Web Miner: A Tool for Discovery of Usage Patterns From Web Data

Roop Ranjan
Student, Department of Computer Science (FMIT), Jamia Hamdard, Hamdard Nagar,
New Delhi, 110062, India
roop.ranjan@gmail.com
http://jamiahamdard.edu

Sameena Naaz
Department of Computer Science (FMIT), Jamia Hamdard, Hamdard Nagar,
New Delhi, 110062, India
snaaz@jamiahamdard.ac.in
http://jamiahamdard.edu

Neeraj Kaushik
Department of Computer Applications, Galgotias College of Engg. & Tech., 1, Knowledge Park-II,
Greater Noida, 201306, India
kaushikneeraj@hotmail.com
http://galgotiacollege.edu

***Abstract* -As there is a huge amount of data available online, the World Wide Web is a fertile area for data mining research. In recent years a various surveys have been performed on static data of web sites to perform web usage mining. This paper deals with the Web usage mining of a website which is hosted on IIS web server. Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to perform improvements in web based applications. Web usage mining consists of three phases, pre-processing, 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 research is being performed on a log file using Log Parser.**

Keywords: W3SVC Log Files; IIS; MVC; Log Parser; .Net Framework.

I. INTRODUCTION

The goal of Web Usage Mining is to find out extract the useful information from web data or web log files. The other goals are to enhance the usability of the web information and to apply the technology on the web applications, for instance, pre-fetching and catching, personalization etc. For decision management, the result of web usage mining can be used for target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise and market analysis. A lot of work has been performed on static log files to process web usage mining. But in this paper, a framework has been provided to process web mining on log files of a dynamic web site. The major problem with Web Mining in general and Web Usage Mining in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of pre-processing and parsing before the actual extraction of the required information.

II. RELATED WORK

There have been some works around content mining, and structure mining, based on the research of Data mining and Information Retrieval, Information Extraction, and Artificial Intelligence. But, in the web usage mining research area, several groups did distinguished work. From the business and applications point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-business, e-services, e-education and so on [1]. Some models like WebSIFT System [2] have also been proposed for detail study of the web mining processes. A model called WHOWEDA (*W*are*ho*use of *We*b *Da*ta) has been proposed by Sanjay Madria, Sourav S Bhowmick [16] in which a discussion has been performed on various issues in web mining area. Various experiments have been performed for implementing web data as a web personalization tool [3] in which they have categorized the process of web mining in five phases i.e. i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications. A model has been proposed to get the benefit of combining the Semantic Web and Web Mining [5]. Accurate Web usage information could help to attract new customers, retain current customers,

improve cross marketing/sales, effectiveness of promotional campaigns, track leaving customers and find the most effective logical structure for their Web space [17]. A very good model has been proposed using decision trees which analyses the hyper links of the pages and their hierarchies of arrangements to analyse the page and their structure [8]. Some of the researchers have analysed the pattern using different algorithms like Apriori, Hash tree and Fuzzy and then we used enhanced Apriori algorithm to give the solution for Crisp Boundary problem with higher optimized efficiency while comparing to other algorithm [9]. Few have given the detailed review of web mining as another form of data mining [10]. Another aspect of web mining has been also given using two different views i.e. process-centric view which defined web mining as a sequence of tasks, and data-centric- view which defined web mining in terms of the types of web data that was being used in the mining process [11]. To facilitate the mining process a tool has been proposed which uses the different phases of the CRISP-DM methodology as data preparation, data selection, modeling and evaluation [4]. Researches also have performed research to use open Web APIs to understand the various aspects of web mining [12]. Chen Ting , Niu Xiao, Yang Weiping has utilized the flexibility and power of XML to solve many problems for mining the competitive intelligence system of enterprise [6]. In a survey paper Naresh Barsagade has discussed about the importance and future directions of Web Mining [7]. Researchers also attempted to provide an outlook on the existing tools, their specialized focus with respect to an applicative objectives and the need for a more comprehensive new entrant in this sphere in the light of the current scenario. In the end, the paper will be concluded by listing some challenges and future trends in this research area [13]. Raymond Kosala, Hendrik Blockeel have given their researches to remove the confusion among Web Usage Mining and other categories of Web Mining [15].

