ISSN: 2321-7782 (Online)
Volume 2, Issue 11, November 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

An Introduction of WUBA (Web User Behavior Analysis)

N. Pushpalatha[1]
Assoc.Professor in CSE
Marri Laxman Reddy Institute of Technology & Management
Hyderabad - India

G. Prabhakara Reddy[2]
Assoc.Professor in CSE
Marri Laxman Reddy Institute of Technology & Management
Hyderabad - India

Abstract: ***Web mining refers to the use of data mining techniques to automatically retrieve, extract and evaluate (generalize/analyze) information for knowledge discovery from Web documents and services. Web data is typically unlabelled, distributed, heterogeneous, semi-structured, time varying, and high dimensional. Categorizing the end user in the web environment is a mind numbing task. Huge amount of operational data is generated when end user interacts in web environment. This generated operational data is stored in various logs and may be useful source of capturing the end user activates. Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analyzing these log files gives a neat idea about the user. This paper gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in Role mining algorithms to address an important access control problem: configuring a role-based access control system. Given a direct assignment of users to permissions, role mining discovers a set of roles together with an assignment of users to roles.***

Keywords: WUBA, Web content Mining, RBAC, Web structure mining, Web usage mining

I. INTRODUCTION

The analysis of human behavior has been conducted within diverse disciplines, such as psychology, sociology, economics, linguistics, and marketing and computer science. Hence, a broad theoretical framework is available, with a high potential for application into other areas, in particular to the analysis of web user browsing behavior. The above mentioned disciplines use surveys and experimental sampling for testing and calibrating their theoretical models. With respect to web user browsing behavior, the major source of data is the web logs, which store every visitor's action on a web site. Such files could contain millions of registers, depending on the web site traffic, and represents a major data source about human behavior. The proposed tool "**Web User Analyzer**" surveys the new trends in analyzing web user behavior and revises some novel approaches, such as those based on the neuron physiological theory of decision making, for describing what web users are looking for in a web site.

The current tool web user behavior analyzer, make use of logs that are logged during web surfing by the internet user, by using web mining techniques those logs are processed in to specific format by deleting un-necessary data from web logs. Log files are files that list the actions that have been occurred. These log files reside in the web server. Computers that deliver the web pages are called as web servers. The Web server stores all of the files necessary to display the Web pages on the user's computer. The browser requests the data from the Web server, and using HTTP, the server delivers the data back to the browser that had requested the web page. In the same way the server can send the files to many client computers at the same time, allowing multiple clients to view the same page simultaneously.

Almost 90% of the data is useless, and often does not represent any relevant information that the user is looking for. Taking into account the huge amount of data storage and manipulation needed for (say) a simple query, the processing

essentially requires adequate tools suitable for extracting only the relevant, sometimes hidden, knowledge as the final result of the problem under consideration.

To mine the interesting data from this huge pool, data mining techniques can be applied. But the web data is unstructured or semi structured. So we can'not apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is the use of data mining techniques to automatically discover and extract information from

Web mining is categorized into 3 types.

1. Content Mining (Examines the content of web pages as well as results of web Searching)

2. Structure Mining (Exploiting Hyperlink Structure)

3. Usage Mining (analyzing user web navigation)

Web User Behavior Analyzer (WUBA) focuses on the development of techniques and tools to study users web navigation behavior. Understanding the visitor's navigation preferences is an essential step in the study of the quality of an electronic commerce site. In fact, understanding the most likely access patterns of the users allows the service provider to customize and adapt the site's interface for the individual user, and to improve the site's static structure within the underlying hypertext system.

When web users interact with a site, data recording their behavior is stored in web server logs. These log files may contain invaluable information characterizing the users experience in the site. WUBA process the logs of logger file and formulate into different categories

1. Data collection – Web log files, which keeps track of visits of all the visitors.

2. Data Integration – Integrate multiple log files into a single file.

3. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction.

4. Pattern extraction – Extracting interesting patterns

5. Pattern analysis and visualization – Analyze the extracted pattern.

6. Pattern applications – Apply the pattern in real world problems.

WUBA uses the Role-Based Access Control (RBAC), is an access control model used in many systems. In WUBA, rather than assigning permissions directly to users, one introduces a set of roles and defines two relations: a user-role relation that assigns users to roles and a role-permission relation that assigns roles to permissions. This decomposition facilitates the administration of authorization policies since roles are (or should be) natural abstractions of functional roles within an enterprise and the two relations are conceptually easier to work with than a direct assignment of users to permissions.

