User Future Request Prediction Using F.P Tree

Rujuta Panvalkar, Namrata Valera,Ami Vashi, Khushali Deulkar
Dwarkadas J. Sanghvi College of Engineering
Mumbai-400 056. India
rujutapanvalkar@gmail.com, valeranamrata@gmail.com,
ami.vashi34@gmail.com, khushalideulkar@gmail.com

Abstract: *The use of Web technology has increased by a great extent in the recent times. Millions of users spend time surfing net to obtain information or for recreational activities. Along with satiating their own purpose, the users leave behind a detailed trail of all the web pages accessed and the frequency with which they are accessed. This information is of paramount use to several commercial domains like e-commerce websites, social networking sites, entrepreneur franchises, etc. Web usage mining satisfies the basic purpose of amalgamating information from a user's web history and finding patterns in the usage characteristics. Web usage mining is the extension to the traditional data mining and also forms the base of our paper. Our paper emphasizes on predicting a user's future request. The attributes of web usage mining help to determine a pattern depending on a user's navigational footprints and actions. This pattern is then analyzed, giving us tools to predict requests that the user is likely to make in the future.*

Keywords: Future Request, Prediction, Surfing, Web History, Web Usage Mining Introduction

Recieved: 12 November 2014, Revised 15 December 2014, Accepted 20 December 2014

1. Introduction

Today Internet has become more of a necessity than a luxury to people. Every day, inestimable websites are created, visited, and merged. People use internet for almost everything and hence, it is but natural that Web domain is a domain that hasn't yet been explored to its full potential. Whenever anything is accessed on web, that particular page is saved in the device's web history. This as we call is the temporary trail of the web accesses. The more permanent trail is that which is stored in the web logs. Web logs act as the chart of user's behavior on a particular machine. Most of the web log data is generally automatically generated by the web servers. A colossal amount of data is uploaded, downloaded or just viewed on the World Wide Web. This also means a number of users are very much interested in carrying out communications and transactions using this medium. This in turn has a tantamount effect on the commercial businesses having a huge interest in using the web services. As users form the base of this commercial build-up, the comfort of the end users is given ultimate importance. Now, as the advances in graphic content can only satiate the user to some extent, new advances are being made in field of web usage mining which is more appealing to the user as it affects in reducing the efforts on the end-user side. Future user prediction is one such field which benefits the user as well as the Web sites.

The basis of the paper is to correctly predict what the end user wants to see on the page he opens. This is accomplished by first

accumulating the users' data from the web logs generated by the servers. This data is then cleaned so that only the relevant information is made available for the further process. Once this step is done, the remaining data is grouped into clusters according to the various users using the same machine and then classified according to the domains to which the accessed information belongs. For example, all the sports data of user A will be classified together, all the data relating to politics will be classified together. In another cluster, will exist the sports data of user B, the educational data of user B and so on. Thus according to the variance and access frequency of the users, the total accessed data will be clustered and classified and will be made available for future processing. Now, when any of the existing users logs in the machine, using the algorithm discussed below, new suggestions will be made for the user's future requests in accordance to the pages accessed by the user on all the previous instances.

We mainly focus on Web Usage Mining. It is a type of web mining which extracts interesting and useful patterns about the user's navigational behavior. This activity helps developers to understand individual user's psyche and helps them to customize the services provided to that particular user. This customization is now an integral and attractive quality about software or website as it makes the user's navigation easy and smooth.

The rest of the paper is divided as follows- Section [2] explains the Review of Literature. Section [3] states the Proposed Solution. Section [4] states the Architectural Design. Section [5] gives the System Implementation. Section [6] states the Working on the project. Section [7] states the Testing Approaches. Section [8] gives the Result and finally, Section [9] states the Conclusion and Future Scope.

2. Review of Literature

Some of the important references used in this project are :

The explosion of data all over the world has led us to strive to find ways to manage and use the available data in more than one way. The aim is to convert heaps of 'data' to '*useful knowledge*'. Hence it was important for researchers all over to create client and server side technologies which can carry out this conversion.

