A FRAMEWORK FOR DATA MINING USER NAVIGATION PATTERNS

V. VIJAYALAKSHMI

Assistant Professor, Department of CSE,
CCET, Pondicherry University, Puducherry.
vivenan09@gmail.com

R. T. SHOWMYA

M.Tech (CSE), SMVEC, Pondicherry University
Puducherry.
showmya.rt@gmail.com

ABSTRACT:

Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every year, data mining is becoming an increasingly important tool to transform this data into information. In this paper, we present a framework for mining user profiles for website management from log files. The website under study is part of a nonprofit organization that does not sell any products but only provides free information. Hence an approach for tracking and discovering evolving user profiles is described. It is also described how the user profiles can be used to enrich the web site management. Keywords present in the search text box are extracted for further updates.

Keywords: clustering, log files, mining, user profiles, mining evolving click streams.

1. INTRODUCTION

Analyzing the web site traffic may be fascinating study of how users traverse the web pages [1]. Ideally one would want to preserve in the form that combines analysis with rapid access. A recent time has made it imperative to use automated data mining techniques to discover Web user profiles. Although there have been considerable advances in Web usage mining, there have been no detailed studies presenting a fully integrated approach for mining the day wise characteristics. This method is complicated in evolving profiles, dynamic content, and the availability of taxonomy or databases in addition to Web logs.

In this paper, a complete framework for and a summary of experiences in mining web usage patterns with day wise characteristics is provided. The web site in study does not sell anything it just provides the information that are complete, accurate and up to date. As a result it validates evolving multifaceted user profiles on web site. A multifaceted user profiles are gathered to get clustered count of the viewed pages for update according to the user needs. Search engine queries are also gathered for further updates.

This paper is organized as following: Section 2 describes closely related works. Section 3 describes the structural design of the proposed system. Section 4 discusses about the steps involved in mining the user profiles. In Section 5 result analysis is found to be carried out. Section 6 concludes the paper and suggests further research extension.

2. RELATED STUDIES

A user profile consists of a username and additional properties we collect and store about a user. These properties can be used to personalize the users experience in our portal [11]. Properties can consists of personal data, work related data, geographic data or something else that logically categories the users. We can create user profiles and edit user profiles default property value. Administrator can edit the profiles property values in the administration console.

A user profile is a collection of user profile value for a user from all available user property set. Each piece of Meta data in a user profile is called a user property. The properties we create can be used to define the rules for personalization, delegated administration or visitor entitlement. Customer Relationship Management is

considered as an information industry term for methodologies, software and usually internet capabilities that help an enterprise manage customer relationship in an organized and efficient manner [12]. In many cases an enterprise builds a database about its customer. Database describes relationships in sufficient details so that the management, sales person and customer service can access information, match customer needs with product plans and offerings. Organizations with global operations must manage customer interactions in different languages, time zones, currencies and regulatory environments. Clustering is a common descriptive task where it seeks to identify a finite set of categories or clusters to describe the data.

3. OVERVIEW OF THE PROPOSED SYSTEM

Design involves identification of classes, their relationships as well as their collaboration. In objectory, classes are divided into entity classes, interface classes and control classes. The Computer Aided Software Engineering (CASE) tools that are available commercially do not provide any assistance in this transition. CASE tools take advantage of Meta modeling that is helpful only after the construction of the class diagram. Objectory used the term “agents” to represent some of the hardware and software system. In Fusion method, there is no requirement phase, where a user will supply the initial requirement document. Any software project is worked out by both the analyst and the designer. The analyst creates the use case diagram. The designer creates the class diagram. But the designer can do this only after the analyst creates the use case diagram. Once the design is over, it is essential to decide which software is suitable for the application.