1.1 Features of WUBA:

1.1.1 Hierarchical View: View reports in hierarchical form when it is appropriate. This is very convenient because most website statistics have hierarchical nature.

View Complete Reports, Not Just A Few Top Records: All reports are active and not just tables in static html. Scroll report easily or sort it by any column with just one mouse click. Open or close hierarchy levels.

1.1.2 Advanced Graphical Charts: All reports are accompanied by graphical charts that represent report data for easier visual perception. View own chart for each level of hierarchy. Bar graphs change immediately as you navigate & hellip;

Narrow your Focus With Category Filter: Select date interval in calendar to view reports for that period immediately. Compare same reports for different time intervals and understand the results of important changes made to the site.

1.1.3 Website Navigation: See how visitors navigate through your web site. Popular Paths through Site report shows the list of typical ways users take while browsing the website. Came from Page and left to page reports.

There are two main reasons for using distributed systems and distributed computing. First, the very nature of the application may require the use of a communication network that connects several computers. For example, data is produced in one physical location and it is needed in another location.

Second, there are many cases in which the use of a single computer would be possible in principle, but the use of a distributed system is beneficial for practical reasons. For example, it may be more cost-efficient to obtain the desired level of performance by using a cluster of several low-end computers, in comparison with a single high-end computer. A distributed system can be more reliable than a non-distributed system, as there is no single point of failure. Moreover, a distributed system may be easier to expand and manage than a monolithic uni processor system.

1. Network applications:

1. World Wide Web and peer-to-peer networks.

2. Massively multiplayer online games and virtual reality communities.

3. Distributed databases and distributed database management systems.

4. Network files systems.

5. Distributed information processing systems such as banking systems and airline reservation systems.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then next steps is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

1. Existing System

In Existing system there are many problems in which one seeks to develop predictive tool to analyze a user behavior in web, statistical tools such as multiple regression or data mining provide mature methods for computing model parameters when the set of predictive covariates and the model structure are pre-specified. Furthermore, recent research is providing new tools for inferring the structural form of non-linear predictive models, given good input and output data. However, the task of choosing which potentially predictive variables to study is largely a qualitative task that requires substantial domain expertise. For example, a survey designer must have domain expertise to choose questions that will identify predictive covariates. An engineer must develop substantial familiarity with a design in order to determine which variables can be systematically adjusted in order to optimize performance.

2. Proposed System

The rapid growth in user-generated content on the Internet is an example of how bottom-up interactions can, under some circumstances, effectively solve problems,Web usage mining model is a kind of mining to server logs. Web User Behavior analysis plays an important role in analyzing the logs captured during user request to several web servers, the browser requests the data from the Web server, and using HTTP, the server delivers the data back to the browser that had requested the web page.

In the same way the server can send the files to many client computers at the same time, allowing multiple clients to view the same page simultaneously. Almost 90% of the data is useless, and often does not represent any relevant information that the user is looking for. Taking into account the huge amount of data storage and manipulation needed for (say) a simple query, the processing essentially requires adequate tools suitable for extracting only the relevant, sometimes hidden, knowledge as the final result of the problem under consideration.

To mine the interesting data from this huge pool, data mining techniques can be applied. But the web data is unstructured or semi structured. So we cannot apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is the use of data mining techniques to automatically discover and extract information from. When web users interact with a site, data recording their behavior is stored in web server logs. These log files may contain invaluable information characterizing the users experience in the site. In addition, since in a medium size site log files amount to several megabytes a day, there is a necessity of techniques and tools to help take advantage of their content.

III. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

1. Economical Feasibility
2. Technical Feasibility
3. Social Feasibility

1. Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

2. Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

3. Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. System design

A variety of implementations and realizations are employed by Web usage mining systems. This section gives a generalized structure of the systems, each of which carries out five major tasks:

Usage data gathering: Web logs, which record user activities on Web sites, provide the most comprehensive, detailed Web usage data.