III. LOG PARSER

It is a splendid tool if you want to parse very large log files and have knowledge of SQL statements. Log Parser parses multiple IIS logs that are really large in size. With the help of Log Parser we can do deep troubleshooting and data mining of the IIS Logs, Event Viewer, the File System, other file types and more. Useful examples of parsing the IIS logs include: finding long running pages, finding all error pages with a 500 status code, all requests from a particular IP (potential hacking attempt), and much more. Log Parser is a powerful, versatile tool that provides universal query access to text-based data such as log files, XML files and CSV files, as well as key data sources on the Windows operating system such as the Event Log, the Registry, the file system, and Active Directory. The results of your query can be custom-formatted in text based output, or they can be persisted to more specialty targets like SQL, SYSLOG, or a chart.

Log Parser can be used with any type of log files such as ADS, BIN, COM, CSV, TSV, ETW, EVT, FS, HTTPERR, IIS, IISODBC, IISW3C, NCSA, NETMON, REG, TEXTWORD, TEXTLINE, URLSCAN, and W3C.

This paper deals with the discovery of usage patterns on W3SVC log file format. The W3C Extended log file format is the default log file format for IIS. It is a customizable ASCII text-based format. You can use IIS Manager to select which fields to include in the log file, which allows you to keep log files as small as possible. Because HTTP.sys handles the W3C Extended log file format, this format records HTTP.sys kernel-mode cache hits.

Table 1. Description of W3SVC Log File Content

Field	Appears As	Description	Default Y/N
Date	date	The date on which the activity occurred.	Y
Time	time	The time, in coordinated universal time (UTC), at which the activity occurred.	Y
Client IP Address	c-ip	The IP address of the client that made the request.	Y
User Name	cs-username	The name of the authenticated user who accessed your server. Anonymous users are indicated by a hyphen.	Y
Service Name and Instance Number	s-sitename	The Internet service name and instance number that was running on the client.	N
Server Name	s-computername	The name of the server on which the log file entry was generated.	N
Server IP Address	s-ip	The IP address of the server on which the log file entry was generated.	Y
Server Port	s-port	The server port number that is configured for the service.	Y
Method	cs-method	The requested action, for example, a GET method.	Y
URI Stem	cs-uri-stem	The target of the action, for example, Default.htm.	Y
URI Query	cs-uri-query	The query, if any, that the client was trying to perform. A Universal Resource Identifier (URI) query is necessary only for dynamic pages.	Y
HTTP Status	sc-status	The HTTP status code.	Y

IV. INTRODUCTION TO .NET FRAMEWORK 4.5

The .NET Framework is a managed execution environment that provides a variety of services to its running applications. It consists of two major components: the common language runtime (CLR), which is the execution engine that handles running applications; and the .NET Framework Class Library, which provides a library of tested, reusable code that developers can call from their own applications. If you are using the Windows operating system, the .NET Framework may already be installed on your computer. The .NET Framework consists of the common language runtime and the .NET Framework class library. The common language runtime is the foundation of the .NET Framework. You can think of the runtime as an agent that manages code at execution time, providing core services such as memory management, thread management, and remoting, while also enforcing strict type safety and other forms of code accuracy that promote security and robustness. In fact, the concept of code management is a fundamental principle of the runtime. Code that targets the runtime is known as managed code, while code that does not target the runtime is known as unmanaged code. The class library is a comprehensive, object-oriented collection of reusable types that you can use to develop applications ranging from traditional command-line or graphical user interface (GUI) applications to applications based on the latest innovations provided by ASP.NET, such as Web Forms and XML Web services.

