RESEARCH ISSUES IN WEB MINING

Dr.S. Vijiyarani[1] and Ms. E. Suganya[2]

[1]Assistant professor, Department of Computer science, School of Computer Science and Engineering,
Bharathiar University, Coimbatore
[2] M.Phil Research Scholar, Department of Computer science, School of Computer science and
Engineering, Bharathiar University, Coimbatore

ABSTRACT

Web is a collection of inter-related files on one or more web servers while web mining means extracting valuable information from web databases. Web mining is one of the data mining domains where data mining techniques are used for extracting information from the web servers. The web data includes web pages, web links, objects on the web and web logs. Web mining is used to understand the customer behaviour, evaluate a particular website based on the information which is stored in web log files. Web mining is evaluated by using data mining techniques, namely classification, clustering, and association rules. It has some beneficial areas or applications such as Electronic commerce, E-learning, E-government, E-policies, E-democracy, Electronic business, security, crime investigation and digital library. Retrieving the required web page from the web efficiently and effectively becomes a challenging task because web is made up of unstructured data, which delivers the large amount of information and increase the complexity of dealing information from different web service providers. The collection of information becomes very hard to find, extract, filter or evaluate the relevant information for the users. In this paper, we have studied the basic concepts of web mining, classification, processes and issues. In addition to this, this paper also analyzed the web mining research challenges.

KEYWORDS

Web Mining, Classification, Application, Tools, Algorithms, Research Issues

1. INTRODUCTION

Web mining is the application of data mining technique which is an unstructured or semi-structured data and it automatically discovers and extracts potentially useful and previously unknown information or knowledge from the web [23]. The significant web mining applications are website design, web search, search engines, information retrieval, network management, E-commerce, business and artificial intelligence, web market places and web communities. Online business breaks the barrier of time and space as compared to the physical office business. Big companies around the world are realizing that e-commerce is not just buying and selling over Internet, rather it improves the efficiency to compete with other giants in the market. This application includes the temporal issues for the users.

Web mining has three classifications namely, web content mining, web structure mining and web usage mining. Each classification is having its own algorithms and tools. Web content mining is nothing but the discovery of valuable information from web documents and these web documents may contain text, image, hyperlinks, metadata and structured records. It is used to look at the information by search engine or web spiders i.e. Google, Yahoo. It is the process of retrieving the useful information from the web content or web documents. Web structure mining is also a process of discovering structured information from the websites. The structure of a graph consists of web pages and hyperlinks where the web pages are considered as nodes and the hyperlinks are edges and these are connecting between related pages. Web usage mining is also called as web log mining. It reflects the user's behaviour which can catch the meaningful patterns from one or more web localities [9].

Web mining process consists of four important steps, they are, resource finding, data selection and pre-processing, generalization and analysis [23]. Resource finding is the process which is used to extract the data either from online or offline text resources. In data selection and pre-processing step, specific information from retrieved web sources are automatically selected and pre-processed. During generalization, data mining and machine learning techniques are used to discover general patterns from individual web sites as well as across multiple sites. Validation and interpretation of the mined patterns are done in analysis step. [1][17]. Web mining is classified into three different categories, they are, web content mining, web structure mining and web usage mining. This is illustrated in Figure 1.