Thus came the concept of web mining. What we are going to deal with is web usage mining, a mining approach for user browsing and access patterns. Analysing this browsing data can help organisations to better understand their customers, design strategies, evaluate user response and conduct surveys. [1]

The major tasks that are to be handled are :

a. Preprocessing

b. Pattern discovery

Preprocessing is mainly data cleaning. It basically retains data that can be useful and discards or eliminates data that is irrelevant to the purpose. The second major preprocessing task is transaction identification. Before any mining is done on Web usage data, sequences of page references must be grouped into logical units representing Web transactions or user sessions. A user session is all of the page references made by a user during a single visit to a site. Identifying user sessions is similar to the problem of identifying individual users, asdiscussed above. A transaction differs from a user session in that the size of a transaction can range from a single page reference to all of the page references in a user session, depending on the criteria used to identify transactions.

Content Preprocessing handles content like Images, text, scripts and other files such as multimedia files are converted into useful data for Web Mining Processes. The process involves classification and clustering. Result of a classification is such that what type of pages has been visited or what class of products has been searched.

Pattern discovery has wide range of applications like on statistical data, data mining and machine learning .The author has limited the coverage of pattern Discovery in the field of Web Domain. The Pattern Discovery in the Web Usage Mining to analyze and Discover the Pattern that has been generated by Server sessions which is the sequence of pages requested by the user. Statistical analysis, association rule, clustering, classification, sequential pattern, etc are some of the techniques used in pattern discovery.

RenataIvancsy and IstvanVajk [4] presented discovering frequent patterns in Web log data is to obtain information about the

navigational behavior of the users. The different patterns in Web log mining are page sets, page sequences and page graphs. Devinder Kaur and Ravneet Kaur [3] talk about today the World Wide Web is popular and interactive medium to distribute information. The web is huge, diverse, dynamic and unstructured nature of web data, web data research encountered lot of challenges for web mining. Information user could encounter following challenges when interacting with web.

a. Finding Relevant Information- People either browse or use the search service when they want to find specific information on the web. Today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.

b. Creating new knowledge out of the information available on the web- This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that already has collection of web data and extract potentially useful knowledge out of it.

c. Personalization of information- When people interact with the web they differ in the contents and presentations they prefer.

d. Learning about Consumers or individual users- This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to website design management and marketing.

3. Proposed Solution

We propose to formulate an improved F.P tree algorithm to implement the solution to the problem formulation. There are multiple steps involved in this process. We aim to make a website by using the basics of ASP.NET and C#. The website will be the host to the implementation of the web usage mining solution. The problem statement of our Web application is to design a Course Recommendation Model using FP Tree. The main aim is to make the website dynamically customized to each particular user according to the web content searched by the users. The next step is to devise the algorithm. That is done by researching the literature studied in part 2 of the report. Finally, the last step is to use the algorithm in the website to categorize the content. The following steps are involved

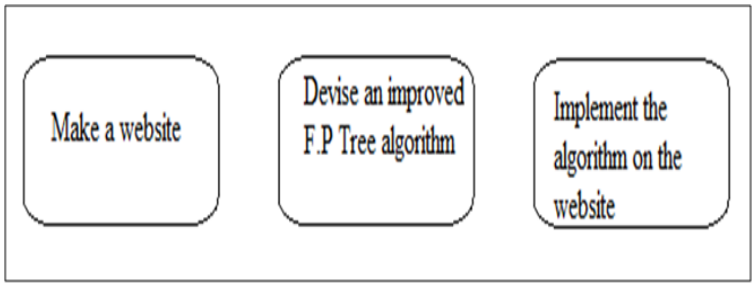

Figure 1. Illustration of algorithm

The F.P Tree algorithm works as follows-

Step1: In the first step we actually build the data for the mining. We build web log content based on different user preferences and navigations. For every access to the course we build a log file. The log file format has been kept very simple, involving date of the log entry and the list of keywords mapped to the relevant Course.

Step 2: In this step there is Pattern Discovery which is performed by the Frequent Pattern (FP) which involves FP Tree which in turn is FP growth .FP tree method is used in Data Mining .It consists of two passes over the Data Set .In the first Pass it scans data and find the minimum support for the each item. The item set whose support is less than minimum is discarded .The Data item that is included is the Course Page that is being visited by the User. Next step is to generate a decreasing order on the basis of frequency of occurrence of the Item Set Which is the keywords of all the course pages visited by the User. In the Second Pass of the FP Tree, Transaction is being read .In this work the Transaction is the number of times the user visited the particular Web Site. The Read Transaction is iterated until all the Transaction is being completed. After Reading all the Transaction the results of the F.P Tree are displayed dynamically.

