International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 3, May 2014, 157- 166

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

HYBRID METHODOLOGY TO ANALYZE WEB USER BEHAVIOR IN WEB MINING AND FUZZY NETWORKS

N. PUSHPA LATHA[1], K. V. N BHANU PRAKASH[2] & K. VENKATESWARA REDDY[3]

[1]Assistant Professor, Marri Laxman Reddy Institute of Technology & Management, Dundigal, Hyderabad, India

[2]Department of CSE, Marri Laxman Reddy Institute of Technology & Management, Dundigal, Hyderabad, India

[3]Principal, Marri Laxman Reddy Institute of Technology & Management, Dundigal, 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: Web Data, IP Address, Using HTTP, URL

INTRODUCTION

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 users 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.

- Content Mining (Examines the content of web pages as well as results of web Searching)
- Structure Mining (Exploiting Hyperlink Structure)
- Usage Mining (analyzing user web navigation)

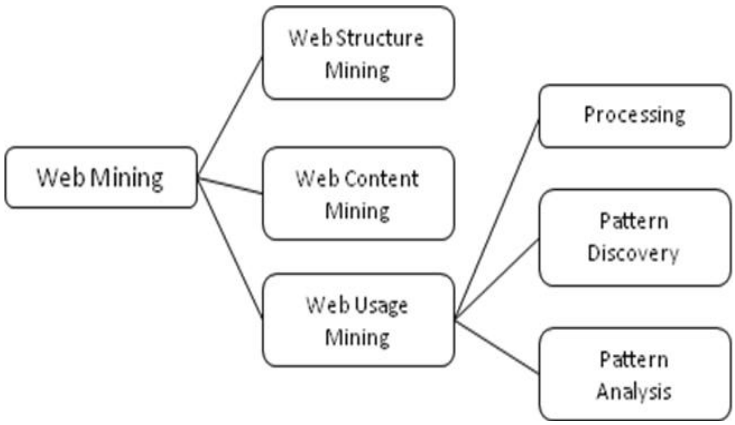

Figure 1

Web usage mining is a research field that focuses on the development of techniques and tools to study users web navigation behavior. Understanding the visitors 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. 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.

Five major steps followed in web usage mining are

- **Data Collection:** Web log files, which keeps track of visits of all the visitors
- **Data Integration:** Integrate multiple log files into a single file
- **Data Preprocessing:** Cleaning and structuring data to prepare for pattern extraction
- **Pattern Extraction:** Extracting interesting patterns
- **Pattern Analysis and Visualization:** Analyze the extracted pattern
- **Pattern Applications:** Apply the pattern in real world problems

Role-Based Access Control (RBAC) is an access control model used in many systems. In RBAC, 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.

A System Structure

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.

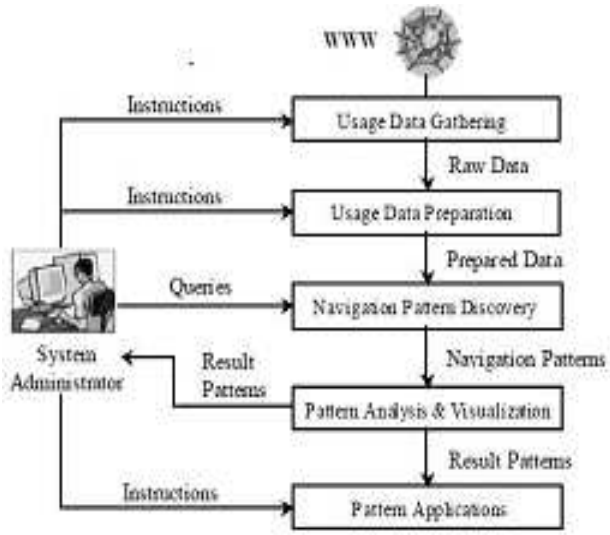

Figure 2: A Web Usage Mining System Structure

Figure 2 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 assignment relations are defined.

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

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.

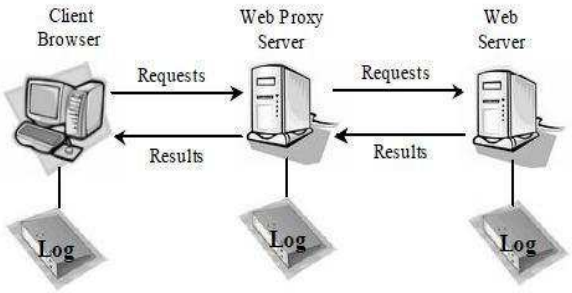

Figure 3: Three Web Log File Locations

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 Web Quilt, 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.