Usage data preparation: Log data are normally too raw to be used by mining algorithms. This task restores the users' activities that are recorded in the Web server logs in a reliable and consistent way.

Navigation pattern discovery: This part of a usage mining system looks for interesting usage patterns contained in the log data. Most algorithms use the method of sequential pattern generation, while the remaining methods tend to be rather ad hoc.

Pattern applications: The navigation patterns discovered can be applied to the following major areas, among others: i) improving the page/site design, ii) making additional product or topic recommendations, iii) Web personalization, and iv) learning the user or customer behavior.

Pattern analysis and visualization: Navigation patterns show the facts of Web usage, but these require further interpretation and analysis before they can be applied to obtain useful results.

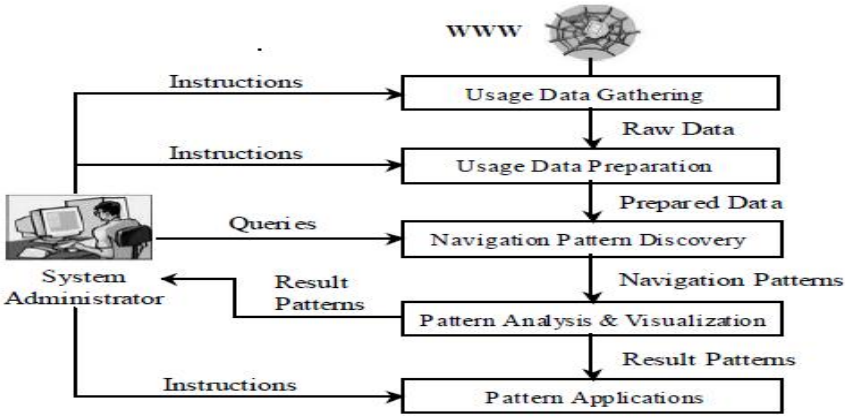


Fig 1.1 A Web usage mining system structure

Figure 1.1 shows a generalized structure of a Web usage mining system; the five components will be detailed in the next five sections. A usage mining system can also be divided into the following two types:

Personal: A user is observed as a physical person, for whom identifying information and personal data/properties are known. Here, a usage mining system optimizes the interaction for this specific individual user, for example, by making product recommendations specifically designed to appeal to this customer.

Impersonal: The user is observed as a unit of unknown identity, although some properties may be accessible from demographic data. In this case, a usage mining system works for a general population, for example, the most popular products are listed for all customers.

Basic definitions of Role Based:

The notation we use is borrowed from the NIST standard for Core Role-Based Access Control (Core RBAC) and it is adapted to our needs. We denote with USERS = {u1 , . . . , un } the set of users, with PMRS = {p1 , . . . , pm} the set of permissions, & with ROLES ={r1 , . . . , rt } the set of roles. The following assignment relations are defined.

- U RA ⊆ U SE RS × ROLE S is a many-to-many map- ping user-to-role assignment relation.

- RPA ⊆ ROLE S × PMRS is a many-to-many map- ping role-to-permission assignment relation.
- U PA ⊆ U SE RS × PMRS is a many-to-many map- ping user-to-permission assignment relation.

Data Gathering:

Web usage data are usually supplied by two sources: trial runs by humans and Web logs. The first approach is impractical an rarely used because of the nature of its high time and expense costs and its bias. Most usage mining systems use log data as their data source. This section looks at how and what usage data can be collected.

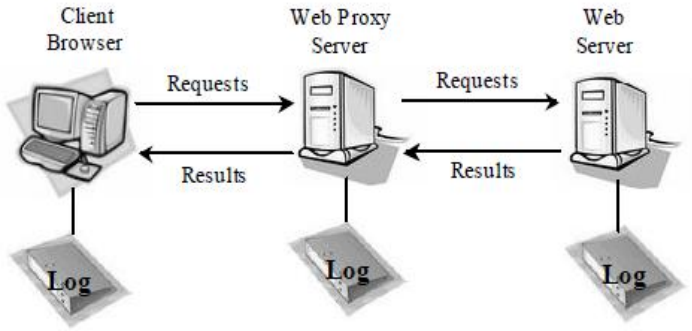


Fig 1.2 Three Web Log File Locations

Web Logs:

A Web log file records activity information when a Web user submits a request to a Web server. A log file can be located in three different places: i) Web servers, ii) Web proxy servers, and iii) client browsers, as shown in Figure 3, and each suffers from two major drawbacks:

Server-side logs: These logs generally supply the most complete and accurate usage data, but their two drawbacks are: These logs contain sensitive, personal information, therefore the server owners usually keep them closed. The logs do not record cached pages visited. The cached pages are summoned from local storage of browsers or proxy servers, not from Web servers.