Step3: In this step Pattern analysis is done and in this Candidate rule is generated and on the basis of candidate rule confidence is generated. On the basis of pattern analysis Prediction is done of the User's Future request. Whenever the user clicks on Recommended Courses, we showed this predictions of the results.

4. Architectural Design

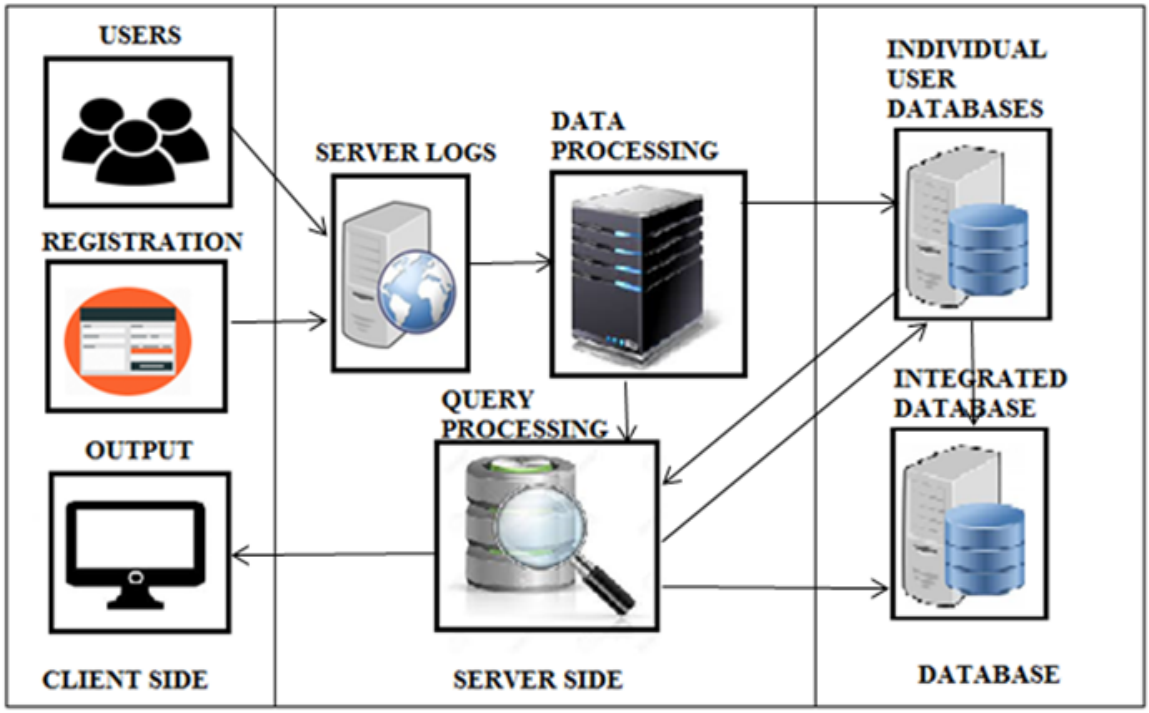

Figure 2. Architectural Design

The above figure depicts the architectural design of the system devised. The system is based on the client server design and incorporates varied components for its functioning. The client side includes the user interface which is used by the users to interact with the system and the registration information which serves as an initialization point for each user's logs. The server side consists of the web server logs which include the default logs created by the server in addition to the user's logs. The information contained in the server logs is subjected to data processing which includes the ETL process. The query processing component is responsible for implementing the queries on the logged data and presenting information according to user needs. There are two databases involved in the overall system. One database is used to store individual user logs and is updated every time the algorithm and the query processing takes place and the second database is the integrated database which includes the data of all the users collected together. Each time the individual databases are updated, the integrated database is also scheduled to update.

5. System Implementation

Basically here we have to define three types of user viz., Admin, Teacher and Student. Being a role based system there are different levels of authorization for the user. For instance, a student cannot access the profile of the faculty. Moreover he cannot upload the files on the server. Hence, we have to implement the role based architecture. For this implementation, we would use the ASP.net membership provider feature that will allow us the entire user management task with one go. We just have to define the restrictions in the web configuration file.

