International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-2 Issue-4, March 2013

An Overview of Preprocessing on Web Log Data for Web Usage Analysis

Naga Lakshmi, Raja Sekhara Rao , Sai Satyanarayana Reddy

Abstract-- ***Web has been growing as a dominant platform for retrieving information and discovering knowledge from web data. Web data is stored in web server log files. Web usage analysis or web usage mining or web log mining or click stream analysis is the process of extracting useful knowledge from web server logs, database logs, user queries, client side cookies and user profiles in order to analyze web users' behavior. Web usage analysis requires data abstraction for pattern discovery. This data abstraction can be achieved through data preprocessing. This paper presents different formats of web server log files and how web server log data is preprocesses for web usage analysis.***

Keywords: ***Web server logs, Web usage analysis, preprocessing, data cleaning, user identification, session identification, path completion, pattern discovery, pattern analysis.***

I. INTRODUCTION

World Wide Web is a huge, interconnected, semi-structured, widely distributed, highly heterogeneous and hypertext information repository. The Web continues to grow at an incredible rate as information gateway. Web mining technologies are the proper solutions for knowledge discovery on the Web. Web mining is the application of data mining techniques to discover patterns from the Web. Web mining can be classified into three different types, which are Web content mining, Web structure mining and Web usage mining. Web content mining is the process of extracting and integration of useful data, information and knowledge from Web page contents. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. Web Usage Mining is a part of web mining which deals with the extraction of interesting knowledge from log files produced by web server. Web Usage Mining is also called Web Log Mining or Web Usage Analysis or Click Stream Analysis. The term Web usage mining was introduced by Cooley in 1997 [1] when a first attempt of taxonomy of Web Mining was done in particular that they define Web mining " as the discovery and analysis of useful information from the World Wide Web". Also it is defined as "the application of data mining techniques to large Web data repositories [2].This type of web mining allows for the collection of Web access data for Web pages.

Manuscript published on 30 March 2013.
*****Correspondence Author(s)
Naga Lakshmi Theerthala, Assistant Professor, Department of Information Technology, Usha Rama College of Engineering, and Technology, Telaprolu, Unguturu Mandal, Krishna District, Andhra Pradesh, India.
Dr. Raja Sekhara Rao Kurra, Professor and Dean., Department of Computer Science and Engineering, KL University, Vaddeswaram, Guntur District, Andhra Pradesh, India.
Dr. Sai Satyanarayana Reddy Seelam, Professor and Head of the Computer Science and Engineering Department, Lakireddy Balireddy College of Engineering, Mylavaram, Andhra Pradesh, India.

This usage data provides the paths leading to accessed Web pages. This data is often gathered automatically into access logs by means of the Web server. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs. The taxonomy of Web Mining is shown in Figure 1 below:

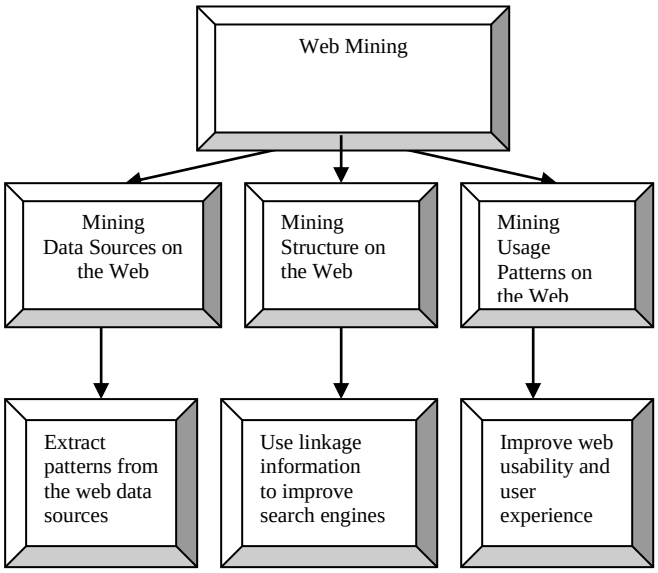

Figure 1: Web Mining Taxonomy

II. CLASSIFICATION OF WEB DATA

Web data is classified as follows:

A. Content data

Content data deals with any complete representation of the resource such as HTML documents, images, sound files etc.