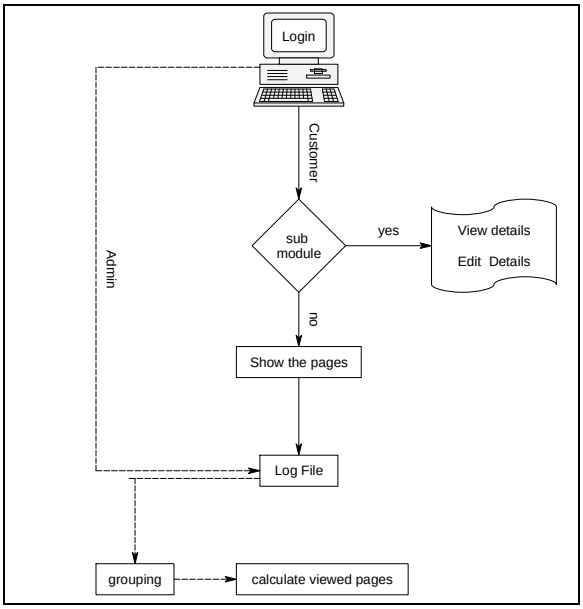

Fig. 1. Overall System Design

Fig. 1 shows the two modules and the operations performed in each module. First module is customer module and second is Admin module. In the customer module it has some sub module after the login success customer can go only in his sub modules. Categories are clusters to describe the data.

4. FRAMEWORK FOR MINING

The Web site is a portal that provides access to news, events, resources, company information (such as companies or contractors supplying related products and services), and a library of technical and regulatory documentation related to corrosion and surface treatment.

Here two main modules are considered. They are customer site and admin site. By the use of these modules only customer views the links and the admin views calculate the viewed pages. Of this under the former part the customer if being a new user is requested to register by providing certain details and granted permission by the admin to use the particular site, by providing each customer a user name and password. In this module only user can login and see the links and can gather information from the site. It is further divided into three sub modules.

The first module is important in the customer module it is registration module. It is important because to view the site every user need username and password so it needs to register. To register user click the registration module in show a page in that page it contains A field like name, username password etc .to register fill the all details and enter before The values going to database it check weather the current username is available or not and then insert in database suppose the username is not available in show the message That the username is already exist .by use of this username and password the user login and view the pages.

The second module is to edit and view the details. This module is use to edit and save the user details, After the registration the user want to do some change in this details means user login by Username and password and select the edit option .The user can edit all the details and then press save .It will update the current details in database.

The third module is user login. This is the main module in the customer module because to view the site user wants to login with the use of his username and password. When the users type his username and password it checks in the database weather the give username and password is correct or not.