Another part of the design is the Recommendation Courses using FP Tree. Here, the users must be able see the recommendations based on their profiling using Web Log Content. We are implementing the FP Growth and FP Tree based on the keywords logged in for the user. These recordings are based on the users' access to different courses.

All these modules would be developed individually and then they are integrated using drivers and stubs. The other feature is the Evaluation and Assessment of the user performance. For this, the faculty can broadcast their feedbacks. He'll write the feedback and broadcast it to the users.

The following figure depicts the flowchart of the entire process-

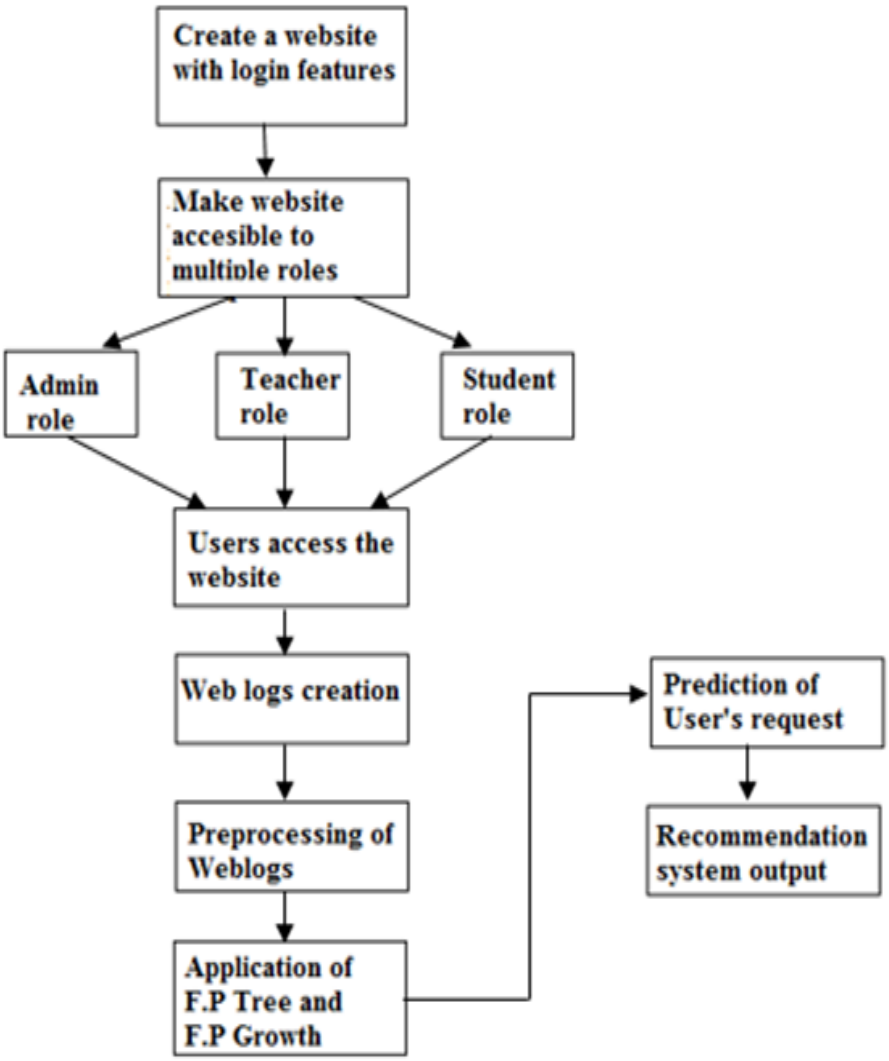

Figure 3. Overall Flowchart

6. Working on the project

There will be basically 3 modules for the project. This is an estimate on a higher level view of the application. They will have multiple corelative sub modules further as per the required coupling and integration. The modules are namely, User Membership, Course Management, and Recommendation System.

User Membership

Basically here we have to define three types of user viz., Admin, Teacher and Student. Being a role based system there different levels of authorization for the user because for instance a student cannot access the profile of the faculty. Moreover he cannot upload the files on the server. Hence, we have to implement the role based architecture. For this implementation, we would use the ASP.net membership provider feature that will allow us the entire user management task with one go.