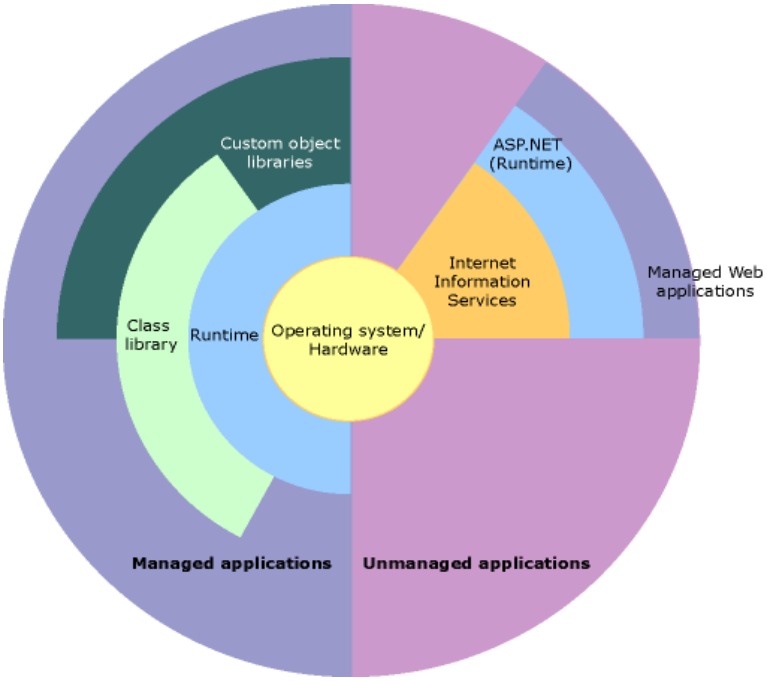


Fig. 1. Taxonomy of .Net Framework 4.5

V. PROCESS OF WEB USAGE MINING IMPLEMENTED

We have created a web based project in ASP .Net 4.5, to read the content of W3SVC log files. So, the process of performing web usage mining by our project is as follows:

A. Importing a log file in our project
B. Using Log Parser 2.2 for parsing the content of log files
C. Code for reading the log file data and display it on the web page in desired format
D. Analysis of log file data
E. Report Generation

A Importing a log file in our project

Following is the code to import the log file in our project.

```
protected void btnupload_Click(object sender, EventArgs e)
  {
    if (FileUpload1.HasFile)
    {
      Savepath  = Server.MapPath("~/Uploaded/" + FileUpload1.FileName);
       FileUpload1.SaveAs(Savepath);
       Session["upoaded"] = Savepath;
    }
  }
```

Here we are saving the log files in our project. The name of the folder in which we are saving the log file is “uploaded”/

B Using Log Parser 2.2 for parsing the content of log files

The code for reading the content of log files is as follows:

```
protected void btnAnalyse_Click(object sender, EventArgs e)
  {
```

```
MSUtil.ILogRecordset rslp;
MSUtil.ILogRecord rowlp;
string queryText;
queryText = String.Format("SELECT TOP 5 cs-uri-stem,cs-method, COUNT(*) as [Count] FROM
c:\\inetpub\\logs\\logfiles\\W3SVC2\\u_ex110108.log GROUP BY cs-uri-stem,cs-method ORDER BY [Count]
DESC", arg0);
rslp = logpr.Execute(queryText, ipcnxt);
Response.Write("<table id=tbl2 runat=server border=1>");
Response.Write("<tr><th>URL</th><th>Method</th><th>Count</th></tr>");
while (!rslp.atEnd())
{
   rowlp = rslp.getRecord();
Response.Write("<tr><td>" + rowlp.getValue(0) + "</td><td>" + rowlp.getValue(1) + "</td>" + "<td>" +
rowlp.getValue(2) + "</td></tr>");

   rslp.moveNext();
}
Response.Write("</table>");

string savp = Session["upoaded"].ToString();
string qurytxt2;
qurytxt2 = String.Format("SELECT distinct cs-uri-stem, time-taken FROM "+ savp +" ORDER BY time-
taken DESC", arg0);

rslp = logpr.Execute(qurytxt2, ipcnxt);

DataTable dt = new DataTable();

TableHeaderRow thr = new TableHeaderRow();
TableHeaderCell thc = new TableHeaderCell();
TableHeaderCell thc2 = new TableHeaderCell();
thc.Text = "URL";
thr.Cells.Add(thc);
tbl3.Rows.Add(thr);

dt.Columns.Add("URL");
dt.Columns.Add("TimeTaken");

Session["dr"] = "URL";
Session["Count"] = "TimeTaken";

thc2.Text = "Time Taken";
thr.Cells.Add(thc2);
tbl3.Rows.Add(thr);

while (!rslp.atEnd())
{
   rowlp = rslp.getRecord();
   TableRow tr1 = new TableRow();

   TableCell cell = new TableCell();
   TableCell cell1 = new TableCell();
   cell.Text = rowlp.getValue(0).ToString();
   tr1.Cells.Add(cell);

   cell1.Text = rowlp.getValue(1).ToString();
   tr1.Cells.Add(cell1);

   tbl3.Rows.Add(tr1);
```

```
        DataRow dr = dt.NewRow();
        dr["URL"] = rowlp.getValue(0).ToString();
        dr["TimeTaken"] = rowlp.getValue(1).ToString();
        dt.Rows.Add(dr);
        rslp.moveNext();

      }
      int a = tbl3.Rows.Count;
      Session["tbl"] = tbl3;
      Session["dt1"] = dt;
  }
```

Here we would like to mention that the steps C, D and E are implemented in the above code only.