B. Structure data

Structure data deals with data describing the structure and the organization of the content through internal tags or hyper-links.

C. Usage data

Usage data deals with collection of available data describing the usage of Web resources.

D. User profile data

Demographic information derived from registration.

III. SOURCES OF DATA FOR WEB USAGE MINING

Data which is used for web usage mining can be collected at three different levels [3]:

A. Server side

Web servers are the common source of data. They store large amounts of information in their log files. These logs generally contain basic information e.g. name and IP of the remote host, date and time of the request etc. The web server stores data regarding request performed by the client. Data can be collected from multiple users on single site. All the click streams are recorded into the web server log file.

B. Client side

It is the client itself which sends information to a repository regarding the users' behavior. This is done either with an ad-hoc browsing application or through client side application running standard browsers. Client level data collection can be implemented by using a remote agent (such as Java applets or Java Scripts).

C. Proxy side

Information about user behavior is stored at proxy level, thus web data is collected from multiple users on several websites, but only users whose web clients pass through the proxy. Proxy servers collect data of groups of users accessing huge groups of web servers.

Proxy level collection is an intermediary between server level and client level. The page load time gets reduce by proxy server, so user experience high performance. In this paper, we have covered only the case of a Web Server (HTTP server) data.

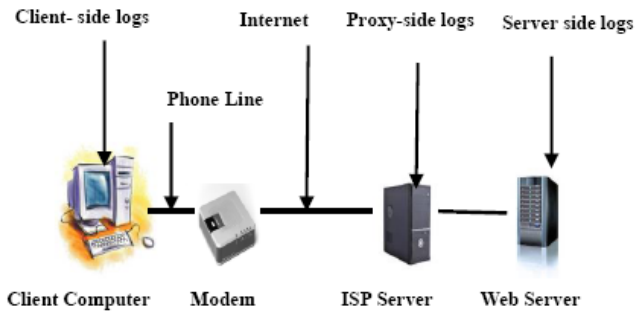

Figure 2: Various Data Sources for Web Usage Mining

IV. WEB LOG FILE STRUCTURE

This paper mainly focuses on analyzation of web server log files. Web server log files are the primary data sources used in web usage mining. These are plain text (ASCII) files. Web server log files store click stream data which can be useful for mining purposes. The data is stored as a result of user's interaction with a website. Web Server log files are records of web server activities. They give information about users file requests to a web server and the server response to those requests. A sample line of a web log file in raw format is shown in Figure 3 below:

217.13.12.211 - - [11/Nov/2011:05:45:26 -0500] "GET

/meta_tags.htm HTTP/1.0 "200 26150

"http://www.google.com/search?q=meta+and+tag"

[Mozilla/11.0(compatible; MSIE 8.0: Windows XP; DigExt)]

Figure 3: Raw Log file of a Web Server

The following line in web server raw log file tells us:

Visitor's IP address or hostname [217.13.12.211]

Login [-]

Authuser [-]

Date and time [11/Nov/2011:05:45:26 -0500]

Request method [GET]

Request path [meta_tags.htm]

Request protocol [HTTP/1.0]

Response status [200]

Response content size [26150]

Referrer path [http://www.google.com/search? q=meta+ and+ tag]

User agent [Mozilla/11.0(compatible; MSIE 8.0; Windows XP; DigExt)]

A. Types of Web log file formats

There are three kinds of log file formats to record log files. They are as follows:

i). Common Log Format (CLF)

This is the most common and standardized text format of a web server log file. This can be produced by several web servers and read by variety of log analysis programs.

The log file entries produced in CLF appear as follows

---

host/ip rfcname logname [DD/MMM/YYYY:HH:MM:SS-0000]
"METHOD/PATH HTTP/1.0" code bytes

Sample line of a Common Log Format

127.0.0.1 - john [12/nov/2011:12:53:46-0700]
"GET/apache_pb.gif HTTP/1.0" 200 2326

Each part of this log entry is described below

127.0.0.1(%h)

This is the IP address of the client (remote host) which made the request to the server.

- (%1)

The "hyphen" in the output indicates that the requested piece of information is not available.

john (%u)

This is the user id of the person requesting the document as determined by HTTP authentication.