Course Management System
All these modules would be developed individually and then they are integrated using drivers and stubs. The feature helps a CRUD system with Courses and Departments. Departments and Courses are added by Admin. Teachers do not have access to these features. But Teachers can add content with the relevant courses. Teachers can add, update, and delete the content for the Courses.

Recommendation System
We have a basic recommendation system in place, which recommends the courses based on the FP Tree Growth and Keyword Mapping per Course. For every course, visited keywords for the course are recorded in the log along with the date time. Then whenever the Recommended Courses are accessed, it reads all the logs for the particular user, sorts the keyword list, per click. After that, it performs the FP Growth and gets a list of frequent Item sets. We pick Frequent Itemsets and use those keyword mapping to find the Courses. We perform the search based on the maximum matching first.

7. Testing Approach

Software Testing is a process of evaluating a system.

- By manual or automatic means
- Verify that the system satisfies the specified requirement
- Identify differences between expected and actual results

Testing analyzes a program with the intent of finding problems and errors that measures system functionality and quality. Testing includes inspection and structured peer reviews of requirement and design, as well as execution test of code. The code developed during coding activity is likely to have some requirement errors and design errors in addition to the errors introduced during the coding activity. Testing perform a very critical role for quality assurance and for ensuring the reliability of software. The system must be tested to evaluate the actual system functionality.

The two basic approaches to testing are:

- Black Box or Functional Testing.
- White Box or Structural Testing.

7.1.1 White Box Testing
White Box Testing is related with structure of the program. To test the logic of the program various test cases are design which takes care of following:

- Every statement in the program is examined at least once.
- Every path in the program is executed at least once.
- Every logical decision is executed on their true or false side.
- Execute all their loops all their boundaries and within operational boundaries.
- Execute internal data structure and formulae to ensure their validity.
- This type of testing is performed by Developers.

White box testing for the system is performed as follows:

- Module wise Unit Testing is done.
- All the query calls were tested.
- All the fields in the web form are specified with certain length which is should be less than the corresponding field's datatype size in the database tables, so that size limit does not exceed.

- All the fields in the web form are test for valid entries like web pages.
- Proper user validation is done.
- The required fields in each form are validated so that the NO NULL fields are never left blank.
- Each and every query is tested using Microsoft Access Query Builder
- And if any error occurs a appropriate message is thrown.

7.1.2 Black Box Testing

- BBT is related with input and output and not related with internal structure of the program.
- In BBT it is checked if some input is given than whether specific output is produce by the program or not.
- The various sets of input test cases are prepared and applied on a program corresponding output are verified.
- This type of testing is done Test Engineers.

7.2.1 Unit Test Cases

Test Case	Expected Output	State
Test whether authenticated login is accepted	User should be authenticated and redirected to his home page	Pass
Try Invalid Login Attempt	User should be given errors	Pass
After login the correct home page with the specific roles should be visible	After login, student should be redirected to his home page and expert to his own home page.	Pass
User Registration	User should be registered with all the validations and the values	Pass
Course Progress	User should be shown his correct course progress	Pass
Start Course	Start course is divided into views like reading the notes, feedback extractor, objective test evaluation and subject evaluation. So everything should be in sequence	Pass
Recommendation Test Evaluation	Checking Recommendation	Pass
Categories	All the categories should be loaded well in the menu for the Admin, Student and Faculty as per the authorization	Pass
Logout	User should be able to logout and shouldn't be allowed access until he log-ins back	Pass

Table 1

7.2.2 Integration System Test Cases

Test Case	Expected Result	State
Authorization	After login, the state of the selection of the courses, etc should be maintained	Pass
View Menu	User should get the list of his menu items only. It means, a user should get access to the modules, that he has opted for and not others. Moreover, the trace of his actions over that database should also be maintained	Pass
The Multiview for the Inprogress course	The intransit view, where a cascaded view of each department can be seen well	Pass

Table 2

8. Results

The following results have been obtained on a small dataset. The dataset has been kept small on purpose for the reason of better understandability; this result can be replicated on even a large database. Figure 1 shows the frequency with which a particular user accesses the courses of his choice over a period of a week.