Proxy-side logs: A proxy server takes the HTTP requests from users and passes them to a Web server; the proxy server then returns to users the results passed to them by the Web server. The two disadvantages are: Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction.

The request interception is limited, rather than covering most requests. The proxy logger implementation in WebQuilt, a Web logging system, can be used to solve these two problems, but the system performance declines if it is employed because each page request needs to be processed by the proxy simulator.

Client-side logs: Participants remotely test a Web site by downloading special software that records Web usage or by modifying the source code of an existing browser. HTTP cookies could also be used for this purpose.These are pieces of information generated by a Web server and stored in the users' computers, ready for future access. The drawbacks of this approach are:The design team must deploy the special software and have the end-users install it.

This technique makes it hard to achieve compatibility with a range of operating systems and Web browsers.

Web Log Information:

A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site.

#Version: 1.0 #Date: 12-Jan-1996

00:00:00 #Fields: time cs-method

cs-uri 00:34:23 GET /foo/bar.html

12:21:16 GET /foo/bar.html

12:45:52 GET /foo/bar.html

12:57:34 GET /foo/bar.html

- Authuser: Username and password if the server requires user authentication.
- Bytes**:** The content-length of the document transferred.
- Entering and exiting date and time.
- Remote IP address or domain name: An IP address is a 32-bit host address defined by the Internet Protocol; a domain name is used to determine a unique Internet address for any host on the Internet such as, cs.und.nodak.edu. One IP address is usually defined for one domain name, e.g., cs.und.nodak.edu points to 134.129.216.100.

4. Implementation

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

4.1 Modules

4.1.1 Registration

4.1.2 Login

4.1.3 Web Sniffing

4.1.4 Log formatting

4.1.5 WUBA Report

4.2 Modules Description

4.2.1 Registration

- In the registration module, only the admin has rights to register the portal.
- In the registration process, the admin has to enter user details such as name, password, confirm password, email id, mobile number and college enrolled number.
- During registration process, the user has to set password which contains alphanumeric and special character.

4.2.2Login

- In the LoginPage,the admin has to enter the user id and the password to enter the portal.
- User is not allowed to enter the Portal if the credentials get mismatched.

4.2.3 Web Sniffing

Browser sniffing is a set of techniques used in websites and web applications in order to determine the web browser a visitor is using, and to serve browser-appropriate content to the visitor.

4.2.3.1 Client-side sniffing:

Web pages can use programming languages such as JavaScript which are interpreted by the user agent, with results sent to the web server. This code is run by the client computer, and the results are used by other code to make necessary adjustments on client-side. In this example, the client computer is asked to determine whether the browser can use a feature called ActiveX.

4.2.3.2 Server-side sniffing:

Extensive browser techniques enable persistent user tracking even when users try to stay pseudonymous. See device fingerprint for more details on browser fingerprinting, a relatively new, extensive browser sniffing on steroids technique.

4.2.4 Log Formatting:

Web server, and using HTTP, the server delivers the data back to the browser that had requested the web page. In the same way the server can send the files to many client computers at the same time, allowing multiple clients to view the same page simultaneously.

Almost 90% of the data is useless, and often does not represent any relevant information that the user is looking for. Taking into account the huge amount of data storage and manipulation needed for (say) a simple query, the processing essentially requires adequate tools suitable for extracting only the relevant, sometimes hidden, knowledge as the final result of the problem under consideration.

4.2.5 WUBA Report:

Based on weblog filtering, on entering the student enrolled number the browsing history details will be listed over the screen in terms of different categories such as Education, Social Networking and others (Sports,News).