[12/nov/2011:12:53:46-0700]

The time which the server finishes processing the request.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The format is as follows:

[day/month/year: hour: minute: second zone]

day = 2*digit
month = 3*letter
year = 4*digit
hour = 2*digit
minute = 2*digit
second = 2*digit
zone = (`+' | `-') 4*digit

"GET/apache_pb.gif HTTP/1.0" (\ "%r\")

The request line from the client is given in double quotes. The request line contains a great deal of useful information. First, the method used by the client is GET. Second, the client requested the resource /apache_pb.gif, and third, the client used the protocol HTTP/1.0

200 (% >s)

This is the status code that the server sends back to the client.

2326(%b)

The last entry indicates the size of the object returned to the client, not including the response headers.

W3C (World Wide Web Consortium) Extended log file format

This is the default log file format used by IIS. It uses ASCII text format and the time recorded as UTC (Greenwich Mean Time). This is the customizable format. Figure.3 shows sample lines in a W3C Extended log file format with the following fields:

Time, Client IP Address, Method, URI Stem, Protocol Status, and Protocol Version.

#Software: Microsoft Internet Information Service 5.1
#Version: 1.0
#Date: 2011-11-11 14:35:15
#Fields: time c-ip cs-method cs-uri-stem sc-status cs-version
6:32:15 172.16.255.255 GET/default.htm 200 HTTP/1.0

Figure 4: W3C Extended log file format

ii). Microsoft IIS (Internet Information Services) log file format

Microsoft IIS log file format is a non-customizable ASCII format used to record more information than the NCSA Common format but less than the W3C format. It uses comma to separate fields and uses the local time. It includes the user's IP address, user name, request date and time, Service status code and number of bytes received, the elapsed time, the number of bytes sent, the action (for example, a download carried out by a GET command) and the target file. Figure 5 shows sample lines in an IIS log file format

192.168.114.201, —, 11/25/2011, 9:45:25,

W3SVC2, SALES1, 192.168.114.201, 4504,163, 3223,200, 0, GET, / SalesDeptLogo.gif, —,172.16.255.255, anonymous, 11/25/2011, 23:58:11, MSFTPSVC, SALES1,192.168.114.201, 60, 275, 0, 0, 0, PASS, /introduction.htm, —,

Figure 5: IIS log file format

iii). NCSA Common log file format

NCSA (National Centre for Supercomputing Applications Common format is a fixed (non-customizable) ASCII format. It does not support FTP sites. Since the entries are small with this format, the storage space required for logging is less compared to other formats. It logs the basic information about user requests such as remote host name, user name, date, time, request type, HTTP status code, and the number of bytes sent by the server. It records the time by using the local time and fields are separated by spaces. Figure 5 shows sample lines in a NCSA log file format with the following fields

172.21.13.45- REDMOND\sam [11/11/2011:25:28:06 -0800] "GET scripts/iisadmin/ism.dll?http/serv HTTP/1.0" 200 3401

Figure 6: NCSA log file format

VI. TYPES OF WEB SERVER LOG FILES

Web server log files comprise access logs, referrer logs, agent logs and error logs.

A. Access logs

The data from Access Logs provides an extensive view of a Web server's and users. Such analysis enables server administrators and decision makers to characterize the users and usage patterns. Access Logs are also called Transfer Logs. It stores information about which files are requested from web server. The Access log format is shown in Figure

7 below.

Client IP address	User ID	Access time	HTTP request method	Protocol Used for the transmission	Status of the request	Amount of data transferred

Figure 7: Access log format

Sample Access log entry

123.45.6.78.9 - [11/May/2011:04:05:45 -0500] "GET/HTTP/1.0" 200 3250

This line consist the following fields

123.45.6.78.9

This is an IP address of the client.

—

This is user id field, here '—'(hyphen) represents anonymous user id.

[11/May/2011:04:05:45 -0500]

This is an access time of the web page.

-0500

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

This is time zone.

GET/HTTP/1.0

Date	Time	Error	IP Address	Error Message

This represents the HTTP request method and protocol used for the transmission

200

This is status code returned by the server.

3250

This is number of bytes transmitted.

B. Referrer logs

It stores information of the URLs of web pages on other sites that link to web pages. That is, if a user gets to one of the server's pages by clicking on a link from another site, the URL of that site will appear in this log. The Referrer log format is given below in Figure 8.