VI. OUTPUT AND RESULTS

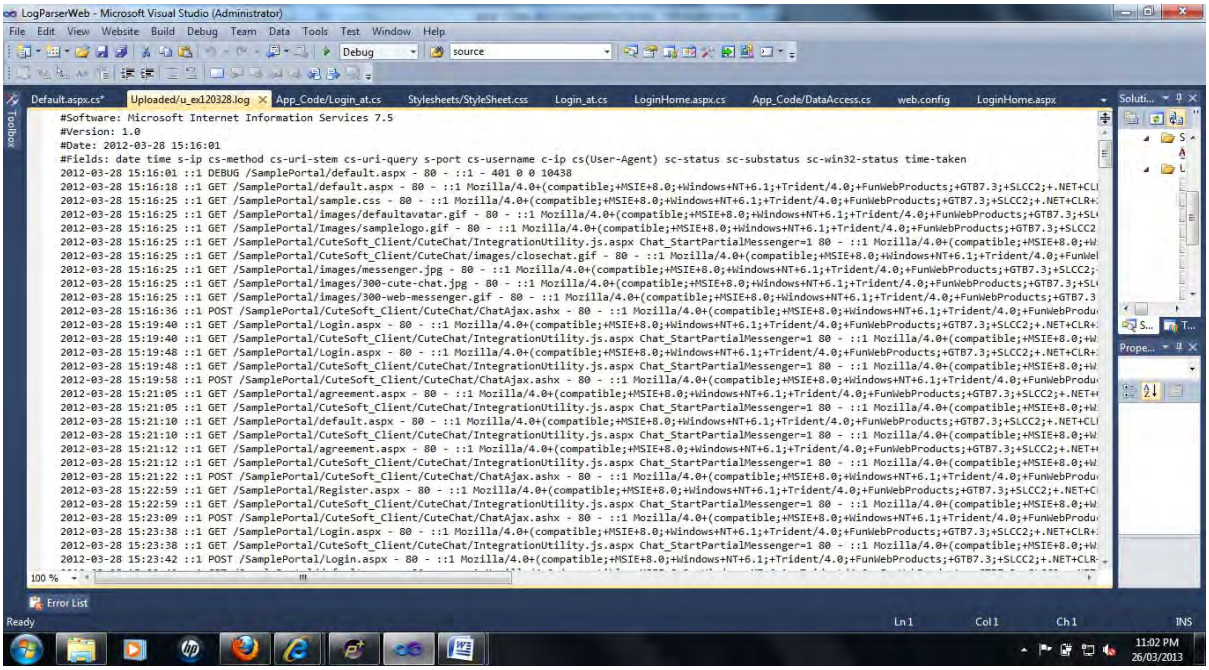
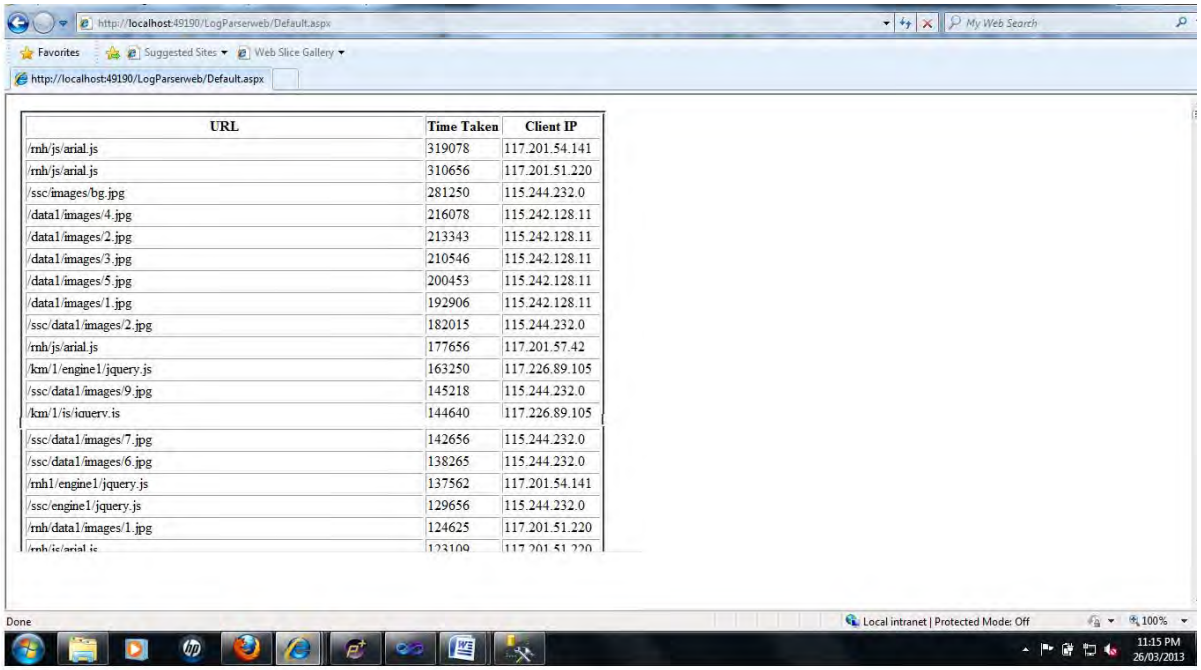