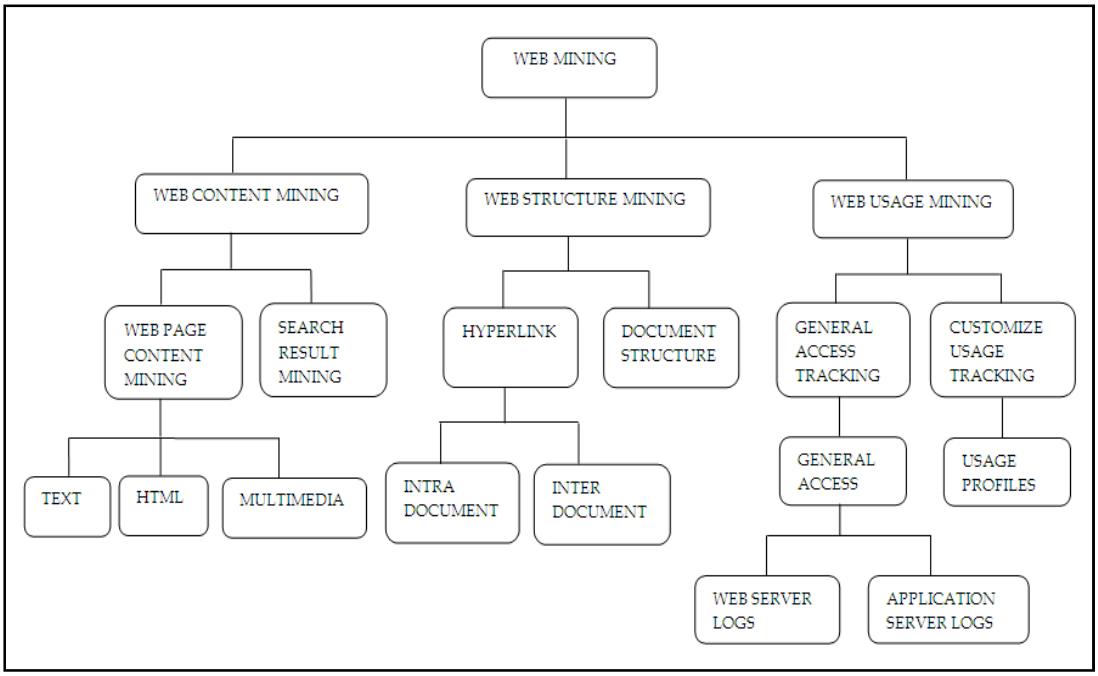

Figure 1. Classification of Web Mining

The remaining section of the paper is organized as follows. Section 2 discusses the research issues in web mining. Web content mining and its research challenges are given in Section 3. Section 4 describes web structure mining. Section 5 provides the details about web usage mining. Conclusions are given in Section 6.

2. RESEARCH ISSUES IN WEB MINING

The web is highly dynamic; lots of pages are added, updated and removed everyday and it handles huge set of information hence there is an arrival of many number of problems or issues. Normally, web data is high dimensional, limited query interface, keyword oriented search and limited customization to individual users. Due to this, it is very difficult to find the relevant information from the web which may create new issues. Web mining techniques are classification, clustering and association rules which are used to understand the customer behaviour, evaluate a particular website by using traditional data mining parameters. Web mining process is divided into four steps; they are resource finding, data selection and pre-processing, generalization and analysis [11] [8]. Web measurement or web analytics are one of the significant challenges in web mining. The measurement factors are hits, page views, visits or user sessions and find the unique visitor regularly used to measure the user impact of various proposed changes. Large institutions and organizations archive usage data from the web sites [10]. The main problem is that, detecting and/or preventing fraud activities. The web usage mining algorithms are more efficient and accurate. But there is a challenge that has to be taken into consideration. Web cleaning is the most important process but data cleaning becomes difficult when it comes to heterogeneous data [20]. Maintaining accuracy in classifying the data needs to be concentrated. Although many classification techniques exist the quality of clustering is still a question to be answered.

2.1 Major issues in Web Mining

- Web data sets can be very large, it takes ten to hundreds of terabytes to store on the database
- It cannot mine on a single server so it needs large number of server
- Proper organization of hardware and software to mine multi-terabyte data sets
- Limited customization, limited coverage, and limited query interface to individual users
- Automated data cleaning
- Over fitting and Under fitting of data
- Over sampling of data
- Scaling up for high dimensional data
- Mining sequence and time series data
- Difficulty in finding relevant information
- Extracting new knowledge from the web

3. WEB CONTENT MINING

Web content mining data may be structured or unstructured/semi structured even though much of web is unstructured. It is the process of retrieving the information from the web into more structured forms and indexing the information to retrieve quickly or finding valuable information from web content or web documents. Web content mining includes the web documents which may consist of text, html, multimedia documents i.e., images, audio, video and sound etc. The search result mining contains the web search results. It may be a structure documents or unstructured documents.