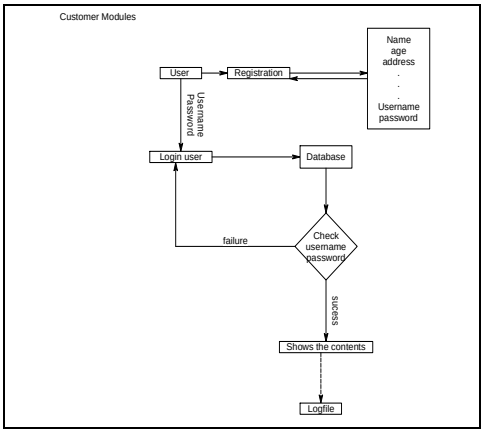

Fig. 2. Customer Module

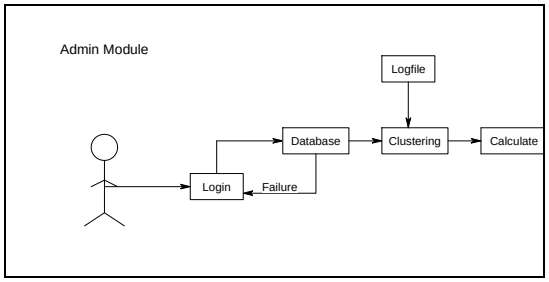

Fig. 3. Admin Module

Next is the admin module. This module is very important in the project because in this module log files is created and calculate the links which set more target increase. In this module it has two sub modules. The first module is log file. By the use of this file only admin calculates the number of page the user requested, The log file is created by admin only .In that file it contains all the information about the site. so admin insert some data in the log file like username, links which the user clicks ,date and time of clicks etc.. [6].In the existing system this information is stored in cookies which are in client side so admin doesn't know what the links the user clicks are. To solve this problem admin created the log file in server side .so it make easy to the admin to see what are the links the users visited. The second module is clustering. In this module only admin make the user in to group (cluster) by Hierarchical unsupervised niche clustering (HUNC) and unsupervised niche clustering (UNC) algorithm.

Hierarchical Unsupervised Niche Clustering (HUNC): In this algorithm the clustering the user data by using of log file is easily done. The log contains user session, set of URLs, ip address, date and time of page clicked .All the data from log file are taken and inserted into the database from that table where the set of values is given (which are in log file) to HUNC to read all the data's and make in to a group because it is easy to calculate .the output of this algorithm is a profiles and a set of URLs which the user reads. The reason using H-UNC instead of other clustering algorithms is that unlike most other algorithms, H-UNC can handle noise in the data and automatically determines the number of clusters.

Unsupervised Niche Clustering (UNC): In this algorithm the data from database is taken and separate the user into group (study, sports, news) and create the tables in to that group and insert the links in that table by finding the username which they select the groups when they register because it is very difficult to calculate the total number of links which are present in the log file .To make it easy separate the user into groups (by UNC) and the links can be calculated.

5. IMPLEMENTATION AND RESULT ANALYSIS

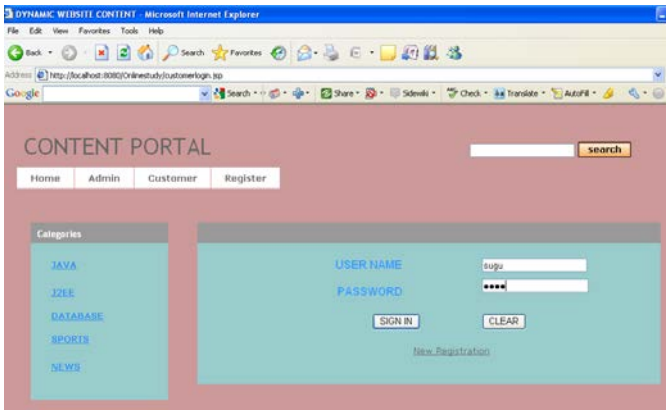

Fig. 4. Customer Login

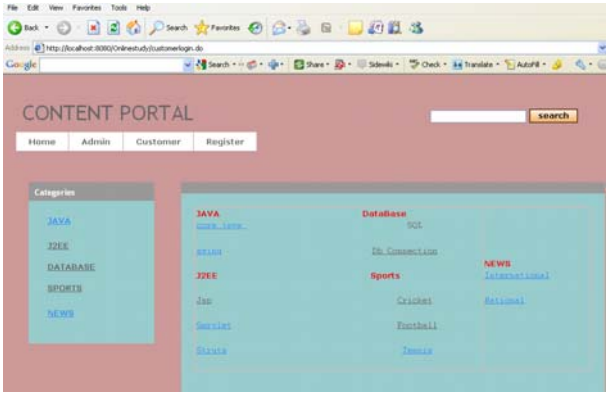

Fig. 5. Contents /Information provided

Fig. 4 &5 explains the customer module. If new user, customer must register. On registering admin will provide the username and password. With the help of the Username and password the customer can enter the particular sit and view the information that is provided. At times of registering users interest is got (study, sports, news) based on these details provided by the user only clustering action is found to be carried out.