6. WUBA Reports:

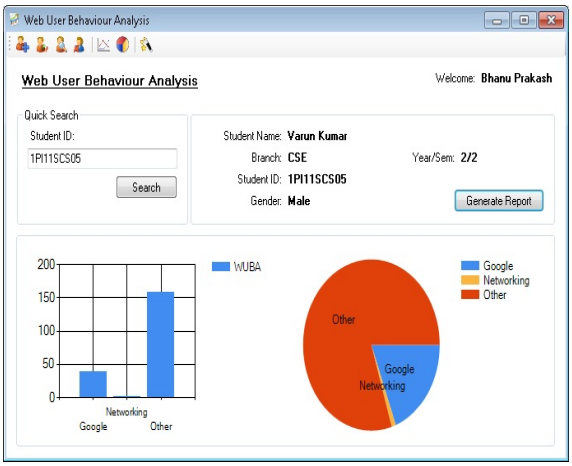
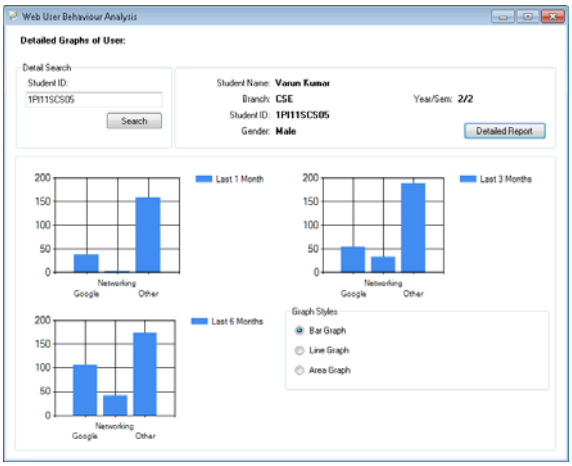
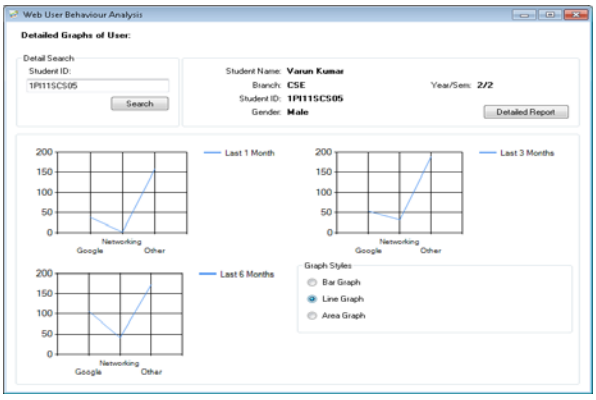