Modus of Request: GET, POST or HEAD method of CGI (Common Gateway Interface).

Number of hits on the page Remote log and agent log. Remote URL

"request:" The request line exactly as it came from the client.

Requested URL z

rfc931: The remote logname of the user.

Status: The HTTP status code returned to the client, e.g., 200 is

"ok" and 404 is "not found."

DATA PREPARATION

The information contained in a raw Web server log does not reliably represent a user session file. The Web usage data preparation phase is used to restore users' activities in the Web server log in a reliable and consistent way. This phase should at a minimum achieve the following four major tasks: i) removing undesirable entries, ii) distinguishing among users, iii) building sessions, and iv) restoring the contents of a session.

Removing Undesirable Entries

Web logs contain user activity information, of which some is not closely relevant to usage mining and can be removed without noticeably affecting the mining.

As much irrelevant information as possible should be removed before applying data mining algorithms to the log data

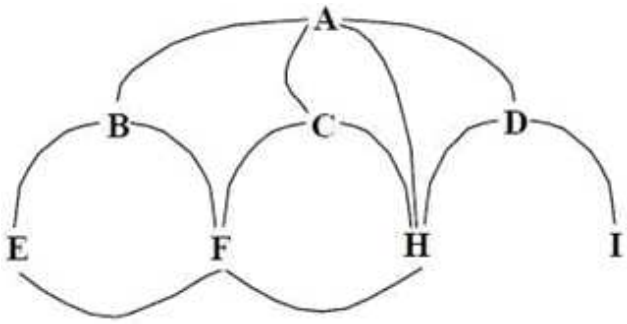

Figure 4: A Sample Website

Distinguishing among Users

A user is defined as a single individual that accesses files from one or more Web servers through a browser. A Web log sequentially records users' activities according to the time each occurred. In order to study the actual user behavior, users in the log must be distinguished. Figure 3 is a sample Web site where nodes are pages, edges are hyperlinks, and node A is the entry page of this site. The edges are bi-directional because users can easily use the back button on the browser to return to the previous page. Assume the access data from an IP address recorded on the log are those given in Table 1. Two user paths are identified from the access data: i) A-D-I-H-A-B-F and ii) C-H-B. These two paths are found by heuristics; other possibilities may also exist.

Table 1: Sample Access Data from an IP Address on the Web Site in Figure 4

No	Time	Requested URL	Remote URL
1	12:05	A	-
2	12:15	D	A
3	12:32	C	-
4	12:45	I	D
5	12:59	H	C
6	01:10	B	A
7	02:21	H	D
8	03:22	A	-
9	03:23	B	A
10	03:49	F	B

Building Sessions

For logs that span long periods of time, it is very likely that individual users will visit the Web site more than once or their browsing may be interrupted. The goal of session identification is to divide the page accesses of each user into individual sessions. A time threshold is usually used to identify sessions. For example, the previous two paths can be further assigned to three sessions: i) A-D-I-H, ii) A-B-F, and iii) C-H-B if a threshold value of thirty minutes is used.

Restoring the Contents of a Session

This task determines if there are important accesses that are not recorded in the access logs. For example, Web caching or using the back button of a browser will cause information discontinuance in logs. The three user sessions previously identified can be restored to obtain the complete sessions:

- A-D-I-D-H,
- A-B-F and
- C-H-A-B because there are no direct links between I and H and between H and B in Figure 4

The SMA Heuristic

In this section we present sma, our Simple role Mining Algorithm. Such an heuristic simply tries to generate a candidate role at time by selecting, according to a given criterion, one of the rows or one of the columns of the UPA matrix. The permissions assigned to the selected user are collected in the new candidate role which is added to the candidate role set.

```
CHOOSEROW(matrix, criterion)
 1  nr ← NUMROWS(matrix)
 2  if (criterion = MINIMUM )
 3     then
 4          m ← min{|matrix[i]| : 1 ≤ i ≤ nr}
 5          candidateRows ← {i : 1 ≤ i ≤ nr and |matrix[i]| = m}
 6  elseif (criterion = MAXIMUM )
 7     then
 8          m ← max{| matrix[i]| : 1 ≤ i ≤ nr}
 9          candidateRows ← {i : 1 ≤ i ≤ nr and |matrix[i]| = m}
10  elseif (criterion = RANDOM )
11     then
12          candidateRows ← {i : 1 ≤ i ≤ nr}
13  numRow ←R candidateRows
14  return numRou
```

The procedure Choose Row is used to determine which row to consider in order to form a candidate role. If such procedure returns index i, then the role that will be added to the list of candidate roles is the one comprising all the permissions pj such that UPA[i][j] = 1. In choosing a row's index (i.e., a user) we can select three different strategies. Indeed, we could choose a user (i.e., a row's index) at random among: i) all users; ii) the users having the minimum number of permissions; iii) the users having the maximum number of permissions.