Web content mining used many algorithms and tools such as Genetic algorithm, Cluster Hierarchy Construction Algorithm (CHCA), Correlation algorithm. Web Info Extractor (WIE), Mozenda, screen-scrapper, ontology based tools; web content extractor and automation anywhere are content mining tools. Cloud users require to extract the information from the cloud provided by web servers can make use of the web mining. For instance, Web communities can be maintained the information such as facebook. That is the users of same field of interest can be grouped and they can communicate through the network. Digital library performs automated citation indexing using web mining techniques. E-services include e-banking, search engines, on-line auctions, on-line knowledge management, social networking, e-learning, blog analysis, and personalization and recommendation systems. This can be analyzed for the customers and enable provision to the customers based on their recommendations [18]. It has two approaches; they are (i) Agent based and (ii) Database Approach. Figure 2 gives the web content mining approaches.

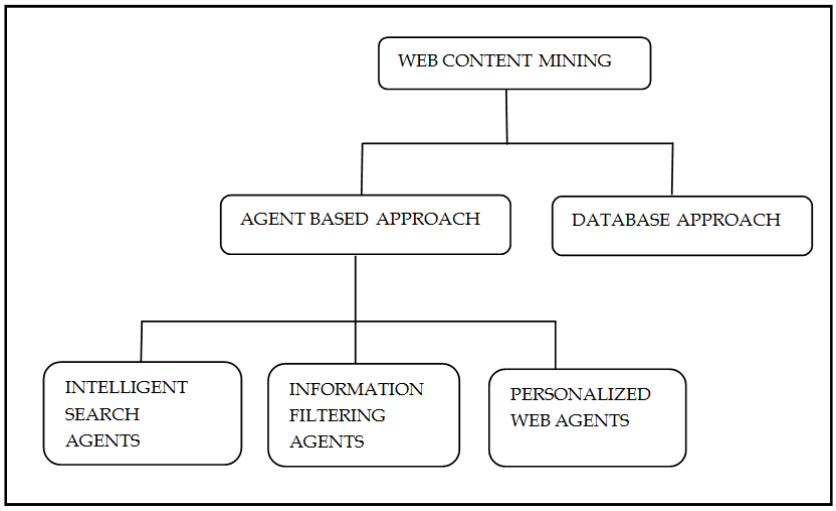

Figure2. Web Content Mining approaches

(i) Agent Based Approach

Agent based approach focuses on searching relevant information from the World Wide Web. Three types of agents they are

(i) Intelligent search agents – Automatically searches for information along with a particular query

(ii) Information filtering/categorizing agents - Filters the data

(ii) Personalized web agents – Discovers the documents those are related to the user profiles

(iii)Database Approach

Database approach consists of databases which contain attributes, tables and schema with defined domains. It focused on techniques for organizing the semi structured data on the web into more collections of resources, and using standard database querying mechanism and data mining techniques to analyze it, for example multilevel database and web querying system [5].

Web content mining has the other approaches to mine the data. These are unstructured text data mining, structure mining, and semi-structure text mining and multimedia data mining [16].

3.1 RESEARCH ISSUES ON WEB CONTENT MINING

Web content mining has number of research issues because it can extract the information from the web search engines.

- Data / Information Extraction concentrate on extraction of structured data from web pages such as products and search results.
- Web information integration and schema matching. The web contains large amount of data, each website accept similar information in a different way. Similar data discovery is an important problem with lots of realistic applications.
- Opinion extraction from online sources i.e. customer makes sure of products, forums, blogs and chat rooms. Mining opinions are of big consequence for marketing intelligence and product benchmarking.
- Automatically segmenting web pages and detecting noise is an interesting problem in web application. It could not have advertisements, navigation links and copyrights notices. Hence, extracting the main content of the web page is important problem in web application [19].

4. WEB STRUCTURE MINING

Web structure mining is the study of data interconnected to the structure of a particular website. It consists of web graph which contains the web pages or web documents as nodes and hyperlinks as edges those are connecting between two related pages [7]. Figure 3 represents the web graph structure.