Fig. 2. The input log file content

The above figure shows the log file which is being input to the project for analysis. This log file will be read by the Log Parser in our project.



URL	Time Taken	Client IP
/rnh/js/arial.js	319078	117.201.54.141
/rnh/js/arial.js	310656	117.201.51.220
/ssc/images/bg.jpg	281250	115.244.232.0
/data1/images/4.jpg	216078	115.242.128.11
/data1/images/2.jpg	213343	115.242.128.11
/data1/images/3.jpg	210546	115.242.128.11
/data1/images/5.jpg	200453	115.242.128.11
/data1/images/1.jpg	192906	115.242.128.11
/ssc/data1/images/2.jpg	182015	115.244.232.0
/rnh/js/arial.js	177656	117.201.57.42
/km/1/engine1/jquery.js	163250	117.226.89.105
/ssc/data1/images/9.jpg	145218	115.244.232.0
/km/1/is/iquery.is	144640	117.226.89.105
/ssc/data1/images/7.jpg	142656	115.244.232.0
/ssc/data1/images/6.jpg	138265	115.244.232.0
/rnh1/engine1/jquery.js	137562	117.201.54.141
/ssc/engine1/jquery.js	129656	115.244.232.0
/rnh/data1/images/1.jpg	124625	117.201.51.220
/rnh/is/arial.is	123109	117.201.51.220