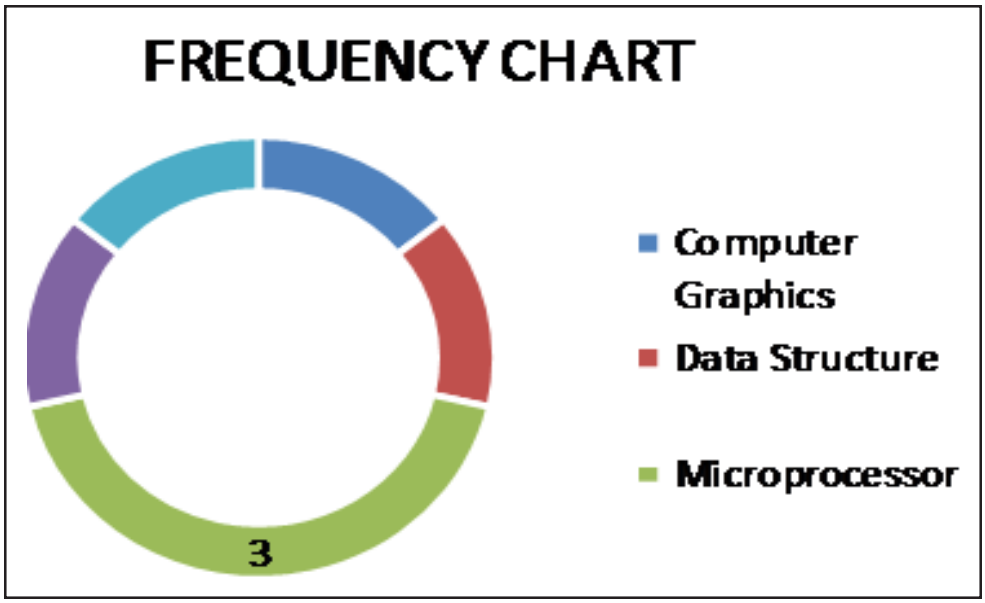

Figure 4 Frequency of the courses accessed

Figure 4 displays the log files and the recommended courses that are obtained after keeping the F.P Tree support as 40%. It can be observed here that the recommended courses are quite concise and only the five most accessed courses are displayed. The advantage of using the F.P Tree algorithm is that along with the recommendations, the order in which the user is mostly likely to view the content is also displayed.

Figure 5 depicts the comparison of the number of recommendations that are displayed as and when the minimum support of the F.P Tree is changed. Inorder to get the maximum number of recommendations, the support has to be kept as minimum as possible. Inorder to get the minimum and the most accurate number of recommendations, the support has to be kept as high as possible.