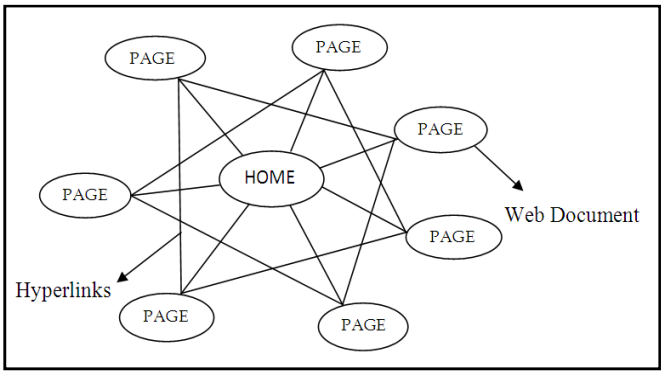

Figure3. Web Graph Structure

Web structure is useful source for extracting information. Web structure is to extract some interesting web graph patterns like co-citation, social choice, complete bipartite graphs, etc [1]. It classifies the web page on various topics and deciding which web page is to be added into the collection of web pages. Web structure mining can be performed either at intra-page level or inter-page level. A hyperlink that connects to a different part of the same page is called intra-page hyperlink. It is a document structure level [22].

A hyperlink that connects two different pages are called inter-page hyperlink which is structure level [12]. Web page is organized in tree structure format based on HTML tags. Here, the documents are extracted automatically by the Document Object Model (DOM). The main reason for developing link mining is to understand the social organization of the web. The research of structure analysis is called Link mining [14] which is located in the connection of work in link analysis, hypertext and web mining, relational learning, inductive logic programming and graph mining. Some of the important tasks of link mining are link based classification, link based cluster analysis, link type, link strength and link cardinality. The research of the hyperlink level is also called hyperlink analysis [22] which can be used to retrieve useful information from the web [13].

Web structure mining is used in search engines such as Google, Yahoo, etc. HITS algorithm was used in clever search engine by IBM and the page rank algorithm is used by Google [11]. Algorithms of web structure mining are HITS (Hypertext Induced Topic Search) algorithm, Max flow- Min cut algorithm, ECLAT algorithm, and Page rank algorithm. Page rank algorithm can be divided into two types. One is weighted page rank algorithm and another one is Topic sensitive page rank algorithm.

4.1 RESEARCH ISSUES ON WEB STRUCTURE MINING

Web structure mining has two issues due to its huge amount of data.

- Reducing irrelevant search results. Relevance of search information becomes unorganized due to the problem search engines often only tolerate for low precision criteria.
- Indexing information on the web [7]. This causes low amount of recall with content mining.

5. WEB USAGE MINING

Web usage mining is also called as web log mining which is used to analyze the behaviour of online users [2]. It fed into two types of tracking; one is general access tracking and another one is customize usage tracking [3]. The general access tracking is used to predict the customer behaviour on the web and it identifies the user while the user interacts with the web. It can store the data automatically when the web server log and application log [15]. The web log is located in three different locations they are web server log, web proxy server and client browser and it contains only plain text file (.txt). The large amounts of irrelevant data are available in the web log file because it contains noisy data, large amount of incomplete, eroded and unnecessary information [6]. Web server log files are used to identify the errors and failed requests were given by the web master and the system administrator. Web usage mining is to extract the data which are stored in server access logs, referrer logs, agent logs and error logs.

Web usage mining generally uses basic data mining algorithms such as association rule mining, sequential rule mining, clustering, and classification. It has several tools to analyze the behaviour of the user. They are KOINOTITES, web SIFT, web usage miner, INSITE, speed tracer, Archcollect, i-Miner, AWUSA, i-JADE web miner, Web Quilt, STRATDYN, SEWeP, webTool, MiDAS, web mate, WebLog Miner, DB2Intelligent miner of Data, Poly Analyst version 6.0, Clemetine, WEBMINER, WEBVIZ. Figure 4 shows the different web server logs [1].