Date	Time	Time Zone	Referrer URL

Figure 8: Referrer log format

Sample Referrer log entry

The following is an example of a record in a Referrer log:

[Wed May 11 17:35:45 2011+0500]
"http://www.ibm.com/index.html"

C. Agent logs

The Agent log provides information on a user's browser including browser version and operating system. It records information about the web clients that sends requests to web server. This is the major information, as the type of browser and the platform determines what a user is able to access on a web site. The Agent log format is given in Figure 9 below.

Date	Time	Time Zone	Version Number	Platform

Figure 9: Agent log format

Sample Agent log entry

The following is an example of a record from an Agent log:

[10/Nov/2011:19:15:06+0500] "Microsoft Internet Explorer – 5.0"

D. Error logs

The Error Log provides the time, domain name of the user, and page on which a user received the error to a server administrator. These error messages inform server administrators of erroneous links on their servers. It stores information about errors and failed requests of the web server. The Error log format is given below in Figure 10.

Figure 10: Error log format

Sample Error log entry

[Wed May 11 17:35:45 2011] [error] [client 132.1.0.1] client denied by server:/export/home/live/ap/htdocs/testdoc

VII. WEB USAGE ANALYSIS

In technical point of view, Web usage analysis is the application of data mining techniques to usage logs or access logs of data repositories to discover or analyze user access patterns from web server log files. The purpose of it is to generate an outcome that can be used to improve and optimize the content of a site [4]. The process of web usage analysis focuses on web usage data or user access data. User's browsing behavior can be captured by Web usage data from web site [5].

Structure of Web Usage Analysis

There are three main phases of Web Usage Mining or Web Usage Analysis. They are Data preprocessing, Pattern discovery and Pattern analysis. This section presents an overview of these phases. In our perspective, the usage data is access logs on server side which keeps information about user navigation. Various phases of Web Usage Analysis is shown in Figure 11 below:

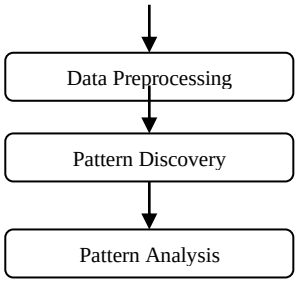

Figure 11: Basic Structure of Web Usage Analysis

A. Data Preprocessing

The raw web log data is generally diverse, incomplete, inconsistent, noisy and difficult to be used directly for pattern mining. The quality data gives quality output. The attributes of the quality data includes accuracy, completeness, consistency, accessibility, and timeliness. In order to obtain the quality data we have to preprocess it. The Data preprocessing phase includes Data cleaning, User Identification, Session Identification, Path completion and Transaction identification. Various steps involved in data preprocessing phase is shown through the Figure 11 below:

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

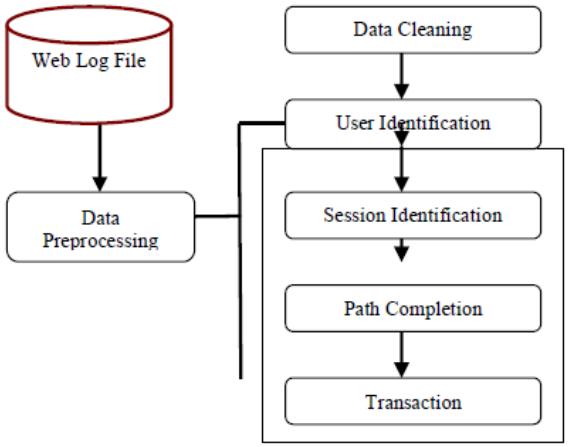

Figure 12: Various Steps of Data Preprocessing

Data Cleaning

In this paper we use web server logs in Common log format. In data cleaning phase irrelevant or redundant information like image, video and sound files which could be downloaded without an explicit user request can be removed. Other removal information includes HTTP errors, records created by spiders, crawlers and robots.

User Identification

User identification is to discover who access web site and which pages are accessed. IP address, User agents and referring URL fields of log files are used to identify user. ISP's (Internet Service Providers) uses DHCP (Dynamic Host Configuration Protocol), a TCP/IP standard that enables the centralized management of IP addresses for client computers on a network. The typical problems encountered in identification of the user are [6]:

i). Single IP address/multiple server session

It is complicated to identify same user through different TCP/IP connections because IP address changes dynamically.

ii). Multiple ID address/single user

IP address of a user changes from connection to connection.

iii) . Multiple IP address/single server session

Different IP addresses can be assigned for every single request performed by the user.

iv). Multiple agent/single users

Same user can access the Web by using different web browsers from the same host.