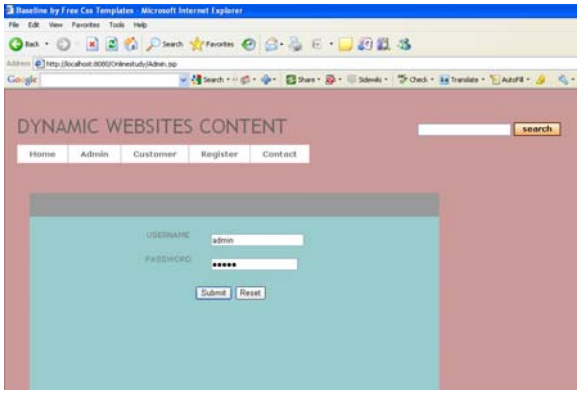

Fig. 6. Administrator Login

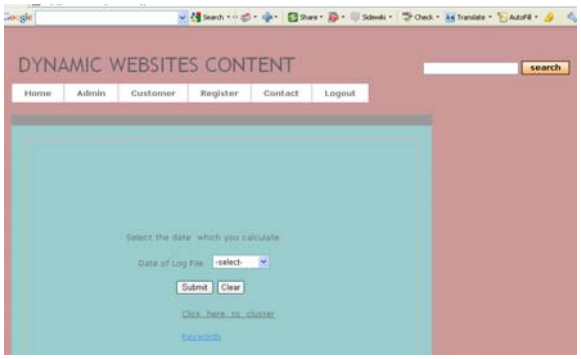

Fig. 7. Cluster Operation

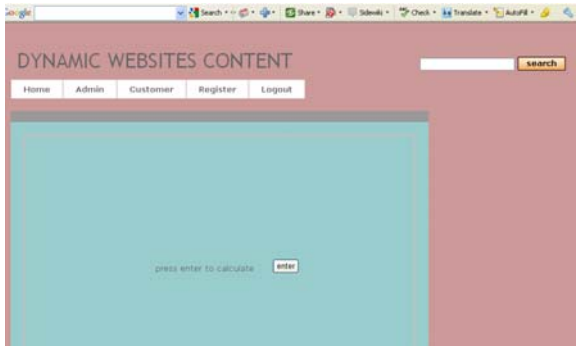

Fig. 8. Calculation

Fig. 6, 7 & 8 explains the admin module. The administrator after performing a successful login is asked to enter the date in order to perform the calculation. On submitting the cluster action is formed to be carried out.

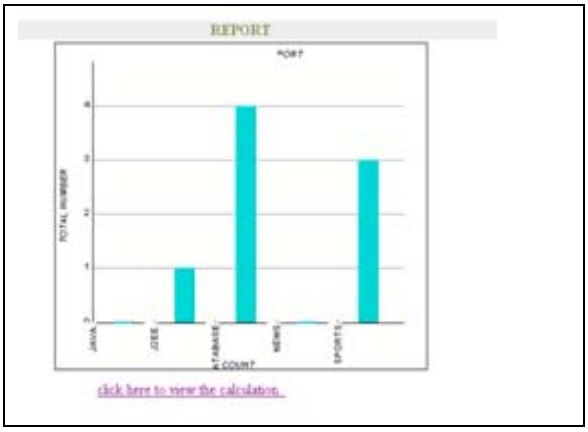

Fig. 9. Complete View

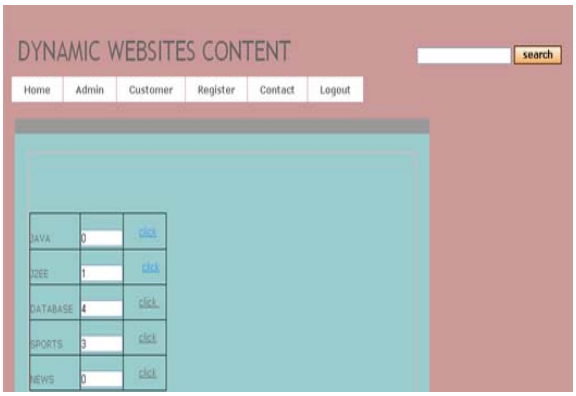

Fig. 10. Detail View

Fig. 9&10 gives the complete view. Fig. 9 generates a graph where it shows the count of total number of users view on particular topic so that further updations can be performed. Fig. 10 shows a more detailed view of viewing a topic deeper.