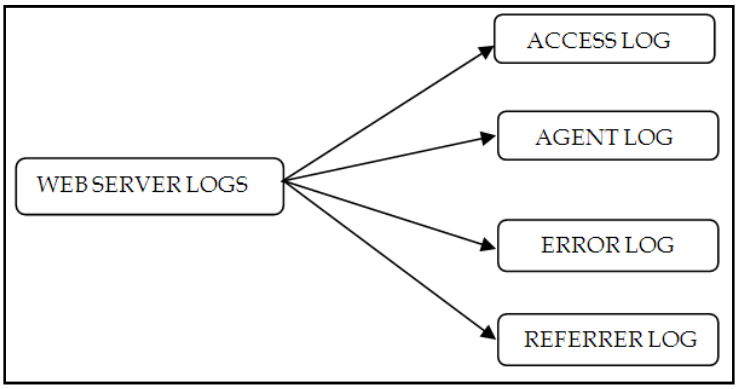

Figure4: Web Server Logs

5.1 WEB SERVER LOGS

5.1.1 Access log

Access log is used to capture the information about the user and it has many numbers of attributes. It will record each click event, hits and access of the user. It is one of the web server logs [16].

5.1.2 Agent log

Agent log is used to record the details about online user behaviour, user's browser, browser's version and operating system. It is a standard log file while comparing the access log.

5.1.3 Error log

When user click on a particular link and the browser does not display the particular page or website then the user receives error 404 not found.

5.1.4 Referrer log

Referrer log is used to store the 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 [21].

5.2 PROCESS OF WEB USAGE MINING

Web usage mining process is generally divided into three tasks:

5.2.1 Data pre-processing

Web log data pre-processing is nothing but, to identify users, sessions, page views and so on. In order to improve the efficiency and scalability many steps are required, these are, data fusion,

data cleaning, user identification by IP address, authentication data, cookies, client information and site topology, session identification, formatting, and path completion [6].

5.2.2 Pattern discovery

The data mining techniques and algorithms are used to perform in the pattern discovery by using clustering, association rules and sequential analysis. The association technique is mostly used in pattern discovery for detection of relation between visited pages by online users. It is used to extract patterns of usage from web data [4]. The extract pattern can be stand for in many ways such as graphs, charts, tables and forms.

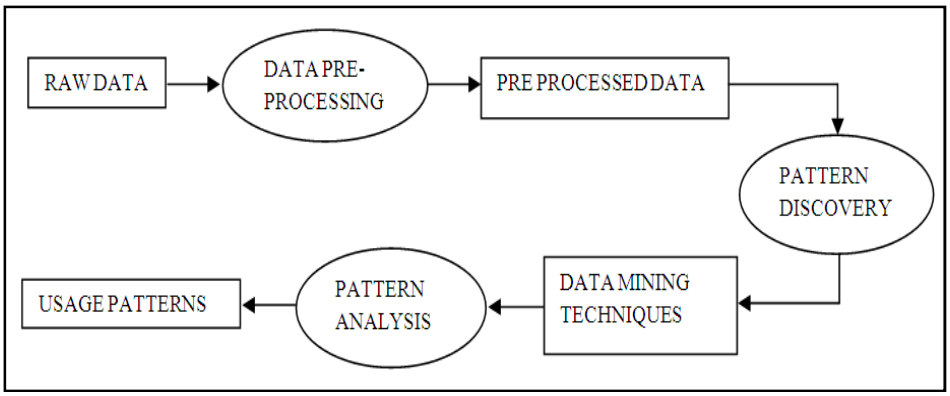

Figure5. Process of web usage mining

5.2.3 Pattern analysis

The last process of the web usage mining is pattern analysis. There are so many techniques are used for pattern analysis such as visualization technique, OLAP technique, data and knowledge querying and usability analysis [4].

5.3 RESEARCH ISSUES ON WEB USAGE MINING

Web usage mining has several issues because it involves number of data mining techniques. The problems are [11]

- Session identification
- CGI data
- Catching
- Dynamic pages
- Robot detection and filtering
- Transaction identification

6. CONCLUSION AND FUTURE WORK

This paper has discussed about the research issues and challenges in web mining and also provided detailed review about the basic concepts of web mining, web content mining, structure mining, usage mining, tools, algorithms and types. Several open research issues and drawbacks which are exists in the current techniques are also discussed. This study and review would be helpful for researchers those who are doing their research in the domain of web mining.