In the following procedures we will use the following notation. Given an a x b binary matrix M, for 1 <= i <= a, with M[i] we denote the M's i-th row; while, with |M[i]| we denote the number of ones appearing in M[i].

PATTERN ANALYSIS & VISUALIZATION

Navigation patterns, which show the facts of Web usage, need further analysis and interpretation before application. The analysis is not discussed here because it usually requires human intervention or is distributed to the two other tasks: navigation pattern discovery and pattern applications. Navigation patterns are normally two-dimensional paths that are difficult to perceive if a proper visualization tool is not supported. A useful visualization tool may provide the following functions:

Displays the discovered navigation patterns clearly.

Provides essential functions for manipulating navigation patterns, e.g., zooming, rotation, scaling, etc.

Pattern Applications

The results of navigation pattern discovery can be applied to the following major areas, among others: i) improving site/page design, ii) making additional topic or product recommendations, iii) Web personalization and iv) learning user/customer behavior. Web caching, a less important application for navigation patterns, is also discussed.

- **Web Site/Page Improvements**

The most important application of discovered navigation patterns is to improve the Web sites/pages by (re)organizing them. Other than manually (re)organizing the Web sites/pages, there are some other automatic ways to achieve this. Adaptive Web sites automatically improve their organization and presentation by learning from visitor access patterns. They mine the data buried in Web server logs to produce easily navigable Web sites. Clustering mining and conceptual clustering mining techniques are applied to synthesize the index pages, which are central to site organization.

- **Topic or Product Recommendations**

Electronic commerce sites use recommender systems or collaborative filtering to suggest products to their customers or to provide consumers with information to help them decide which products to purchase. For example, each account owner at Amazon.com is presented with a section of Your Recommendations, which suggests additional products based on the owner's previous purchases and browsing behavior. Various technologies have been proposed for recommender systems and many electronic commerce sites have employed recommender systems in their sites [28]. For further studies, the Group Lens research group at the University of Minnesota is known for its successful projects on various recommender systems.

- **Web Personalization**

Web personalization (re)organizes Web sites/pages based on the Web experience to fit individual users' needs. It is a broad area that includes adaptive Web sites and recommender systems as special cases. The Web Personalizer system uses a subset of Web log and session clustering techniques to derive usage profiles, which are then used to generate recommendations. An overview of approaches for incorporating semantic knowledge into the Web personalization process is given in the article by Dai and Mobasher.

- **User Behavior Studies**

Knowing the users' purchasing or brows ing behavior is a critical factor for the success of E-commerce. The 1:1Pro system constructs personal profiles based on customers' transactional histories. The system uses data mining techniques to discover a set of rules describing customers' behavior and supports human experts in validating the rules.

CONCLUSIONS

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. 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
2. 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
3. Mario Frank, Andreas P. Streich, David Basin, Joachim M. Buhmann "A Probabilistic Approach to Hybrid Role Mining" Department of Computer Science, ETH Zurich, Switzerland
4. 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
5. 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: 2278-8727. Volume 13, Issue 1 (Jul. - Aug. 2013), PP 22-26
6. Robert Rinnan "Benefits of Centralized Log file Correlation" Master's Thesis, Master of Science in Information Security ECTS, Department of Computer Science and
7. Media Technology Gjøvik University College, 2005.
8. 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 DETAILS

N. Pushpa latha. working as a Assistant professor in Marri Laxman Reddy Institute of Technology and Management, Hyderabad. She has 7+ 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 pursuing Ph. D from JNTU, Hyderabad

Mr. K. V. N. Bhanu Prakash had his education in Computer Science and Engineering in MLR Institute of Technology and Masters degree in Computer Science and Engineering from MLR Institute of Technology and Management.

Dr. K. Venkateswara Reddy is a principal of MLRITM, Hyderabad, received his M. Tech and Ph. D from JNTU and Osmania University. He checquered a dynamic career in reputed engineering colleges as a professor, Head and Vice-principal and is found to be disciplinary and dynamic personality in academic and administrative spheres. He bagged several national and international journals to his credit in 20 years length of his service. He is life member of ISTE. Presently he is working in Cloud Computing, Network Security, MANET and other emerging fields of computer science.