrelevant so filtering process is efficient for web usage mining process. Enhanced EP Log Cleaner filtering method can filter irrelevant, noisy data of server log. This method reduces the records after filtering and resize of log file after reducing the records. Enhanced log cleaner method more than 85% compresses the original log sizes of data. This method finds the unique user and pages and also identifies who access the web page of log file. Filtering method is improving the data quality and efficiency of log file. Results of a web log mining can be used for various application like as web personalization, site recommendation, site improvement etc. Enhanced method improves the relevancy of the pages and thus reduces the time user spends in seeking the required information and also improves the quality and efficiency of data.

REFERENCES

[1] Agarwal R., Arya K.V., Shekhar S. and Kumar R. 2011. An Efficient Weighted Algorithm for Web Information Retrieval System, IEEE.

[2] Alassi D. and Alhajj R. 2012. Effectiveness of template detection on noise reduction and website summarization, Elsevier.

[3] Cooley R. 2000. Web Usage Mining: Discovery and Application of Interesting

[4] Patterns from Web Data. Phd Thesis, Department of Computer Science, University of Minnesota, May.

[5] Cooley R., Mobasher B., and Srivastava J. 1999a. Automatic Personalization Based on Web Usage Mining, Technical Report.

[6] Etzioni O., 1996. The World Wide Web: Quagmire or Gold Mine, Communication of ACM, 39(11): p. 65-68

[7] Forsati R., Mohammad, Meybodi R. and Rahbar A. 2009. An Efficient Algorithm for Web Recommendation Systems, IEEE.

[8] Hussain, T., and Asghar, S., and Masood, N. 2010. "Web Usage Mining: A Survey on Preprocessing Of Web Log File", In Proceedings of: International Conference on Information and Emerging Technologies (ICIET), pp-1-6.

[9] Losarwar V. and Joshi Dr. M. 2012. Data Preprocessing in Web Usage Mining, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, Singapore.

[10] Maheswari B. U. and Sumathi Dr. P. 2014. A New Clustering and Preprocessing for Web Log Mining, IEEE.

[11] Munk M., Kapustaa J. and Šveca P. 2012. Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor, International Conference on Computational Science, ICCS 2011 Procedia Computer Science.

[12] Nithya P. and Sumathi Dr. P. 2012. Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots, IEEE National Conference on Computing and Communication Systems (NCCCS)

[13] Pabarskaite,Z. 2002. Implementing advanced cleaning and end–user interpretability technologies in Web log mining, in Proc. of the 24th International Conference on Information Technology Interfaces (IEEE Cat. No.02EX534). 109–113.

[14] Pabarskaite Z., Raudys A. 2002. Advances in Web usage mining, in 6th WorldMulticonference on Systemics, Cybernetics and Informatics. 11(2): 508–512.

[15] Reddy K. S., Reddy M. K., and Sitaramulu V. 2013. An effective Data Preprocessing method for Web Usage Mining, IEEE.

[16] Shaa H., Liub T., Qinb P., Sunb Y. and Liub Q. 2013. EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining, Information Technology and Quantitative Management, ITQM Procedia Computer Science 17.

[17] Singh Yumnam A., Sreeram Y. C., and Shaik A. N. 2014. Overview: Weblog Mining, Privacy Issues and Application of Web Log Mining, IEEE 978-93-80544-12-0/14/$31.00_c.

[18] Soundarya M. and Balakrishnan R. 2014. Survey on Classification Techniques in Data mining, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7.

[19] Sujhatha V. and Punithavalli 2012. Improved user navigation pattern Prediction technique from web log data, International Conference on Communication Technology and System Design 2011 Procedia Engineering 30, 92.

[20] Taucher L. and Greenberg S. 1997. Revisitation patterns in World Wide Web navigation, Proc. of Int. Conf. CHI'97, Atlanta.

[21] Theint Aye T. 2011. Web Log Cleaning for Mining of Web Usage Patterns, IEEE.

[22] Tyagi N., A. Solanki and Tyagi S. 2010. An Algorithmic Approach to Data Preprocessing in Web Usage Mining, International Journal of Information Technology and Knowledge Management 2 (2) 279–283.

[23] Valera M.and Rathod K. 2013. A Novel Approach of Mining Frequent Sequential Pattern from Customized Web Log Preprocessing, International Journal of Engineering Research and Applications (IJERA) Vol. 3, Issue 1, pp.269-380 369.

[24] Yuhefizar, S. Budi, P. I. Ketut Eddy and S. Yoon K. 2014. Two Level Clustering Approach for Data Quality Improvement in Web Usage Mining, Journal of Theoretical and Applied Information Technology 20th April.

[25] Wang Y. 2000. Web Mining and Knowledge Discovery of Usage Patterns, CS 748T Project.

[26] Zheng L., Gui H. and Li F. 2010. Optimized Data Preprocessing Technology for Web Log Mining, International Conference On Computer Design And Applications (ICCDA).

Author Profile

Ms. Shashi Sahu received the B.E. degree from Chhattisgarh Swami Vivekanand Technical University, Bhilai (C.G.) India in Computer Science & Engineering in the year 2013. She is currently pursuing M.Tech. Degree in Computer Science Engineering with specialization in Software Engineering from CSVTU Bhilai (C.G.), India. Her research area includes Data Mining.

Mr.Omprakash Dewangan is currently Reader in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. He completed his MCA and M.Tech. in Computer Science and Engineering Branch. His research area includes Image processing, Data Mining etc. He has published many Research Papers in various reputed National & International Journals, Conferences, and Seminars.