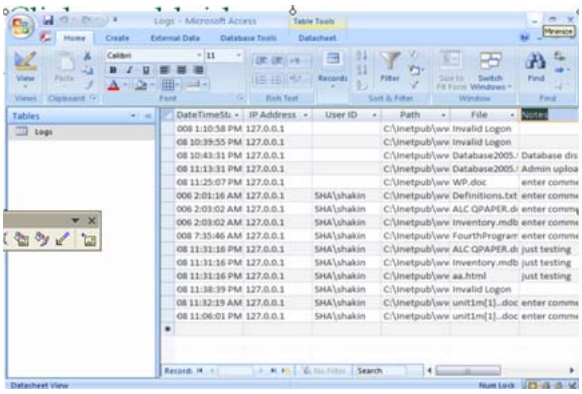

Fig. 11. Existing Log File

Fig. 11 shows the existing log file. This kind of log file is found to be having records of user in particular document or file it. It was considered to be complicated in evolving profiles, dynamic content.

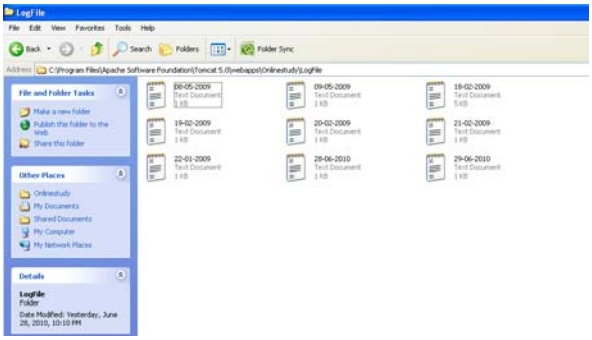

Fig. 12. Proposed Log File

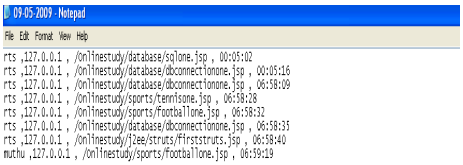

Fig. 13. Details in the log file

Fig. 12 shows the new form of log files. The proposed log file stores each information about user session, URL, ip, address, date and time of the page clicked in a complete, accurate and up to date manner. Data mining techniques have been applied to extract usage patterns from Web log data. This process, known as Web usage mining, is traditionally performed in several stages [1] [3] to achieve its goals. Collection of Web data such as activities/click stream recorded in Web server logs. Preprocessing of Web data such as filtering crawlers requests, request to graphics, and identifying unique sessions. Analysis of Web data also known as Web Usage Mining to discover interesting usage patterns or profiles. Interpretation /evaluation of the discovered profiles. Tracking the evolution of the discovered profiles is found out.

Clustering is performed of the user sessions extracted from the Web logs to partition the users into several homogeneous groups with similar activities and then extract user profiles from each cluster as a set of relevant URLs. Discovered user profiles are tracked, and their evolution pattern is categorized. When clustering the user sessions, the Web site hierarchy is exploited to give partial weights in the session similarity between URLs that are distinct and yet located closer together on this hierarchy.

Sample:

On User Registering, their interest details are being collected based on which clustering carried out. For example interest is found to have following categories(sports, news, studies).The website is found to have the information and links about JAVA,J2EE,Asp.

Of this the users interest under sports ,news and studies categories are determined

Sports —> Jack, Mathew

News —> Hena, Kelvin

Studies → Cathy

It is found out that the users have viewed the pages in the following manner.