Session Identification

A session is a sequence of web pages user browse in a single access. Session identification is to discover different user sessions from the web access log. The goal of session identification is to divide the page accesses of each user into individual sessions [7]. Log entries of the same user are divided in to sessions or visits. Normally a time out of 30 minutes between sequential requests from the same user it taken in order to close a session.

Path Completion

This is important and difficult phase because it involves the use of referring URLs and site topology. Path completion is used to obtain the complete user access path. The incomplete access path of every user session is recognized based on user session identification. There are chances of missing pages after constructing transactions due to proxy servers and caching problems [8] [9]. However by examining the site topology and the referrer field it is possible to rebuild the path followed by the user. At the end of this stage the user session file is ready.

Transaction Identification

The goal of transaction identification is creating significant clusters of references for each user. After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. So, the transaction identification is done by merge or divides approaches. An input for both types of approaches is a transaction list and some parameters and output is a transaction list that has been operated on by the function in the module in the same format as the input**.**

B. Pattern Discovery

Pattern discovery deals with extracting information from preprocessed data. There are several techniques that are deduced from different fields such as data mining, statistics, machine learning and pattern recognition are applied to web usage data to discover user access patterns of the web. Statistical Analysis tools can be used to give a description of the traffic on a web site e.g. most visited pages, average daily hits etc. Association Rules [10] consider every URL requested by a user in a visit as item and find out relationships between them with a minimum support level. Sequential Patterns [11] are used to discover time ordered sequence of URL's followed by past users in order to predict future ones. Clustering [12, 13] forms meaningful clusters of URL's by discovering similar attributes between them according to user behavior.

C. Pattern Analysis

This is the final stage of Web Usage Analysis. The goal of this process is to extract the interesting patterns from the output of the pattern discovery process by eliminating the irrelative patterns. Pattern Analysis involves the validation and interpretation of the mined patterns. Validation can be used to remove the irrelative patterns and to extract the interesting patterns from the output of the pattern discovery process. The output of mining algorithms is in mathematic form and not suitable for direct human interpretations. So, Visualization techniques are used to interpret the results. The most general ways of analyzing user access patterns are either by using a knowledge query mechanism on a database such as SQL or data cubes to perform OLAP operations. Visualization techniques, such as graphing patterns are used for an easier interpretation of the results.

VIII. EXPERIMENTAL RESULT

We have made an experiment to prove the efficiency of our methodology mentioned above, with the web server log of the central library of Lakireddy Balireddy College of Engineering.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The original data source of our experiment is from February 20th, 2013 to February 28th, 2013, which size is 150 MB. Our experiments were performed on a 2.8 GHz Pentium IV CPU, 120 GB RAM, Windows XP, Oracle 9i and JDK 1.5. Our experimental results are shown in Table 1, after performing data cleaning the number of requests are reduced from 861275 to 231265.

Entries in raw web log	Entries after data cleaning	Number of users	Number of sessions
861275	231265	53012	55225

Table 1: Results of Data Preprocessing in Web Usage Analysis

IX. CONCLUSION

The web pages are one of the most important advertisement tool in international area for foundation, institutions etc. Therefore, the suitability to W3C standards [14], content and design of web pages are very important for system administrator and Web designer. This paper has provided the details of data preprocessing steps that are essential for performing Web Usage Analysis. The WWW is a critical resource to carry out business and commerce. Therefore, the design of web pages is highly essential for the system administrators and web site creators. These characteristics have enormous impact on the number of users who access the page. So the web analyzer has to examine with the data of server log file for identifying the navigation pattern. There are number of methods proposed by different researchers for the web usage mining This paper has presented various formats of web server log files and the process of web usage analysis.

X. ACKNOWLEDGEMENT

We thank, Lakireddy Balireddy College of Engineering for providing the web server user access log files to us.