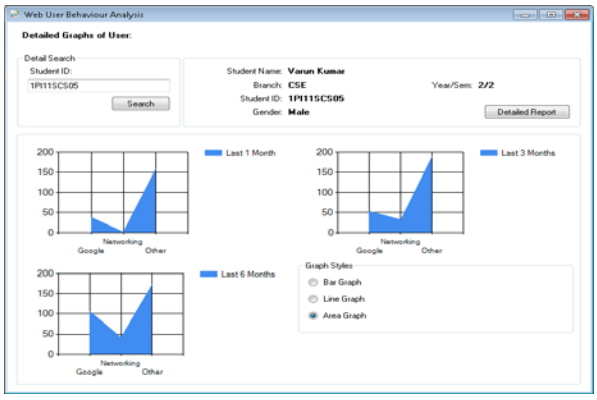
Fig 6.1 WUBA Student Report



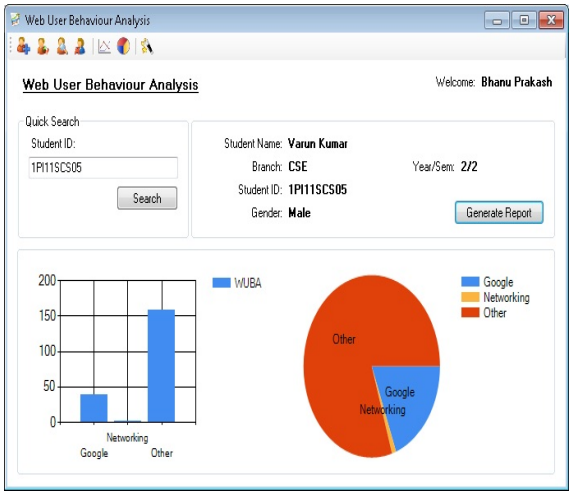
6.2 WUBA Bar Graph of Student



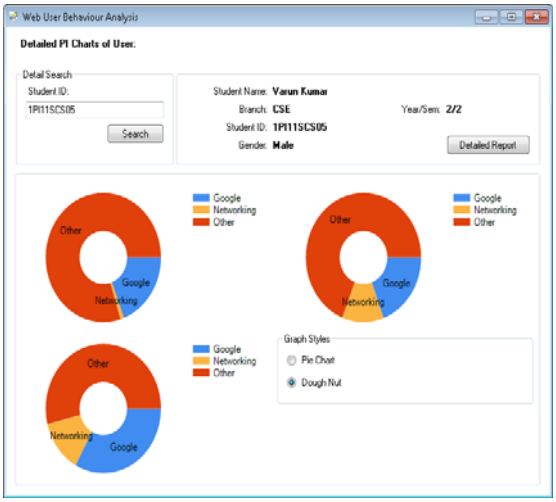
6.3 WUBA Line Graph of Student:



6.4 WUBA Area Graph of Student



6.5 WUBA Pie Graph of Student



6.6 WUBA Dough Nut Graph of Student

IV. CONCLUSION

Web usage mining model is a kind of mining to server logs. Web Usage Mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. In this paper we tried to give a clear understanding of the web server logs, their types and interesting patterns extracted from the web logs. Discovering such information that can be used to improve a business's performance or increase the effectiveness of a particular website. We have divided the hybrid role mining problem into two parts and provided solutions for them: determining the relevance of business information for role mining, and incorporating this information into a hybrid role mining algorithm. We solved the first problem with an entropy-based measure of relevance and the second by deriving an objective function that combines a probabilistic model of RBAC with business information.

References

1. User Interfaces in C#: Windows Forms and Custom Controls by Matthew MacDonald.
2. Applied Microsoft® .NET Framework Programming (Pro-Developer) by Jeffrey Richter.
3. Practical .Net2 and C#2: Harness the Platform, the Language, and the Framework by Patrick Smacchia.
4. Data Communications and Networking, by Behrouz A Forouzan.
5. Computer Networking: A Top-Down Approach, by James F. Kurose.
6. Operating System Concepts,by Abraham Silberschatz.
7. Nikhil Kumar Singh, Deepak Singh Tomar & Bhola Nath Roy "An Approach to Understand the End User Behavior through Log Analysis" International Journal of Computer Applications (0975 – 8887) Volume 5– No.11, August 2010
8. L.K. Joshila Grace, V.Maheswari & Dhinaharan Nagamalai "ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING" International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011
9. Mario Frank, Andreas P. Streich, David Basin, Joachim M. Buhmann "A Probabilistic Approach to Hybrid Role Mining" Department of Computer Science, ETH Zurich, Switzerland.
10. Aditi Shrivastava & Nitin Shukla "Extracting Knowledge from User Access Logs " International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012 1 ISSN 2250-3153
11. V.S.Thiyagarajan & Dr.K.Venkatachalapathy "Web Data mining-A Research area in Web usage mining" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 13, Issue 1 (Jul. - Aug. 2013), PP 22-26
12. Robert Rinnan "Benefits of Centralized Log file Correlation" Master's Thesis, Master of Science in Information Security ECTS, Department of Computer Science and Media Technology Gjøvik University College, 2005.
13. Muhammad Kamran Ahmed, Mukhtar Hussain and Asad Raza "An Automated User Transparent Approach to log Web URLs for Forensic Analysis" Fifth International Conference on IT Security Incident Management and IT Forensics 2009.

AUTHOR(S) PROFILE

Mrs. N. Pushpalatha working as a Assoc. professor in Marri Laxman Reddy Institute of Technology and Management, Hyderabad. She has 8 years teaching experience and good knowledge in computer subjects. She completed master degree in computer science and engineering dept. from University College of Engg, JNTU Campus, Kakinada. Presently Pre Ph.D completed from JNTU, Hyderabad.