Fig. 3. The content of the log file which has been read by the code written

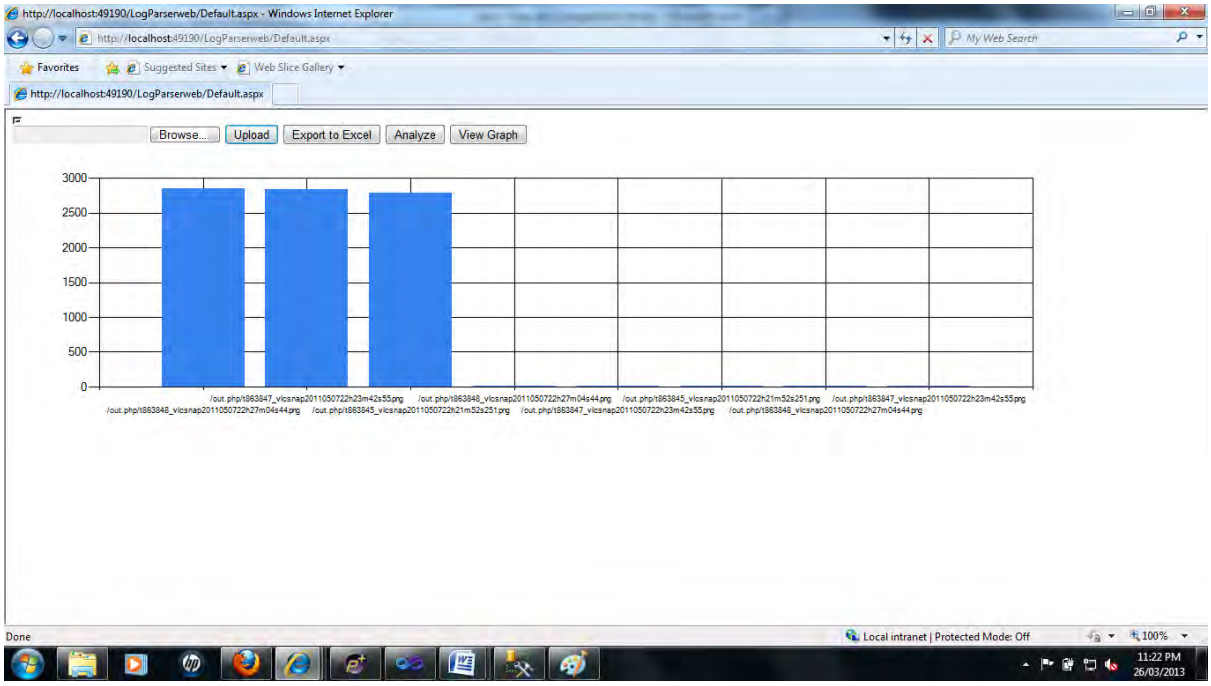


Fig. 4. Graphical analysis of the log file

In the above figure we have shown the graphical analysis of the contents of the log file. In the above figure X-axis represents the page or image i.e. the all the content of the web page stored in the log file which has been accessed by the users and Y-axis represents the time taken by the content to be downloaded at the clients end from the server in UTC time format.

VII. CONCLUSION

From the above experiments and developing the project for parsing the log files, we came to conclusion that using Log Parser tool not only we can parse W3SVC format log files but also a number of different log file like NCSA , another common log file format generated by the Tomcat Server. In our research, we have just shown few of the queries to just demonstrate how we can use .Net Framework, C# and SQL Server to parse the log files and discover a lot of useful information from the web data. And thus we have implemented Web Usage Mining.

REFERENCES

[1] L. Chen, and K. Sycara, WebMate: A Personal Agent for Browsing and Searching, *Proceedings of the 2nd International Conference on Autonomous Agents*, Minneapolis MN, USA, 1999, 132-139.

[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, PangNing Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" in SIGKD Explorations.Copyright19ACM SIGKD,Jan2000.

[3] A.Jebaraj Ratnakumar,"An Implementation of Web Personalization Using Web Mining Techniques", Journal Of Theoretical And Applied Information Technology", 2005 - 2010 JATIT

[4] Jose M. Domenech and Javier Lorenzo, "A Tool for Web Usage Mining", 8th International Conference on Intelligent

[5] Data Engineering and Automated Learning (IDEAL'07), 16-19 December, 2007, Birmingham, UK.

[6] Gerd Stumme, Andreas Hotho, Bettina Berendt, "Semantic Web Mining State Of The Art And Future Directions", Stumme, Hotho, And Berendt: Semantic Web Mining — A Survey.

[7] Chen Ting , Niu Xiao, Yang Weiping, The Application Of Web Data Mining Technique In Competitive Intelligence System Of Enterprise Based On Xml, Research Paper From IEEE.

[8] Naresh Barsagade, "Web Usage Mining and Pattern Discovery: A Survey Paper ", December 8, 2003

[9] Thales Sehn Korting, "C4.5 algorithm and Multivariate Decision Trees", Image Processing Division, National Institute for Space Research – INPES" ao Jos´e dos Campos – SP, Brazil
S.Veeramalai , N.Jaisankar and A.Kannan, "Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[10] Mr. Dushyant Rathod, "A Review On Web Mining ",International Journal of Engineering Research and Technology (IJERT) Vol. 1 Issue 2, April – 2012 , SSN: 2278-0181

[11] *Jaideep Srivastava, Prasanna Desikan, Vipin Kumar,* "Web Mining— Concepts, Applications, and Research Directions*", Page 400-417*

[12] Hsinchun chen, Xin Li, Michael Chau, Yi-jen Ho, Chunju Tseng, "Using Open Web APIs in teaching web mining" , The University of Arizona, The University of Hong Kong

[13] Chhavi Rana, "A Study of Web Usage Mining Research Tools", Int. J. Advanced Networking and Applications Volume:03 Issue:06 Pages:1422-1429 (2012) ISSN : 0975-0290

[14] eyalatha SIVARAMAKRISHNAN, Vijayakumar BALAKRISHNAN, "Web Mining Functions in an Academic Search Application", *Informatica Economică* vol. 13, no. 3/2009

[15] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, July 2000

[16] Sanjay Madria, Sourav s Bhowmick, w. -k ng, e. P. Lim, "Research Issues in Web Data Mining"

[17] J. I. Hong, , J. Heer, S. Waterson, and J. A. Landay,WebQuilt: A proxy-based approach to remote web usability testing, *ACM Transactions*