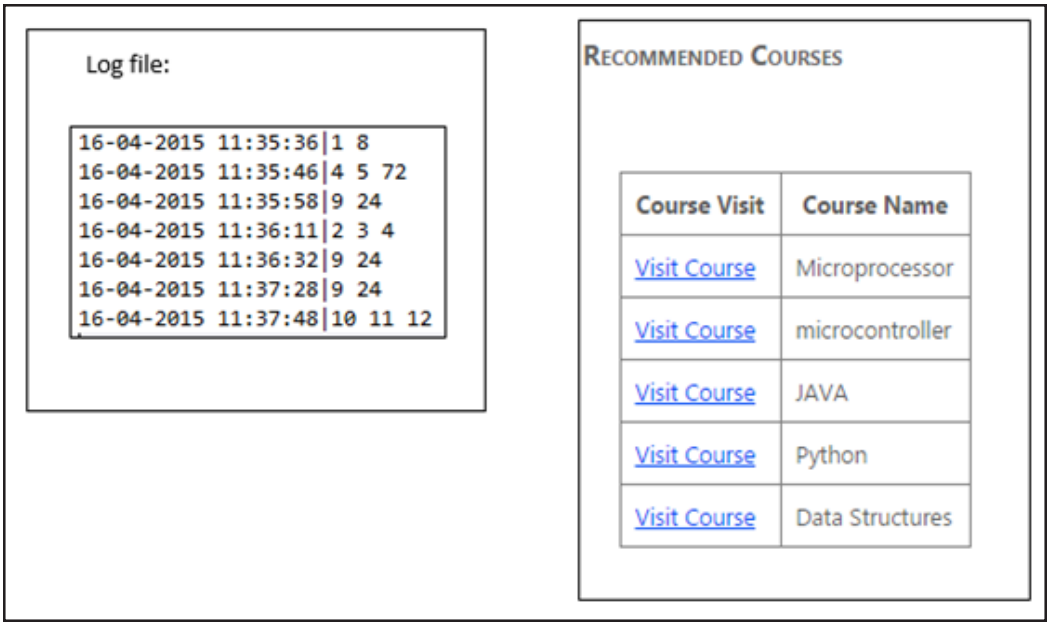

Figure 5. Log files and output recommendations

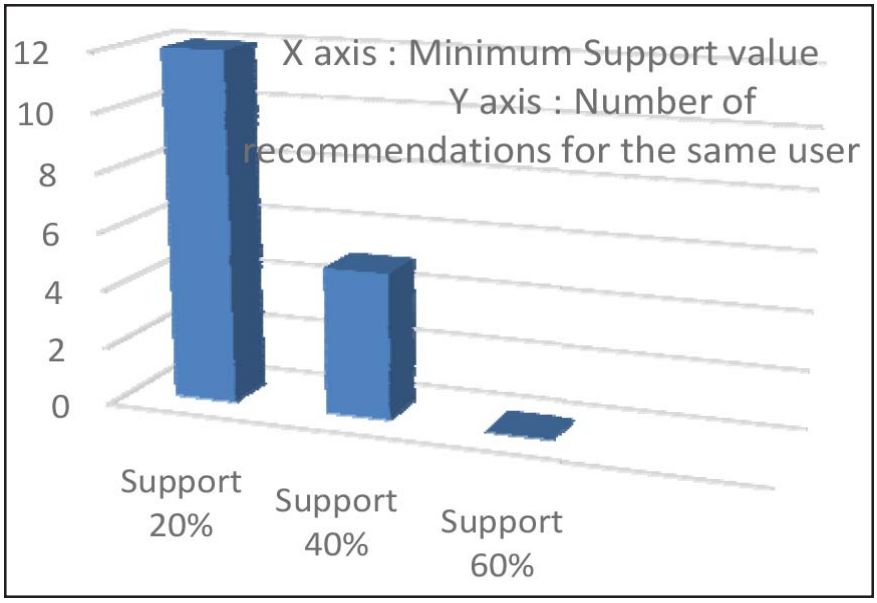

Figure 6. Comparison of recommendations at different thresholds

9. Conclusion and Future Work

The simulation result shows that the FP-Growth algorithm is used for finding the most frequently access pattern generated from the web log data. By using the concept of web usage mining we can easily find out the user's interest and we can modify and make our web site more valuable and more easily accessible for the users. The main goal of the proposed system is to identify usage pattern from web log files. If a large number of patterns and/or long patterns exist, the FP-growth algorithm is one of the fastest approaches for frequent item set mining. The FP-growth algorithm uses the FP-tree data structure to achieve a condensed representation of the database transaction using a divide-and conquer approach to decompose the mining problem.

Our experimental result shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns. In future the algorithm can be extended to web content mining, web structure mining.

References

[1] Metev, S. M., Veiko, V. P. (1998). Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag.

[2] Breckling, J., Ed. (1989). The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: *Springer*, 61.

[3] Zhang, S., Zhu, C., Sin, J. K. O., Mok, P. K. T. (1999). A novel ultrathin elevated channel low-temperature poly-Si TFT, *IEEE Electron Device Lett*., 20, p. 569–571, Nov.

[4] Wegmuller, M., von der Weid, J. P., Oberson, P., Gisin, N. (2000). High resolution fiber distributed measurements with coherent OFDR, *In*: Proceedings ECOC'00, paper 11.3.4, p. 109.

[5] Sorace, R. E., Reinhardt, V. S., Vaughn, S. A. (1997). High-speed digital-to-RF converter, U.S. Patent 5 668 842, Sept. 16.

[6] The IEEE website. (2002). [Online]. Available: http://www.ieee.org/

[7] Shell. M. (2002). IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[8] FLEXChip Signal Processor (MC68175/D), Motorola, 1996.

[9] PDCA12-70 data sheet, Opto Speed SA, Mezzovico, Switzerland.

[10] Karnik, A. (1999). Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP, M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan.

[11] Padhye, J., Firoiu, V., Towsley, D. (1999). stochastic model of TCP Reno congestion avoidance and control, Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02.

[12] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.