REFERNCES

[1] Joy Shalom Sona, Prof. Asha Ambhaikar" A Reconciling Website System to Enhance Efficiency with Web Mining Techniques" International Journal Of Scientific & Engineering Research Volume 3, Issue 2, February-2012 1 ISSN 2229-5518

[2] Aparna Ranade, Abhijit R. Joshi, Ph. D," Techniques for Understanding User Usage Behavior on the Internet" International Journal of Computer Applications (0975 – 8887) Volume 92 – No.7, April 2014

[3] Karan Bhalla & Deepak Prasad," Data Preparation and Pattern Discovery For Web Usage Mining"

[4] Amit Pratap Singh1, Dr. R. C. Jain 2," A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)Volume 3, Issue 3, May – June 2014 ISSN 2278-6856

[5] R. Lokeshkumar1, R. Sindhuja2, Dr. P. Sengottuvelan, "A Survey on Pre-processing of Web Log File in Web Usage Mining to Improve the Quality of Data" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014

[6] Mitali Srivastava, Rakhi Garg, P. K. Mishra," Preprocessing Techniques in Web Usage Mining: A Survey" International Journal of Computer Applications (0975 – 8887) Volume 97– No.18, July 2014

[7] http://www.slideshare.net/akhanna3/discovering-knowledge-using-web-structure-mining-27488978

[8] Ashish Kumar Garg, Mohammad Amir, Jarrar Ahmed, Man Singh, Sham Bansa," Implementation of a Search Engine" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

[9] C.Gomathi, M. Moorthi," Web Access Pattern Algorithms in Education Domain" Computer and information science journal vol. 1, No.4, November 2008

[10] Md. Zahid Hasan, Khawja Jakaria Ahmad Chisty and Nur-E-Zaman Ayshik, "Research Challenges in Web Data Mining", International Journal of Computer Science and Telecommunications Volume 3, Issue 7, July 2012

[11] Jaideep Srivastava, "Web Mining: Accomplishments & Future Directions", University of Minnesota USA, srivasta@cs.umn.edu, http://www.cs.umn.edu/faculty/srivasta.html

[12] http://www.slideshare.net/AmirFahmideh/web-mining-structure-mining

[13] http://www.slideshare.net/Tommy96/web-mining-tutorial

[14] http://www.faadooengineers.com/threads/2177-PPT-Link-Mining

[15] Deepti Kapila, Prof. Charanjit Singh, "Survey on Page Ranking Algorithms for Digital Libraries", International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 6, June 2014 ISSN: 2277 128X

[16] Alberto Sillitti, Marco Scotto, Giancarlo Succi, Tullio Vernazza," News Miner: a Tool for Information Retrieval"

[17] Sandhya, Mala chaturvedi, "a survey on web mining algorithms", The International Journal Of Engineering And Science (IIJES) Volume 2 Issue 3

[18] Ananthi.J," A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014

[19] S.Balan, "A Study of Various Techniques of Web Content Mining Research Issues and Tools", International journal of innovative research and studies ISSN 2319-9725

[20] D.Jayalatchumy, Dr. P.Thambidurai, "Web Mining Research Issues and Future Directions – A Survey", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 3

[21] Naga Lakshmi, Raja Sekhara Rao , Sai Satyanarayana Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4

[22] Mamta M. Hegde, Prof. M.V.Phatak, "Developing an approach for hyperlink analysis with noise reduction using Web Structure Mining", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012

[23] Mr. Dushyant B.Rathod, Dr.Samrat Khanna, "A Review on Emerging Trends of Web Mining and its Applications" ISSN: 2321-9939

BIOGRAPHY

Dr. S. Vijayarani has completed MCA, M.Phil and Ph.D in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and nationalconferences.

Ms. E. Suganya has completed M.Sc in Computer Science. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining and Web Mining.