Jack - java, j2ee

Siva - java

Kelvin – asp, java

Cathy – asp, java

Now the total number of users of different categories viewing a particular page i.e. the count is determined as follows

Based on the above details it is determined that under

Sports category -> Java is viewed by 2 users ->Is named as J1

News category -> Java is viewed by 1 user ->Is named as J2

Studies category -> Java is viewed by 1 user ->Is named as J3

Sports category -> J2EE is viewed by 1 user ->Is named as JE1

News category -> asp is viewed by 1 user ->Is named as A1

Studies category -> asp is viewed by 1 user ->Is named as A2

Total:

Java : J1+J2+J3->(2+1+1) = 4

J2EE : JE1 ->(1) = 1

ASP : A1+A2 ->(1+1) = 2

6. CONCLUSION

We presented a framework for mining, tracking and validating evolving multifaceted user profiles on web sites. A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engines queries. It is noted that update is carried out in such a manner all users are capable to view. In future enhancement update is to be carried out individually for each user profiles present for some users might have viewed the content some might have not. So with the help of this log file records the user who have viewed the details must alone be individually updated with new details.

7. REFERENCES

[1] R.Cooley, B.Mobasher, and J.Srivastava, "Web Mining:Information and Pattern Discovery on the World Wide Web,"Proc. Ninth IEEE Int'l Conf. Tools with AI (ICTAI '97), pp. 558-567,1997.

[2] O. Nasraoui, R. Krishnapuram, and A. Joshi, "Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator," Proc. Eighth Int'l World Wide Web Conf. (WWW '99),pp. 40-41, 1999.

[3] O.Nasraoui, R.Krishnapuram, H. Frigui, and A. Joshi, "Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering," Int'l J. Artificial Intelligence Tools, vol. 9, no. 4, pp. 509-526, 2000.

[4] J.Srivastava, R.Cooley, M.Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations , vol. 1, no. 2, pp. 1-12, Jan. 2000.

[5] M. Spiliopoulou and L.C. Faulstich, "WUM: A Web Utilization Miner," Proc. First Int'l Workshop Web and Databases (WebDB '98), 1998.

[6] T.Yan, M.Jacobsen, H.Garcia-Molina, and U.Dayal, "From User Access Patterns to Dynamic Hypertext Linking," Proc. Fifth Int'l World Wide Web Conf. (WWW '96), 1996.

[7] M.Perkowitz and O.Etzioni, "Adaptive Web Sites: Automatically Learning for User Access Pattern," Proc. Sixth Int'l WWW Conf. (WWW '97), 1997.

[8] J.Borges and M.Levene, "Data Mining of User Navigation Patterns," Web Usage Analysis and User Profiling, LNCS, H.A. Abbass, R.A. Sarker, and C.S.Newton, eds. pp. 92-111, Springer-Verlag, 1999.

[9] O.Zaiane, M.Xin, and J.Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," Proc. Advances in Digital Libraries (ADL '98), pp. 19-29, 1998.

[10] O.Nasraoui and R.Krishnapuram, "A New Evolutionary Approach to Web Usage and Context Sensitive Associations Mining," Int'l J.Computational Intelligence and Applications, special issue on Internet intelligent systems, vol. 2, no. 3, pp. 339-348, Sept. 2002.

[1] http://en.wikipedia.org/wiki/User_profile

[2] http://www.twocrows.com/crm-dm.pdf. Building profitable customer relationship with data mining. Herb Edelistine, President.

Authors

Mrs. V. Vijayalakshmi working as Assistant Professor, Department of Computer Science and Engineering in Christ College of Engineering and Technology, Pondicherry University, Puducherry. She completed M.Tech (CSE) in Sri Manakula Vinayagar Engineering College, Pondicherry University. She holds a MCA can be reached via vivenan09@gmail.com.

Mrs. R.T. Showmya completed M.Tech(CSE) in Sri Manakula Vinayagar Engineering College, Pondicherry University, Puducherry. She holds a B.Tech can be reached via showmya.rt @gmail.com.