REFERENCES

1. Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, “Grouping Web page references into transactions for mining World Wide Web browsing patterns”, 1997.
2. Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, “Data preparation for mining World Wide Web browsing patterns”, 1999.
3. Jaideep Srivastva, Robert Cooly, Mukand Deepande, Pang-MingTan , “Web Usage Mining : Discovery and Applications of usage Patterns from Web Data “ , 2000
4. Drott, M. C, “Using Web Server Logs to Improve Site Design”. Association for Computing Machinery (ACM) Proceeding of the Sixteenth Annual International Conference on Computer Documentation, 1998, pp. 43 – 50.
5. Navin Kumar Tyagi, A.K. Solanki & Sanjay Tyagi, “An Algorithmic Approach to Data Preprocessing in Web Usage Mining”, International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No. 2, pp. 279-283
6. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan, “Web usage mining: Discovery and applications of usage patterns from Web data”, SIGKDD Explorations, 2000, Vol.1.pp. 12-23
7. Navin Kumar Tyagi, , A. K. Solanki and Manoj Wadhwa, “ Analysis of Server Log by Web Usage Mining for Website Improvement”, International Journal of Computer Science Issues (IJCSI), Jul2010, Vol. 7 Issue 4, p17, 2010
8. Yan Li, Boqin FENG and Qinjiao MAO, “Research on Path Completion Technique in Web Usage Mining”, International Symposium on Computer Science and Computational Technology, IEEE, 2008.
9. Yan Li and Bo-qin FENG “The Construction of Transactions for Web Usage Mining”, International Conference on Computational Intelligence and Natural Computing, IEEE, 2009.
10. Karuna P. Joshi, Anupam Joshi and Yelena Yesha, “On using a Ware-house to Analyze Web Logs. Distributed and Parallel Databases, 13(2): pp. 161-180, 2003.
11. Eleni Stroulea Nan niu and Mohammad El-Ramly., “Understanding Web Usage for Dynamic Web Site Adaptation” A Case Study in Proceedings of Fourth International Workshop on Web Site Evolution (WES, 02). Pages: 53-64, IEEE, 2002.
12. A.Banarjee and J.Ghosh, "Clickstream Clustering using Weighted Longest Common Subsequences" , in Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 2001.
13. Jeffery Heer and Edti. Chi, "Mining the Structure of User Activity using Cluster Stability", in Proceeding of the Workshop on Web Analytics, Second SIAM Conference on Data Mining. ACM press, 2002.
14. Internet: Hypertext Transfer Protocol Overview, http://www.w3.org/Protocols/, http://www.w3.org/Protocols/rfc2616/rfc261 6-sec1.html, 1995.
15. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan,"WebUsage Mining: Discovery and applications of usage patterns from Web data", 2000.

AUTHOR PROFILE

Naga Lakshmi Theerthala. She is working as Assistant Professor in Information Technology in Usha Rama College of Engineering, Telaprolu, Krishna District, Andhra Pradesh. She did her M.Tech in Computer Science and Engineering in Koneru Lakshmaiah College of Engineering, Vaddeswaram and pursuing Ph.D in Data Warehousing and Data Mining in JNTU. She has 9 years of teaching experience. She is member of ISTE and life member of CSI.

Dr. Raja Sekhara Rao. Kurra. He is working as Professor and Dean in Koneru Lakshmaiah University, Vaddeswaram, Guntur District, Andhra Pradesh. He did his MS in BITS Pilani and Ph.D in Acharya Nagarjuna University. He has 27 years of teaching experience. He published more than 20 technical papers in International Journals and more than 25 technical papers in National Journals and International Conferences. He is the Chairman of Koneru Chapter, CSI. He is the member in Board of Studies of CSE & IT, Acharya Nagarjuna University. He received “Best Teacher Award” five times in the years 1998-1999, 2003-2004, 2004-2005, 2005-2006 and 2006-2007. He is the life member of ISTE and CSI and fellow of the IETE.

Dr. Sai Satyanaryana Reddy. Seelam. He is working as Professor and Head of the Computer Science and Engineering Department, Lakireddy Balireddy College of Engineering, Mylavaram, Krishna District, Andhra Pradesh. He did his M.E in BITS Pilani and Ph.D in Bharat University, Chennai. He has 20 years of teaching experience. He received “VIDYA RATAN” Award from the Economic for Health and Educational Growth, New Delhi for “Excellence in Chosen filed of activity”. He is active member of IEEE and life member of ISTE and CSI. He published more than 10 technical papers in International Journals and more than 15 technical papers in International conferences.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)